DATA516/CSED516
Scalable Data Systems and Algorithms

Lecture 6 Advanced Query Processing (cont'd)

Announcements

- Project feedbacks in your repo
  Milestone due on 11/29
- HW3 due on 11/16
- May postpone next week's paper review

Today's Lecture

Part 1: guest lecturer:

- Shan Shan Huang on Cloud Databases

Part 2: Advanced Query Processing (cont'd)

Review: Limitations of Traditional Query Processing

- Poor cardinality estimator may lead to poor query plan

- If the query is *cyclic*, then *any* plan is bad

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:

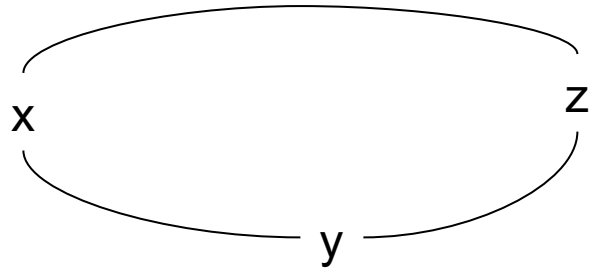
$$Q(X_1, \ldots, X_k) = R_1(Vars_1) \wedge \cdots \wedge R_m(Vars_m)$$

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

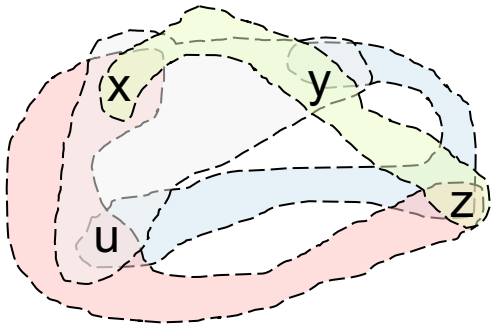
$$|Q| \leq |R_1| \times \cdots \times |R_m|$$

Conjunctive Queries are Hypergraphs

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



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



Simple Fact #2

- Consider any query:

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

- 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:

$$|Q| \leq |R_{i_1}| \times \cdots \times |R_{i_n}|$$

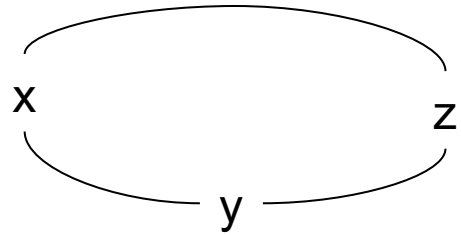
Why?

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

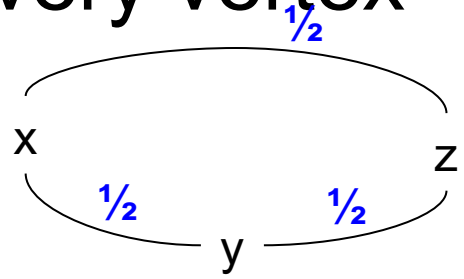
Fractional Edge Cover

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



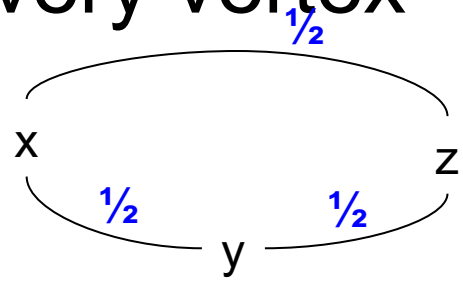
Fractional Edge Cover

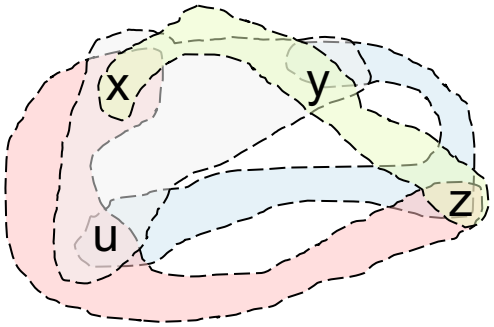
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:v \in e} w_e \geq 1$



Fractional Edge Cover

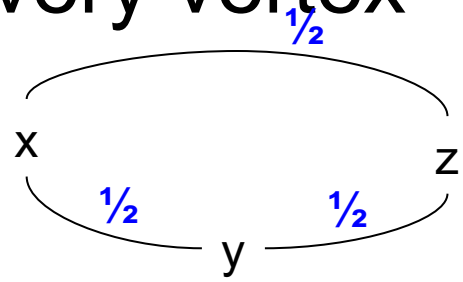
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:v \in e} w_e \geq 1$

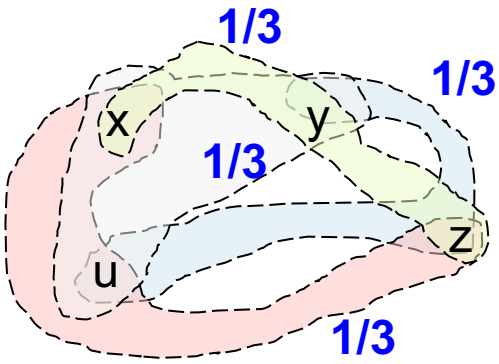




Fractional Edge Cover

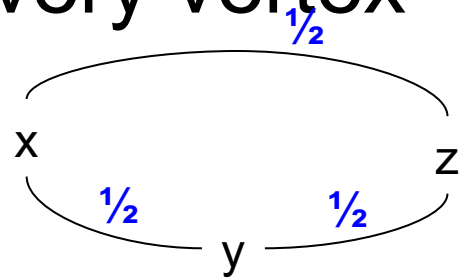
- A *fractional edge cover* of a (hyper)graph is a set of non-negative numbers w_e, one for each edge e, such that, for every vertex v: $\sum_{e:v \in e} w_e \geq 1$



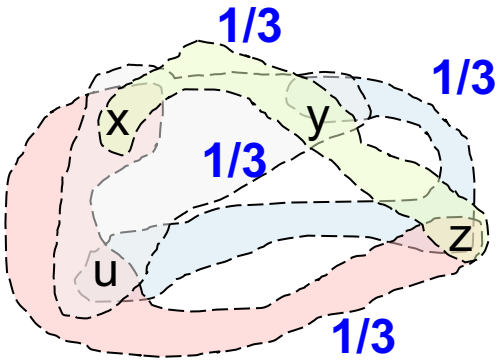


Fractional Edge Cover

- 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:

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

- 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

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

x —— y —— z —— u —— v

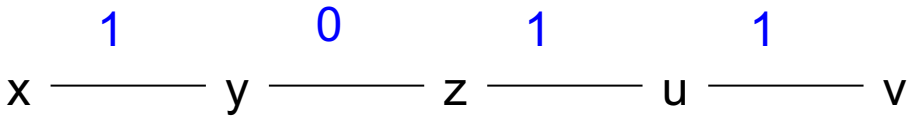
Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$

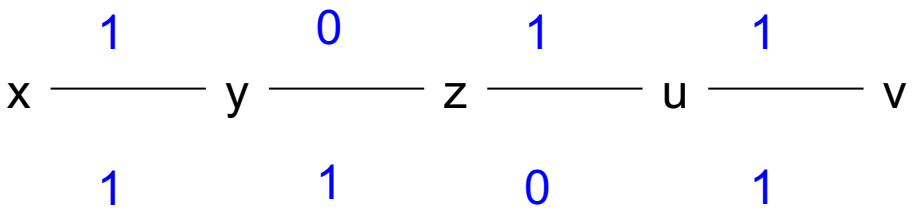


Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$$

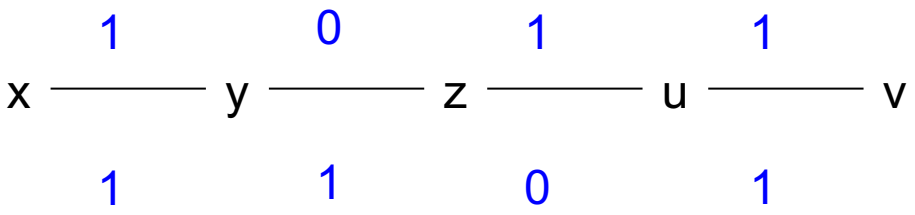
$$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$$

Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$$

$$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$$

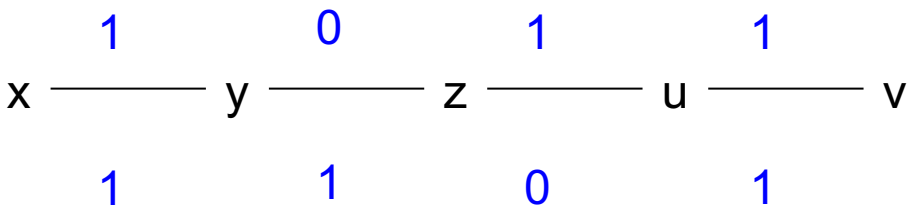
$$|Q| \leq \min(|\mathrm{R}| \cdot |T| \cdot |K|, |\mathrm{R}| \cdot |S| \cdot |K|)$$

Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

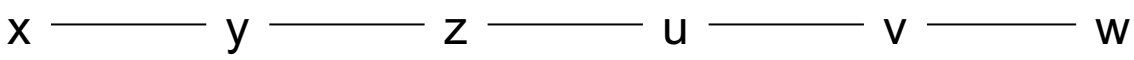


$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$

$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$

$|Q| \leq \min(|\mathrm{R}| \cdot |T| \cdot |K|, |\mathrm{R}| \cdot |S| \cdot |K|)$

- $Q(x, y, z, u, v, w) =$
  $R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v) \wedge L(v, w)$

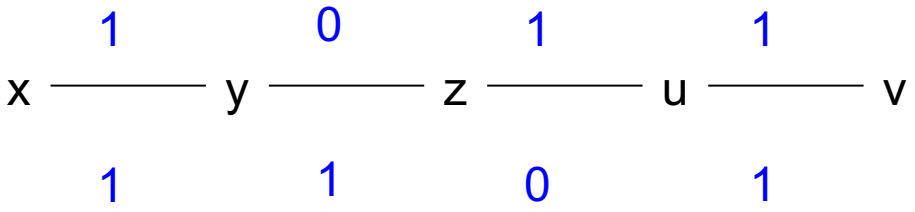


Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

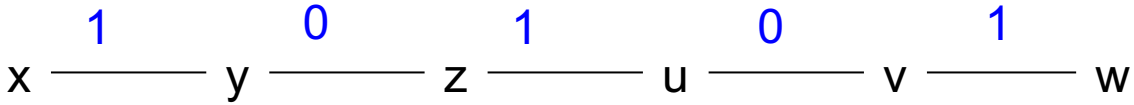


$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$

$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$

$|Q| \leq \min(|\mathrm{R}| \cdot |T| \cdot |K|, |\mathrm{R}| \cdot |S| \cdot |K|)$

- $Q(x, y, z, u, v, w) =$
  $R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v) \wedge L(v, w)$

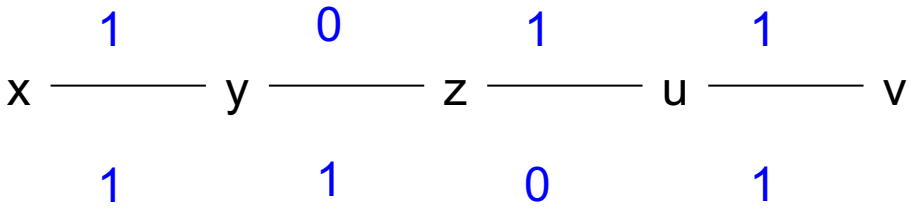


Examples

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

What are the maximum output sizes?

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

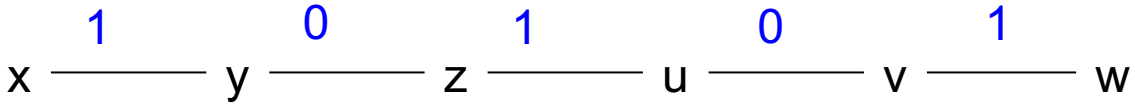


$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$

$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$

$|Q| \leq \min(|\mathrm{R}| \cdot |T| \cdot |K|, |\mathrm{R}| \cdot |S| \cdot |K|)$

- $Q(x, y, z, u, v, w) =$
  $R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v) \wedge L(v, w)$



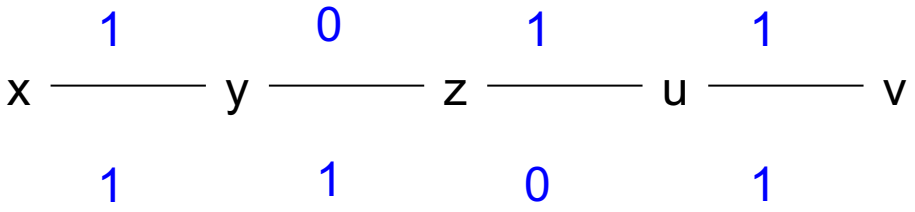
$|Q| \leq |R| \cdot |T| \cdot |L|$

Examples

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

What are the maximum output sizes?

- $Q(x,y,z,u,v) = R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,v)$

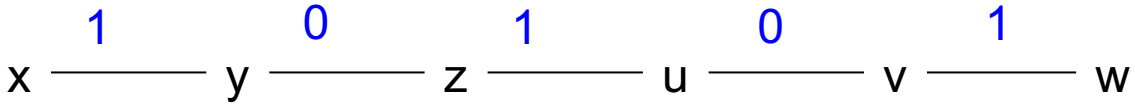


$|Q| \leq |\mathrm{R}| \cdot |T| \cdot |K|$

$|Q| \leq |\mathrm{R}| \cdot |S| \cdot |K|$

$|Q| \leq \min(|\mathrm{R}| \cdot |T| \cdot |K|, |\mathrm{R}| \cdot |S| \cdot |K|)$

- $Q(x,y,z,u,v,w) =$
  $R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,v) \wedge L(v,w)$



$|Q| \leq |R| \cdot |T| \cdot |L|$

S, K may be arbitrarily large!

Discussion

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

When all relations have same cardinality N:

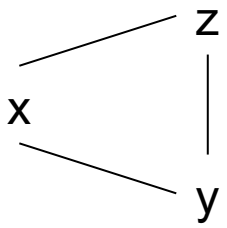
- $|Q| \leq N^{w_1+w_2+\cdots+w_m}$
- Edge covering number:
  $$\rho^* = \min(w_1 + w_2 + \cdots + w_m)$$
- $|Q| \leq N^{\rho^*}$

Examples

Assume all relations have size N

What are the maximum sizes?

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



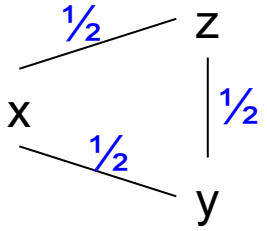
- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

Examples

Assume all relations have size N

What are the maximum sizes?

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



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

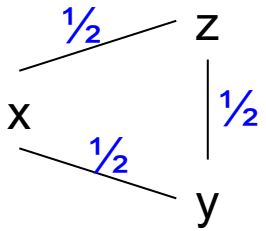
- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

Examples

Assume all relations have size N

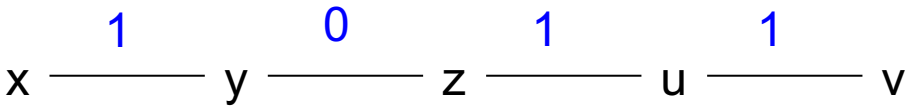
What are the maximum sizes?

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



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

- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



Answer: $|Q| \leq N^3$

Examples

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

Examples

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

Examples

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

Examples

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

Examples

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

Examples

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

Examples

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

Examples

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

Examples

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

Examples

Query	$w_1, w_2, \ldots, w_m$	$\lvert R_1\rvert^{w1} \times \cdots \times \lvert R_m\rvert^{wm}$	$\lvert Q\rvert \leq \ldots$
$R(x,y) \wedge S(y,z)$	1,1	$\lvert R\rvert \times \lvert S\rvert$	$\leq \lvert R\rvert \times \lvert S\rvert$
$R(x,y) \wedge S(y,z) \wedge T(z,x)$	½, ½, ½	$(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$	$\leq \min($ $(\lvert R\rvert \times \lvert S\rvert \times \lvert T\rvert)^{½}$, $\lvert R\rvert \times \lvert S\rvert$, $\lvert R\rvert \times \lvert T\rvert$, $\lvert S\rvert \times \lvert T\rvert)$
	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	$\lvert R\rvert \times \lvert S\rvert \times \lvert K\rvert$	
	1, ½, ½, 1	(no need; why?)	

Examples

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

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 $\tilde{O}(AGM(Q))$
- "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_{i_1}, R_{i_2}, ... be all relations containing x

Compute $D = \Pi_x(R_{i_1}) \cap \Pi_x(R_{i_2}) \cap \cdots$

for every value $v \in D$ do:

    Compute Q,

    where R_{i_1}, R_{i_2}, ... are restricted to $x = v$

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

Generic Join Example

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

Generic Join Example

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

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

Generic Join Example

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

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

Generic Join Example

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

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

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

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

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

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

Runs in time $\tilde{O}(AGM(Q))$

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 $|R_1| = \cdots = |R_m| = N$
- In this case $AGM(Q) = N^{\rho^*}$

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

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 *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: $\sum_x v_x \leq \sum_e w_e$

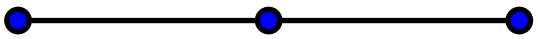
Theorem $\max_v \sum_x v_x = \rho^* = \min_w \sum_e w_e$

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



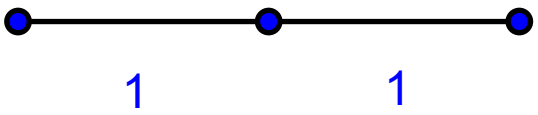
Fractional Vertex Packing

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

$\rho^* = 1$

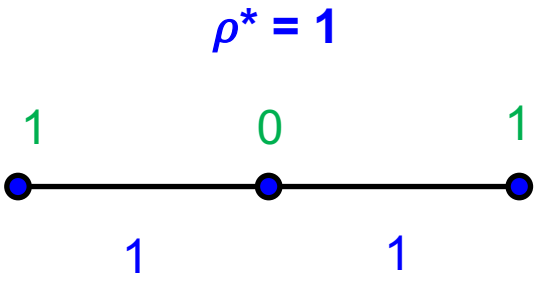


Fractional Vertex Packing

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

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

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

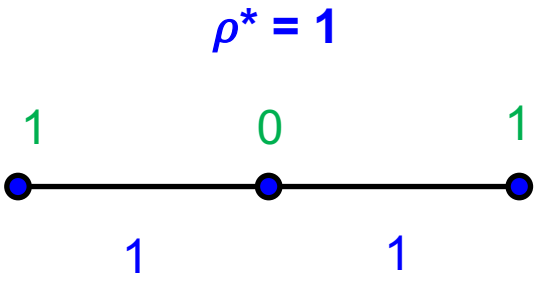


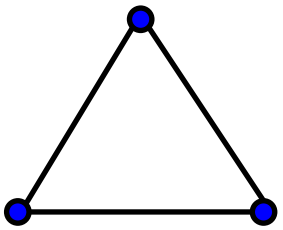
Fractional Vertex Packing

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



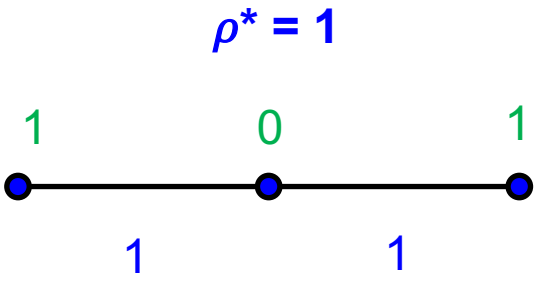


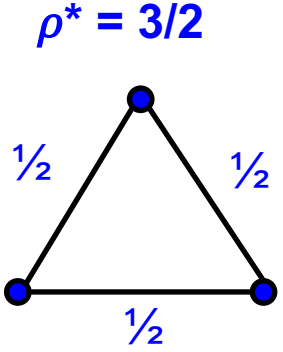
Fractional Vertex Packing

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





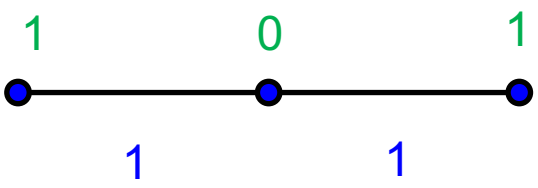
Fractional Vertex Packing

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

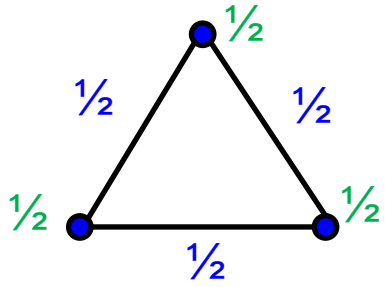
Fact For any v, w: $\sum_x v_x \leq \sum_e w_e$

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

$\rho^* = 1$



$\rho^* = 3/2$

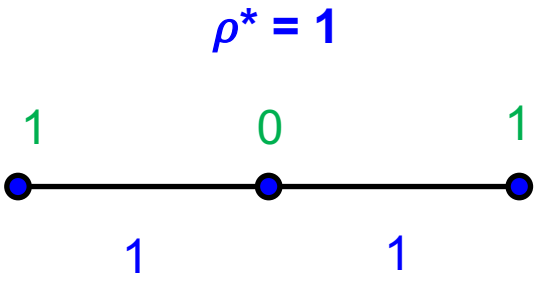


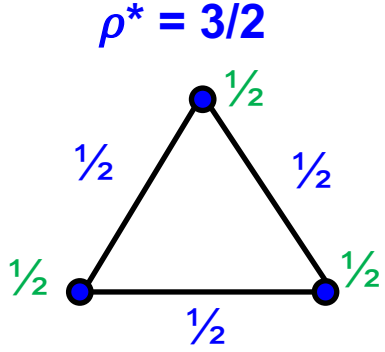
Fractional Vertex Packing

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





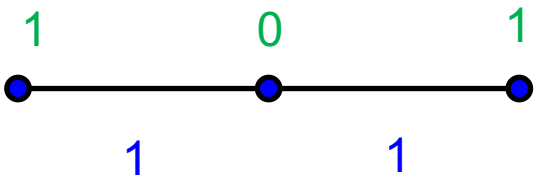
Fractional Vertex Packing

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

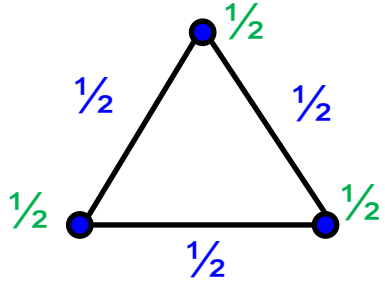
Fact For any v, w: $\sum_x v_x \leq \sum_e w_e$

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

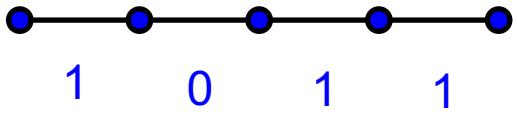
$\rho^* = 1$



$\rho^* = 3/2$



$\rho^* = 3$

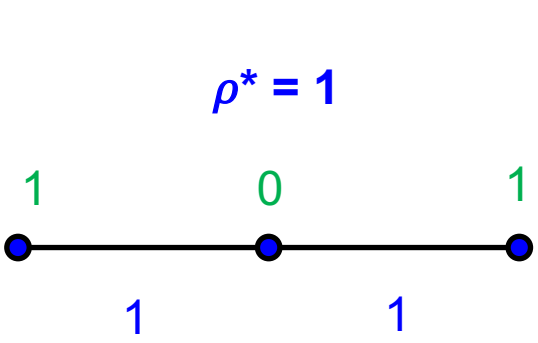


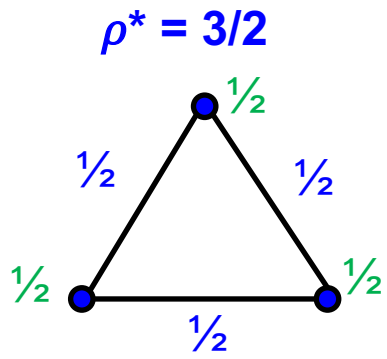
Fractional Vertex Packing

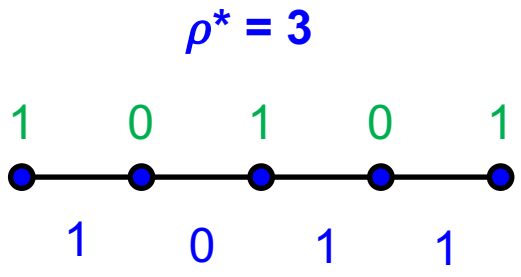
- A *fractional vertex packing* of a (hyper)graph is a set of non-negative numbers v_x, one for each node x, such that, for every edge e: $\sum_{x:\, x \in e} v_x \leq 1$

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
$R_1| \leq N, \ldots, |R_m| \leq N$ 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
$R_1| \leq N, \ldots, |R_m| \leq N$ 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
$R_1| \leq N, \ldots, |R_m| \leq N$ 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:
$(a)|R_j| = N^{v_{i_1} + v_{i_2} + \cdots} \leq N$ (why?)

The Bound is Tight

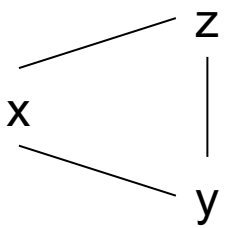
Fact Fix a fractional vertex packing v = $(v_x)_{x \in Nodes}$.
Then there exists a database such that
$R_1| \leq N, \ldots, |R_m| \leq N$ and $|Q| = N^{\Sigma_x v_x}$

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:
$(a)|R_j| = N^{v_{i_1} + v_{i_2} + \cdots} \leq N$ (why?)
$(b)|Q| = N^{\Sigma_x v_x}$(why?)

Examples

Assume all relations have size N

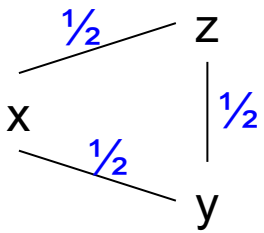
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



Examples

Assume all relations have size N

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

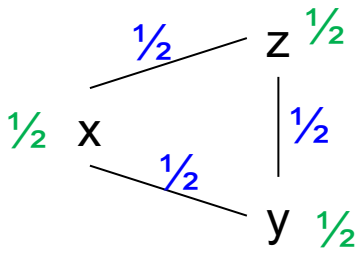


$|Q| \leq N^{3/2}$

Examples

Assume all relations have size N

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



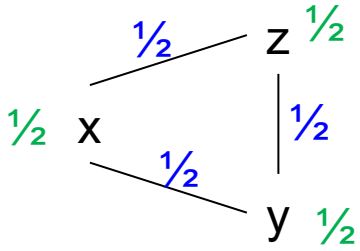
$|Q| \leq N^{3/2}$

Examples

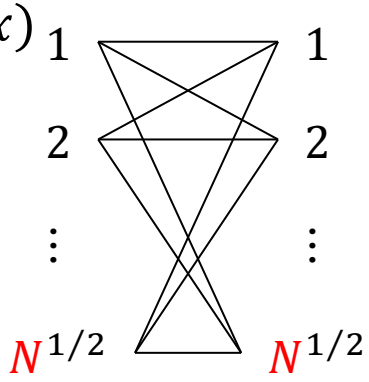
$R =$ $[N^{1/2}] \times [N^{1/2}]$

Assume all relations have size N

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



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

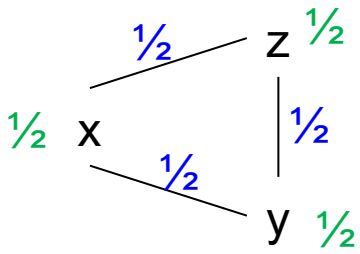


Examples

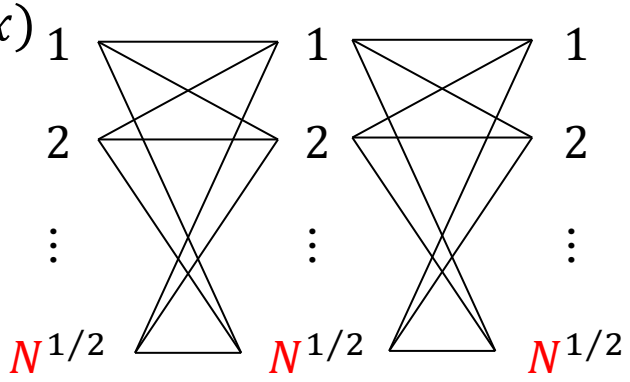
$R = S =$ $[N^{1/2}] \times [N^{1/2}]$

Assume all relations have size N

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



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

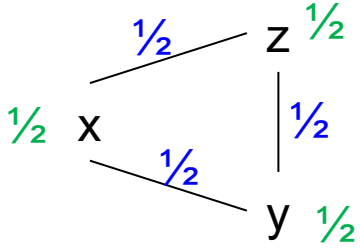


Examples

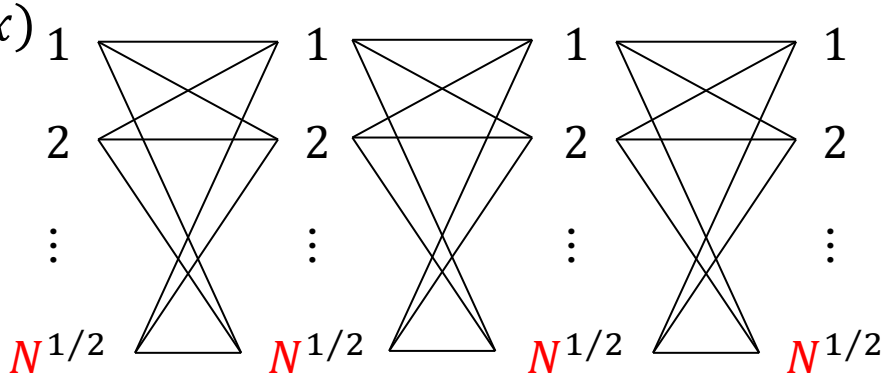
$$R = S = T = [N^{1/2}] \times [N^{1/2}]$$

Assume all relations have size N

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



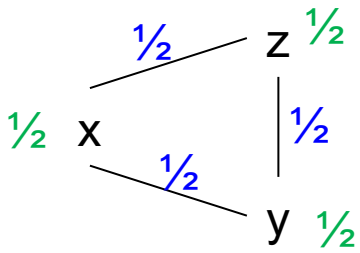
$$|Q| \leq N^{3/2}$$



Examples

Assume all relations have size N

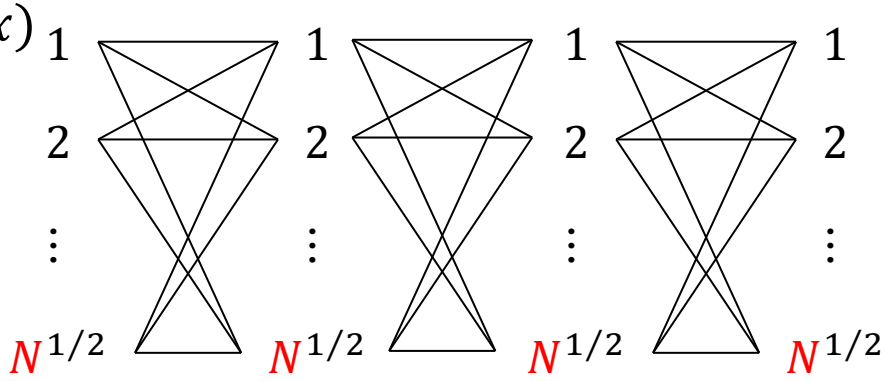
- $Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$



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

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

$|Q| = N^{3/2}$



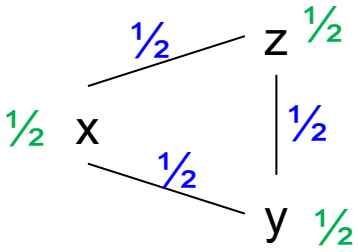
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

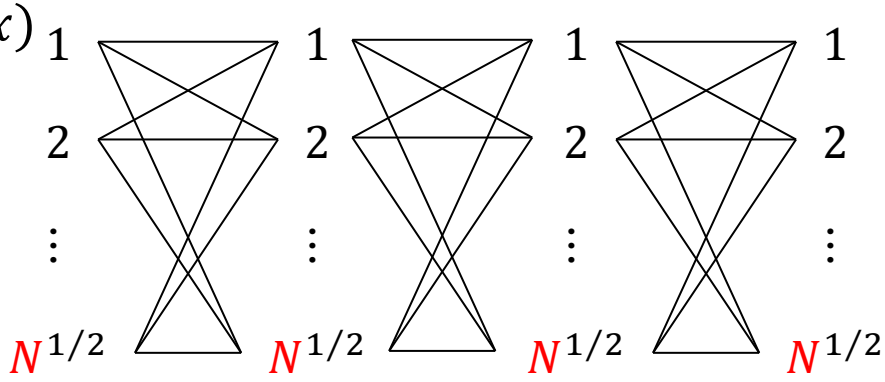
$|Q| = N^{3/2}$

Assume all relations have size N

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



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



- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

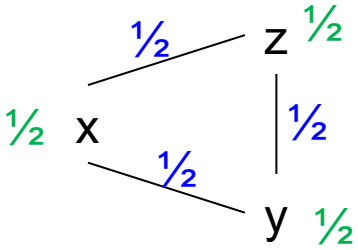
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

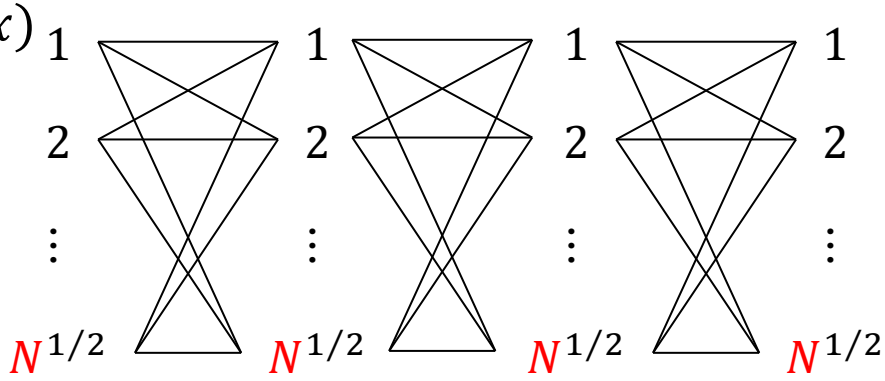
$|Q| = N^{3/2}$

Assume all relations have size N

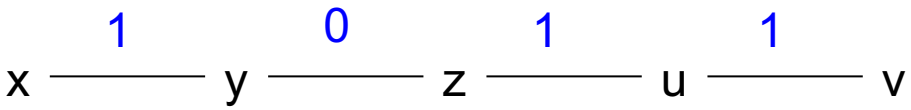
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



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



- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$|Q| \leq N^3$

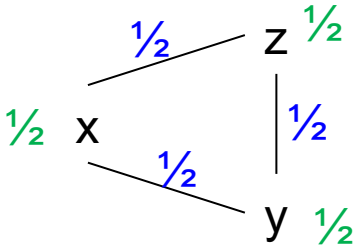
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

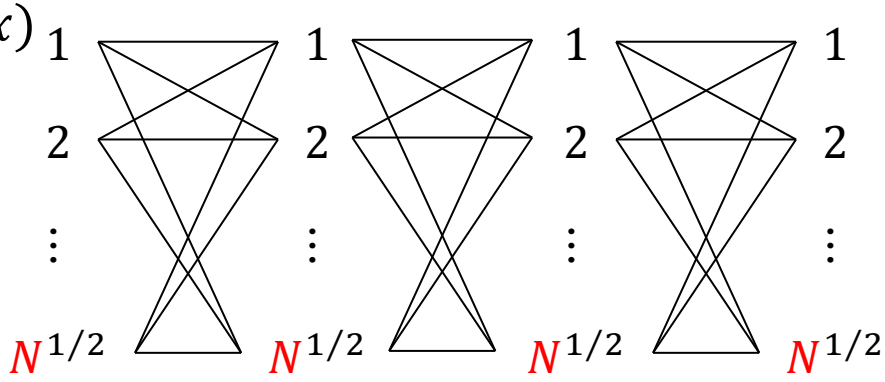
$|Q| = N^{3/2}$

Assume all relations have size N

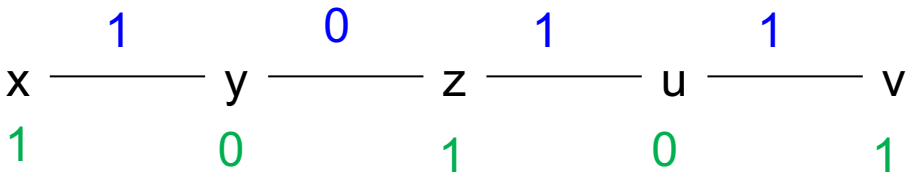
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



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



- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$|Q| \leq N^3$

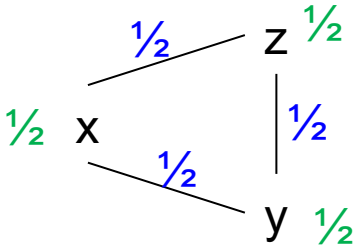
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

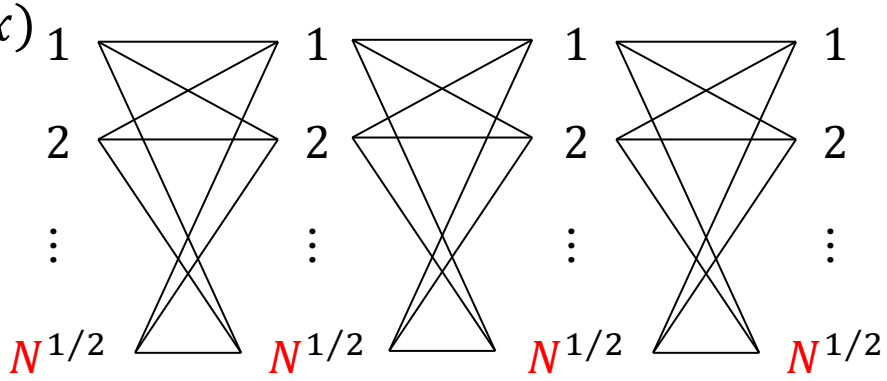
$|Q| = N^{3/2}$

Assume all relations have size N

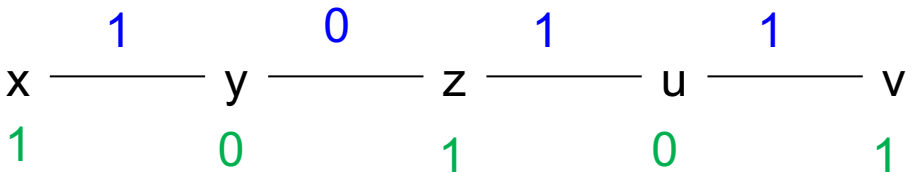
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



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

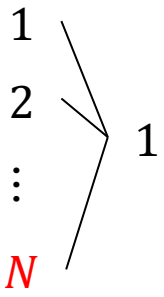


- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$|Q| \leq N^3$

$R = [N] \times [1]$



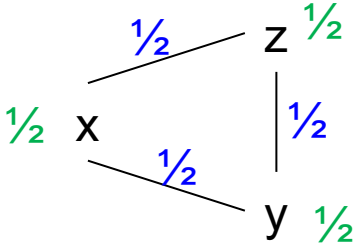
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

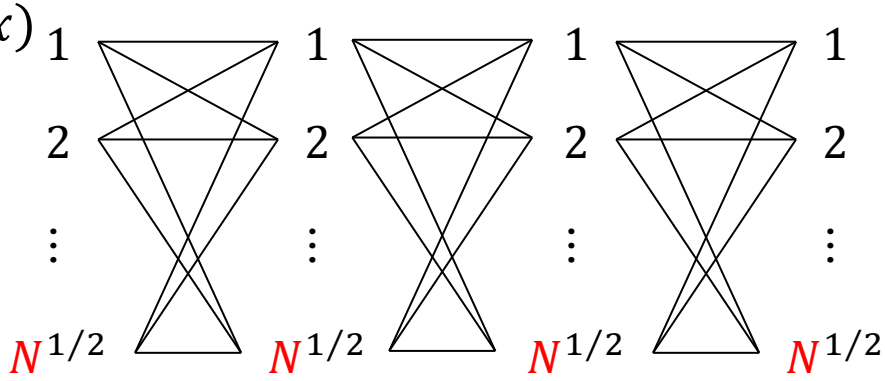
$|Q| = N^{3/2}$

Assume all relations have size N

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

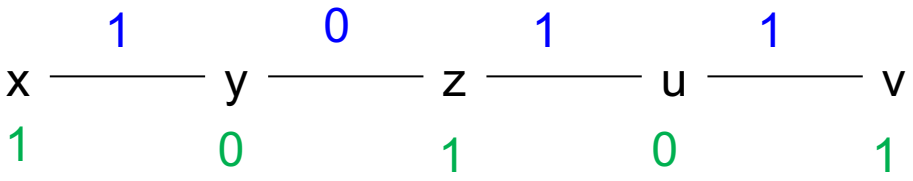


$|Q| \leq N^{3/2}$

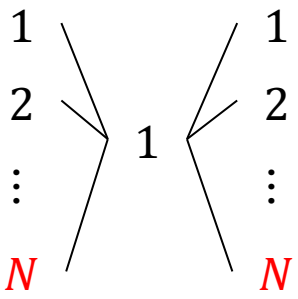


- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

$R = [N] \times [1]$, S= $[1] \times [N]$, …



$|Q| \leq N^3$



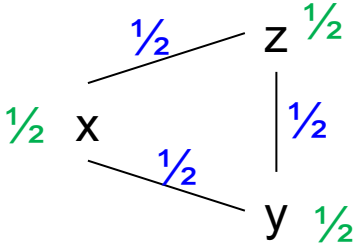
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

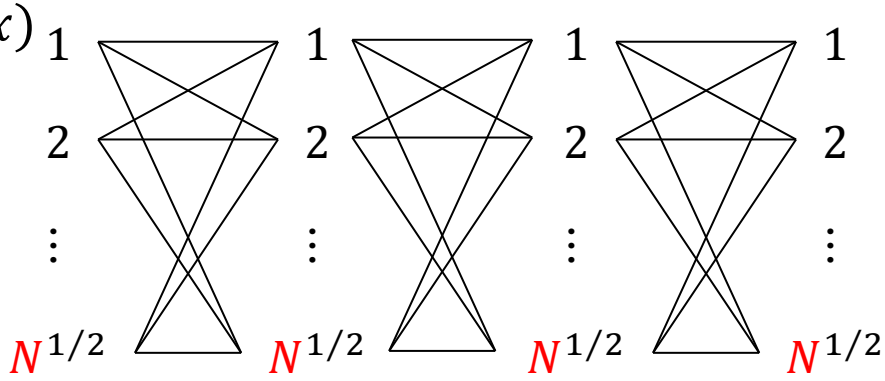
$|Q| = N^{3/2}$

Assume all relations have size N

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

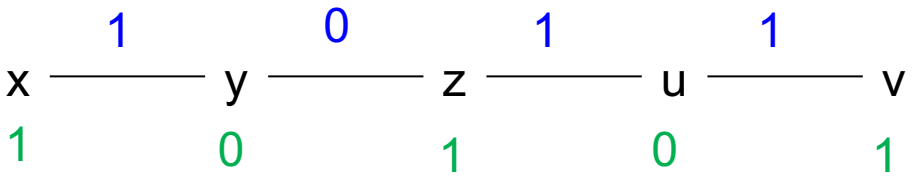


$|Q| \leq N^{3/2}$

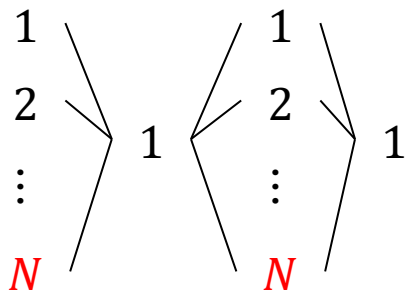


- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$

$R = [N] \times [1]$, S= $[1] \times [N]$, …



$|Q| \leq N^3$



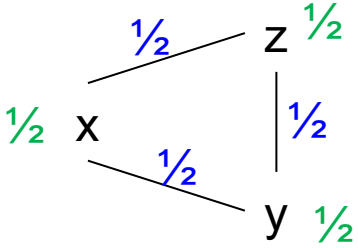
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

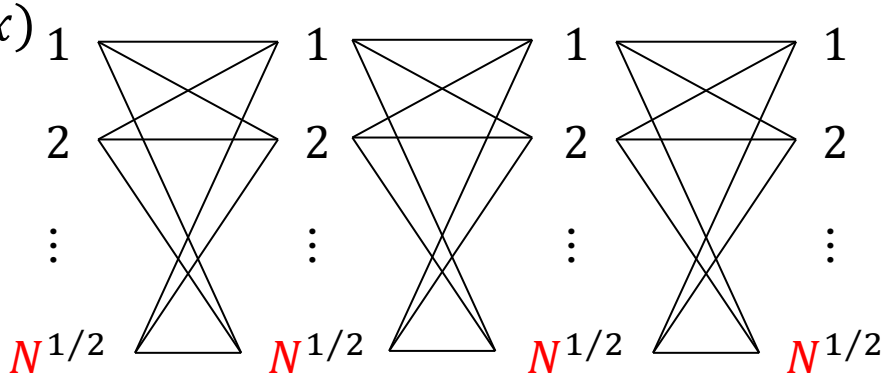
$|Q| = N^{3/2}$

Assume all relations have size N

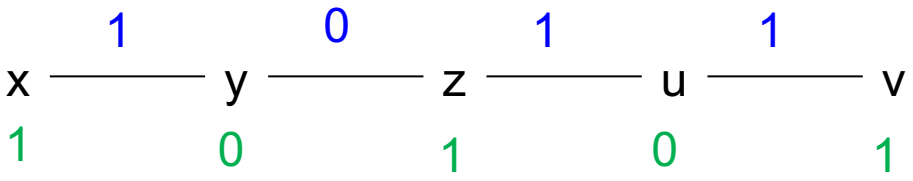
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



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

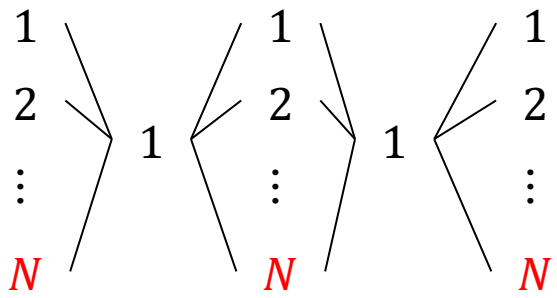


- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$|Q| \leq N^3$

$R = [N] \times [1]$, $S = [1] \times [N]$, ...



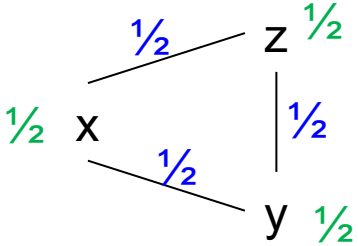
Examples

$R = S = T = [N^{1/2}] \times [N^{1/2}]$

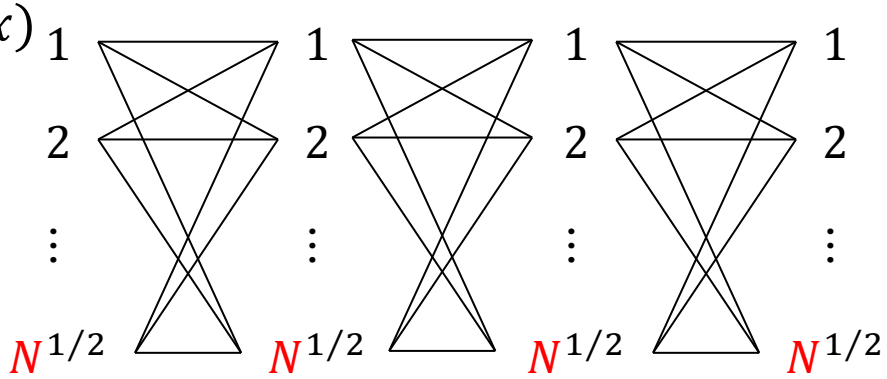
$|Q| = N^{3/2}$

Assume all relations have size N

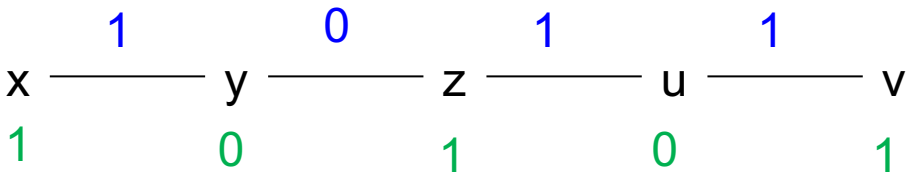
- $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$



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

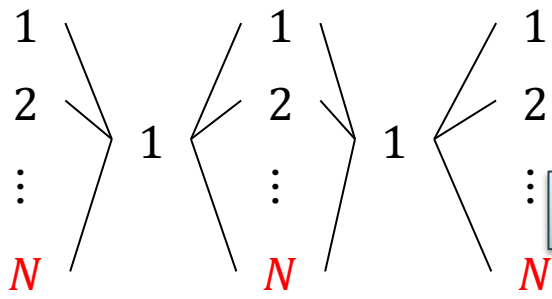


- $Q(x, y, z, u, v) = R(x, y) \wedge S(y, z) \wedge T(z, u) \wedge K(u, v)$



$|Q| \leq N^3$

$R = [N] \times [1]$, $S = [1] \times [N]$, ...



$|Q| = N^3$

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 to all relations containing Y
- Compute AGM($Q^{expanded}$)

Examples

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

Y is a key

Examples

Y is a key

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

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

Examples

Y is a key

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

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

Examples

Y is a key

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

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

Examples

Y is a key

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

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

Y is a key

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

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

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

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

Examples

Y is a key

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

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

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

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