# Data Management Meets Information Theory

Dan Suciu

U. of Washington and RelationalAI

Joint work with M. Abo Khamis, H. Ngo, B. Kenig

# Background

Information theory:

- Routinely used in ML (e.g. decision trees)
- But not in data management
- Recent advances in IT shed deep insight

This talk

- IT in (1) query processing (2) constraints

# Part 1: From Proof to ~~Algorithms~~

Query Plans

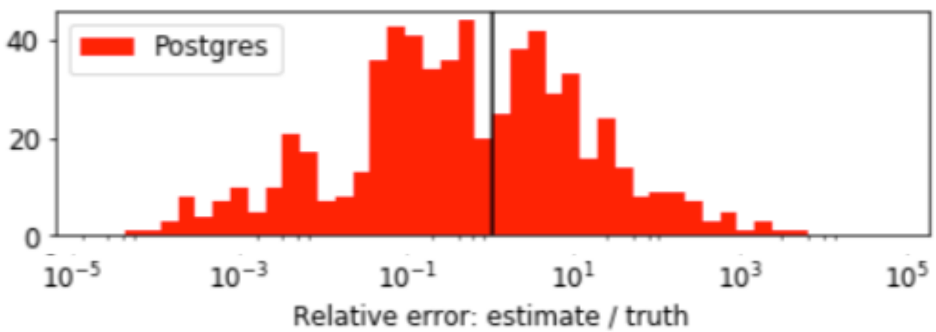
# Query Processing 101

- SQL → Query Plan
- Query Plan → Optimized Query Plan

- Two major problems:
  - Cardinality estimation sucks
  - Optimal plans don't exists

# Example

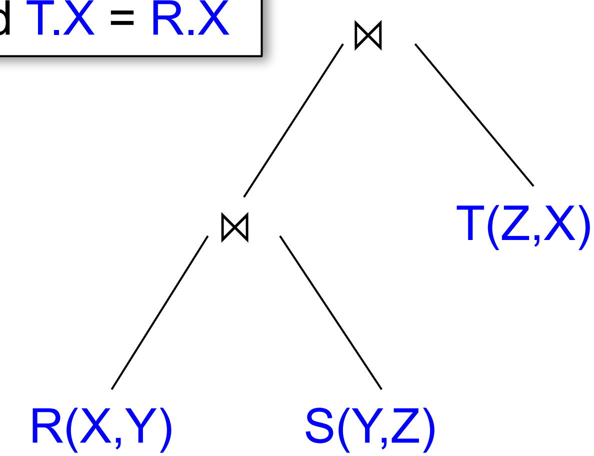
Cardinality estimation sucks




Optimal plans don't exists

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

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




Every query plan $O(N^2)$
Largest output $O(N^{1.5})$

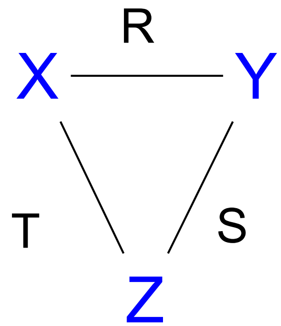
# New Paradigm

- Find information-theoretic *proof* of the upper bound, or the submodular width
- Convert *proof* to *algorithm*

# Two Running Examples

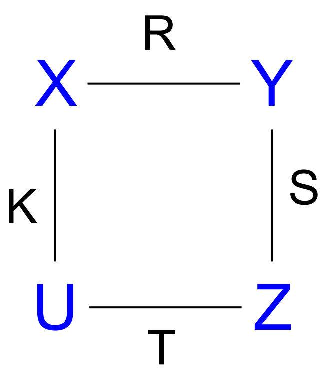
Full Conjunctive Query

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




Boolean Query

$\exists x \exists y \exists z \exists u \; R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,x)$



# Statistics

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

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

# Statistics

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

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

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

# Statistics

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

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:

# Statistics

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

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: $|Q(D)| \leq N$

# Statistics

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

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 $\leq d$:

# Statistics

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

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 $\leq d$: $|Q(D)| \leq d \times N$

# Statistics

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

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

# Statistics

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

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: $|Q(D)| \leq N^{3/2}$

# Entropy

Let $\mathbf{V} = \{X_1, X_2, \ldots\}$ be a set of random variables.

The *entropy* of $\mathbf{X} \subseteq \mathbf{V}$ is

$$H(\mathbf{X}) = - \Sigma_{i=1,N}\, p_i \log p_i$$

$H: 2^{\mathbf{V}} \rightarrow R_+$ is called *entropic*

The *conditional entropy*

$$H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{X} \cup \mathbf{Y}) - H(\mathbf{X})$$

# Shannon Inequalities

Monotonicity

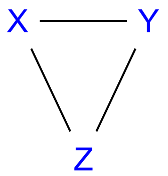
$H(\mathbf{U} \cup \mathbf{V}) \geq H(\mathbf{U})$

$H(\mathbf{U}) + H(\mathbf{V}) \geq H(\mathbf{U} \cap \mathbf{V}) + H(\mathbf{U} \cup \mathbf{V})$

Submodularity

$H: 2^{\mathbf{V}} \rightarrow R_{+}$ is called *polymatroid*

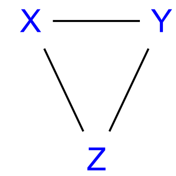
# *Proof* of Upper Bound




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

Database D → entropic function H

# *Proof* of Upper Bound




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

Database D → entropic function H

Database D

R(X,Y)

| X | Y |
|---|---|
| a | 3 |
| a | 2 |
| b | 2 |
| d | 3 |

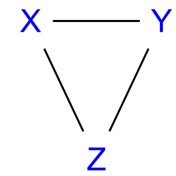
S(Y,Z)

| Y | Z |
|---|---|
| 3 | m |
| 2 | q |
| 3 | q |
| 2 | m |

T(Z,X)

| Z | X |
|---|---|
| m | a |
| q | a |
| q | b |
| m | d |

# *Proof* of Upper Bound




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

Database D → entropic function H

Output Q(D)

| X | Y | Z |
|---|---|---|
| a | 3 | m |
| a | 2 | q |
| b | 2 | q |
| d | 3 | m |
| a | 3 | q |

Database D

R(X,Y)

| X | Y |
|---|---|
| a | 3 |
| a | 2 |
| b | 2 |
| d | 3 |

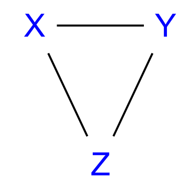
S(Y,Z)

| Y | Z |
|---|---|
| 3 | m |
| 2 | q |
| 3 | q |
| 2 | m |

T(Z,X)

| Z | X |
|---|---|
| m | a |
| q | a |
| q | b |
| m | d |

# *Proof* of Upper Bound




Q(X,Y,Z) = R(X,Y) ∧ S(Y,Z) ∧ T(Z,X)

Database D → entropic function H

Output Q(D)

| X | Y | Z | |
|---|---|---|---|
| a | 3 | m | 1/5 |
| a | 2 | q | 1/5 |
| b | 2 | q | 1/5 |
| d | 3 | m | 1/5 |
| a | 3 | q | 1/5 |

Database D

R(X,Y)

| X | Y |
|---|---|
| a | 3 |
| a | 2 |
| b | 2 |
| d | 3 |

S(Y,Z)

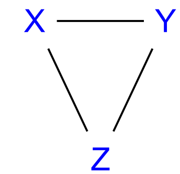
| Y | Z |
|---|---|
| 3 | m |
| 2 | q |
| 3 | q |
| 2 | m |

T(Z,X)

| Z | X |
|---|---|
| m | a |
| q | a |
| q | b |
| m | d |

H(XYZ) = log |Q(D)|

# *Proof* of Upper Bound




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

Database D → entropic function H

Output Q(D)

| X | Y | Z | |
|---|---|---|---|
| a | 3 | m | 1/5 |
| a | 2 | q | 1/5 |
| b | 2 | q | 1/5 |
| d | 3 | m | 1/5 |
| a | 3 | q | 1/5 |

Database D

R(X,Y)

| X | Y | |
|---|---|---|
| a | 3 | 2/5 |
| a | 2 | 1/5 |
| b | 2 | 1/5 |
| d | 3 | 1/5 |

S(Y,Z)

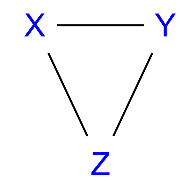
| Y | Z | |
|---|---|---|
| 3 | m | 2/5 |
| 2 | q | 2/5 |
| 3 | q | 1/5 |
| 2 | m | 0 |

T(Z,X)

| Z | X | |
|---|---|---|
| m | a | 1/5 |
| q | a | 2/5 |
| q | b | 1/5 |
| m | d | 1/5 |

$H(XYZ) = \log |Q(D)|$

# *Proof* of Upper Bound




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

Database D → entropic function H

Output Q(D)

| X | Y | Z | |
|---|---|---|---|
| a | 3 | m | 1/5 |
| a | 2 | q | 1/5 |
| b | 2 | q | 1/5 |
| d | 3 | m | 1/5 |
| a | 3 | q | 1/5 |

$H(XYZ) = \log |Q(D)|$

Database D

R(X,Y)

| X | Y | |
|---|---|---|
| a | 3 | 2/5 |
| a | 2 | 1/5 |
| b | 2 | 1/5 |
| d | 3 | 1/5 |

S(Y,Z)

| Y | Z | |
|---|---|---|
| 3 | m | 2/5 |
| 2 | q | 2/5 |
| 3 | q | 1/5 |
| 2 | m | 0 |

T(Z,X)

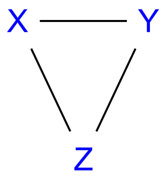
| Z | X | |
|---|---|---|
| m | a | 1/5 |
| q | a | 2/5 |
| q | b | 1/5 |
| m | d | 1/5 |

$H(XY) \le \log N_R$ $H(YZ) \le \log N_S$ $H(XZ) \le \log N_T$

$H(Z|Y) \le \log deg_S(z|y)$

Cardinalitites, functional dependences, max degrees

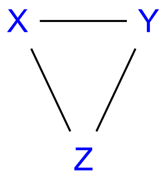
# *Proof* of Upper Bound




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

$|R|,|S|,|T| \leq N$ → $|Q(D)| \leq N^{3/2}$

# *Proof* of Upper Bound

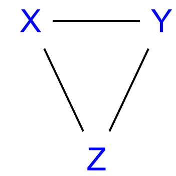



$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

$|R|,|S|,|T| \leq N \quad \rightarrow \quad |Q(D)| \leq N^{3/2}$

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

# *Proof* of Upper Bound



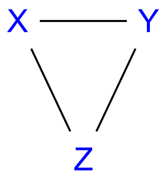

$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

$|R|,|S|,|T| \leq N \quad \rightarrow \quad |Q(D)| \leq N^{3/2}$

submodularity

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

# *Proof* of Upper Bound




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

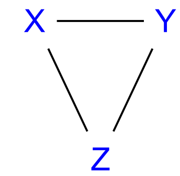
$|R|,|S|,|T| \leq N$ $\rightarrow$ $|Q(D)| \leq N^{3/2}$

submodularity

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

$\geq h(XYZ) + h(Y) + h(XZ)$

# *Proof* of Upper Bound




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

$|R|,|S|,|T| \leq N \rightarrow |Q(D)| \leq N^{3/2}$

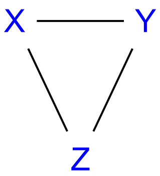
submodularity

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

submodularity

$\geq h(XYZ) + h(Y) + h(XZ)$

# *Proof* of Upper Bound




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

$|R|,|S|,|T| \leq N \rightarrow |Q(D)| \leq N^{3/2}$

submodularity

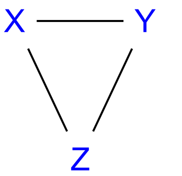
$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

submodularity

$\geq h(XYZ) + h(Y) + h(XZ)$

$\geq h(XYZ) + h(XYZ) + h(\varnothing)$

# *Proof* of Upper Bound




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

$|R|,|S|,|T| \leq N \quad \rightarrow \quad |Q(D)| \leq N^{3/2}$

submodularity

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

submodularity

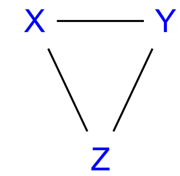
$\geq h(XYZ) + h(Y) + h(XZ)$

$\geq h(XYZ) + h(XYZ) + h(\emptyset)$

$= 2\, h(XYZ)$

$= 2 \log |Q(D)|$

# *Proof* of Upper Bound




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

$|R|,|S|,|T| \leq N$ → $|Q(D)| \leq N^{3/2}$

submodularity

$3 \log N \geq h(XY) + h(YZ) + h(XZ)$

submodularity

$\geq h(XYZ) + h(Y) + h(XZ)$

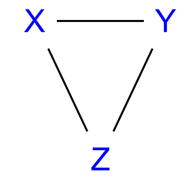
$\geq h(XYZ) + h(XYZ) + h(\varnothing)$

$= 2\, h(XYZ)$

Shearer's inequality

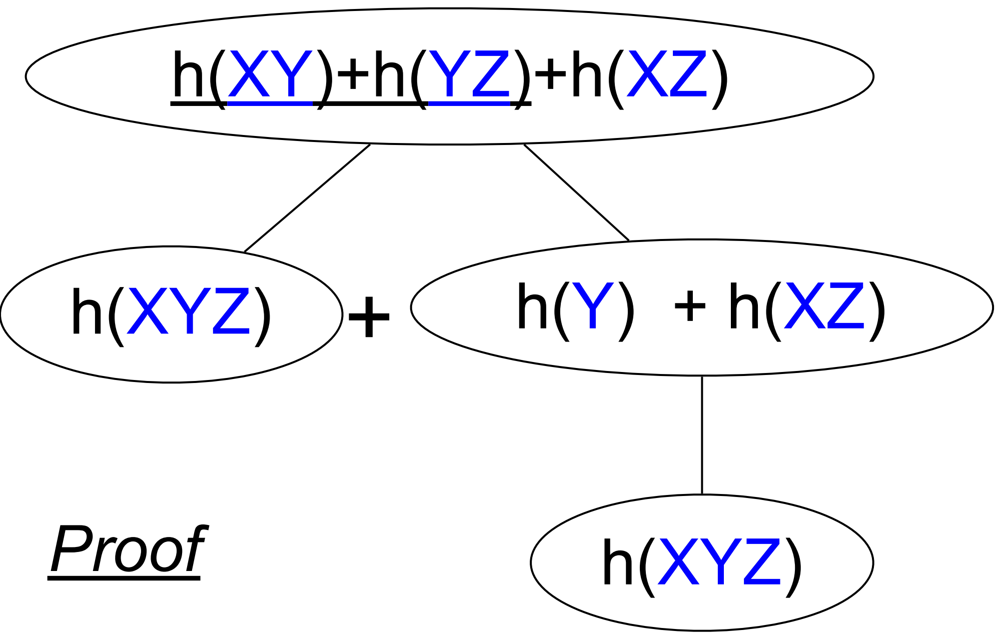
$= 2 \log |Q(D)|$

# *Proof* to *Algorithm*



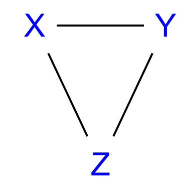

Q(X,Y,Z) = R(X,Y) ∧ S(Y,Z) ∧ T(Z,X)

h(XY) + h(YZ) + h(XZ) ≥ 2 h(XYZ)




*Proof*

# *Proof* to *Algorithm*

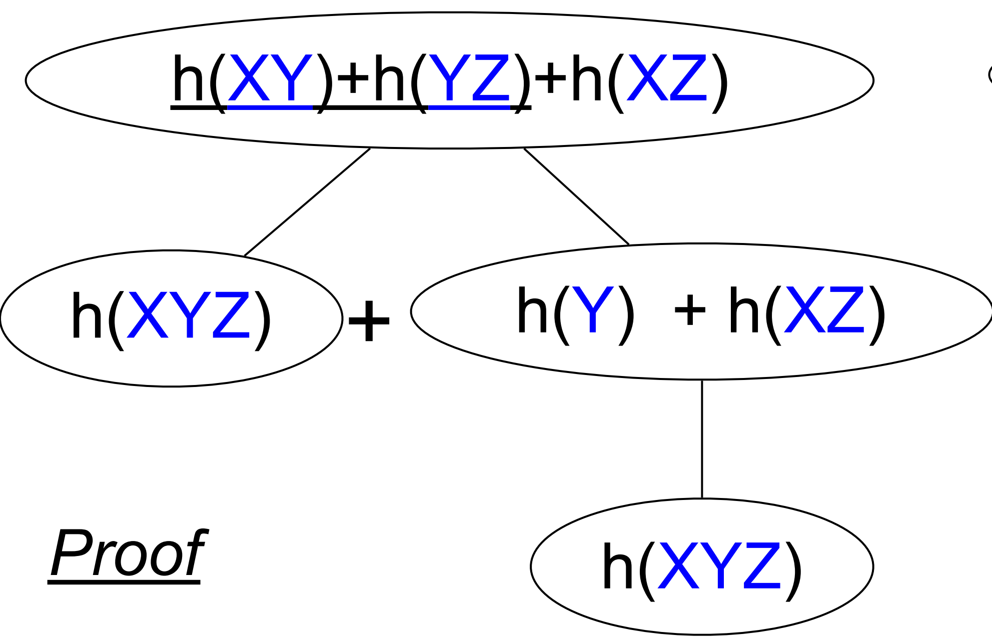



$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

$h(XY) + h(YZ) + h(XZ) \geq 2\, h(XYZ)$

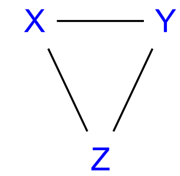
*Algorithm*

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




*Proof*

# *Proof* to *Algorithm*




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

$h(XY) + h(YZ) + h(XZ) \geq 2\, h(XYZ)$

*Algorithm*

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

$N^{3/2}$

$R_{light}(X,Y) \wedge S(Y,Z)$

$\underline{h(XY)+h(YZ)}+h(XZ)$

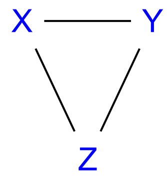
$h(XYZ)$ + $h(Y) + h(XZ)$

$h(XYZ)$

*Proof*

$R_{light}$ or $R_{heavy}$: degree(Y) ≤ or > $N^{1/2}$

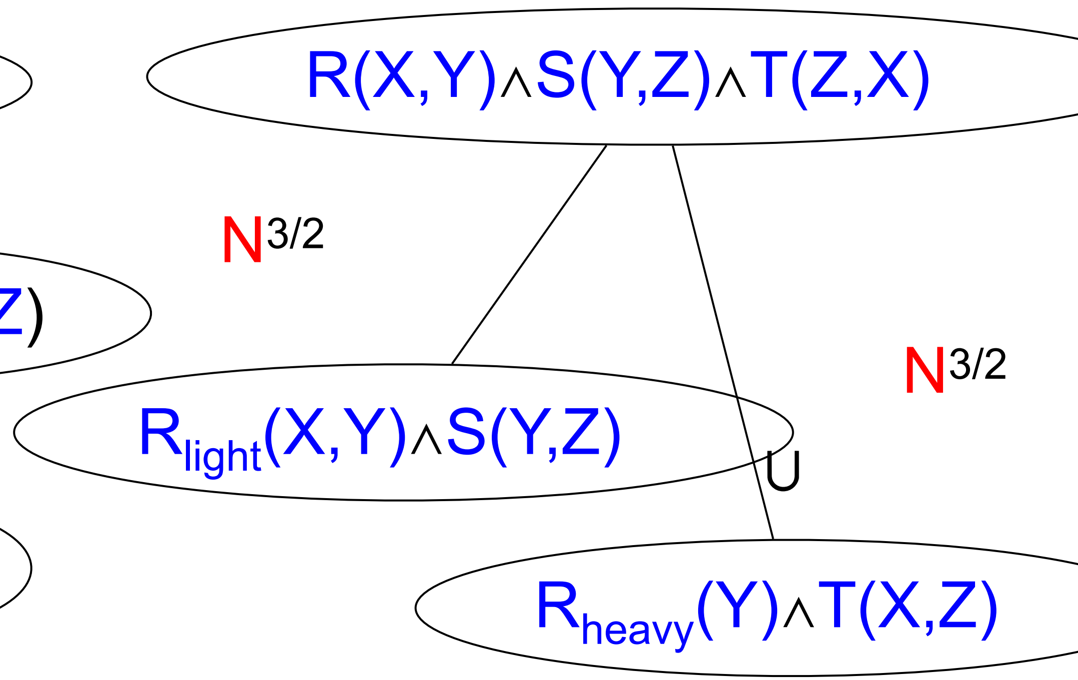
# *Proof* to *Algorithm*

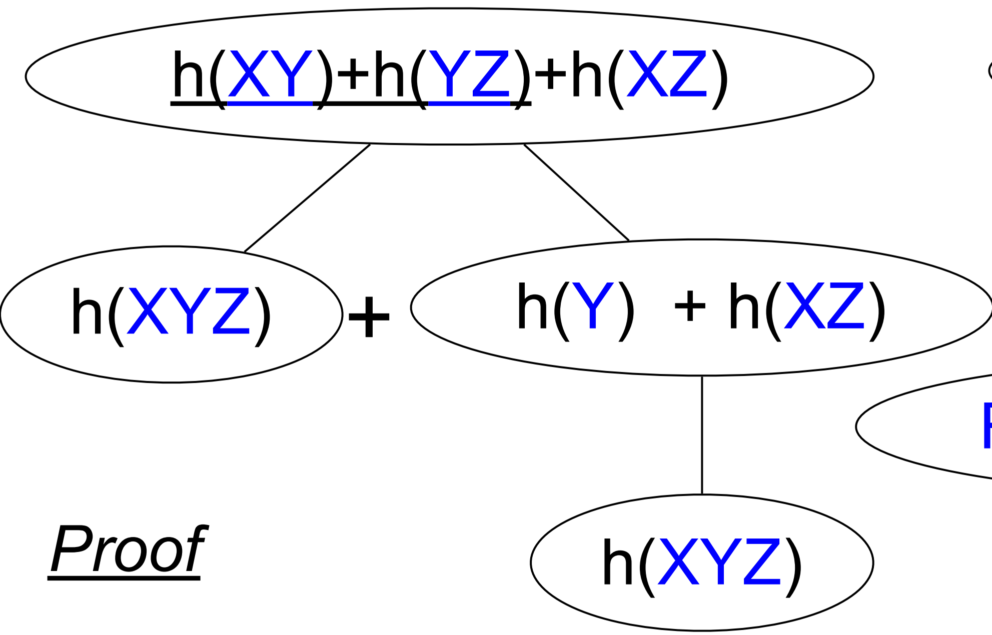



Q(X,Y,Z) = R(X,Y) ∧ S(Y,Z) ∧ T(Z,X)

h(XY) + h(YZ) + h(XZ) ≥ 2 h(XYZ)

*Algorithm*

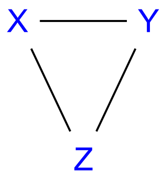



*Proof*

$R_{light}$ or $R_{heavy}$: degree(Y) ≤ or > $N^{1/2}$

# *Proof* to *Algorithm*




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

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

$h(XY) + h(YZ) + h(XZ) \geq 2\, h(XYZ)$

*Proof*

$h(XY)+h(YZ)+h(XZ)$

$h(XYZ)$ + $h(Y) + h(XZ)$

$h(XYZ)$

*Algorithm*

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

$N^{3/2}$

$R_{light}(X,Y) \wedge S(Y,Z)$

$\cup$

$N^{3/2}$

$R_{heavy}(Y) \wedge T(X,Z)$

$R_{light}$ or $R_{heavy}$: degree(Y) ≤ or > $N^{1/2}$

# Full Conjunctive Query

Asymptotically tight, but open if computable

**Theorem** $\forall$D that satisfies the statistics

$$\log |Q(D)| \le \max_{H \text{ entropic satisfying stats}} H(\mathbf{X})$$

$$\le \max_{H \text{ polymatroid satisfying stats}} H(\mathbf{X})$$

Computable in EXPTIME, but not tight

**Thm** Q(D) computable in time Õ(Polymatroid-bound)

# Discussion

AGM Bound [Atserias,Grohe,Marx'08, Ngo,Re,Rudra'13]

- Entropic bound = polymatroid bound
- Algorithm (NPRR) for Q(D) has single log factor

Cardinalities + FDs + max degrees [AboKhamis,Ngo,S'17]

- Entropic bound $\lneqq$ polymatroid bound
- Algorithm (PANDA) for Q(D) has polylog factor

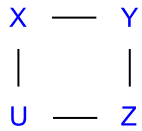
# Boolean Query

Tree decomposition (TD) = a tree where each node t is a full conjunctive query

- Fractional hypetree width [Grohe,Marx'14]
  $\min_{\text{tree}} \max_{\text{node } t} \max_{D}$

- Submodular width [Marx'13,ANS'17]
  $\max_{D} \min_{\text{tree}} \max_{\text{node } t}$

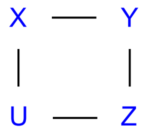
$Q() = \exists x \exists y \exists z \exists u\ R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,x)$




$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

$\min_{\text{tree}} \max_{\text{node } t} \max_{D}$

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

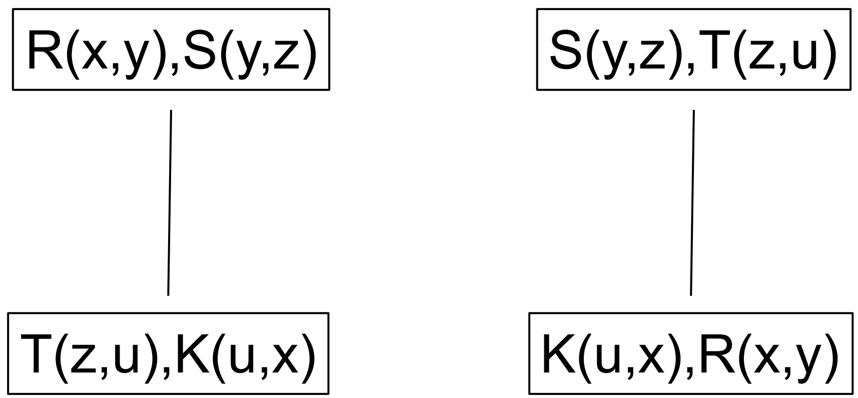



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

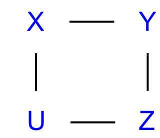
---

$\min_{tree} \max_{node\ t} \max_{D}$

Tree decompositions



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

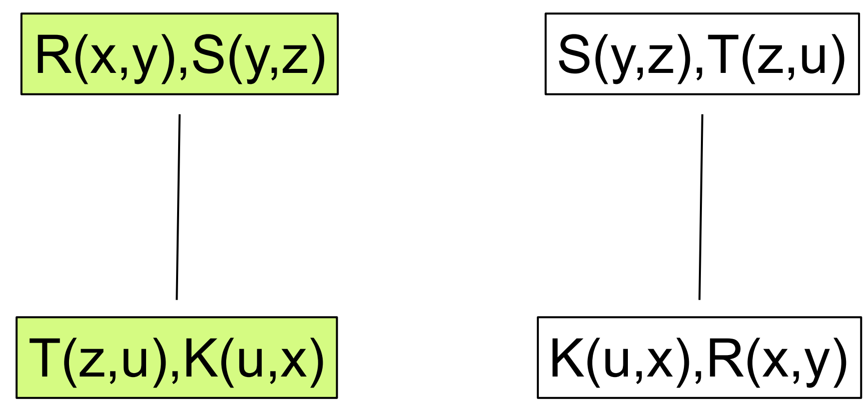



$O(N^{3/2})$ algorithm [Alon, Yuster, Zwick'97]

---

$\min_{tree} \max_{node\ t} \max_{D}$

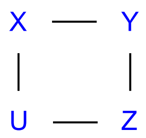
Tree decompositions




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

(suboptimal)

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

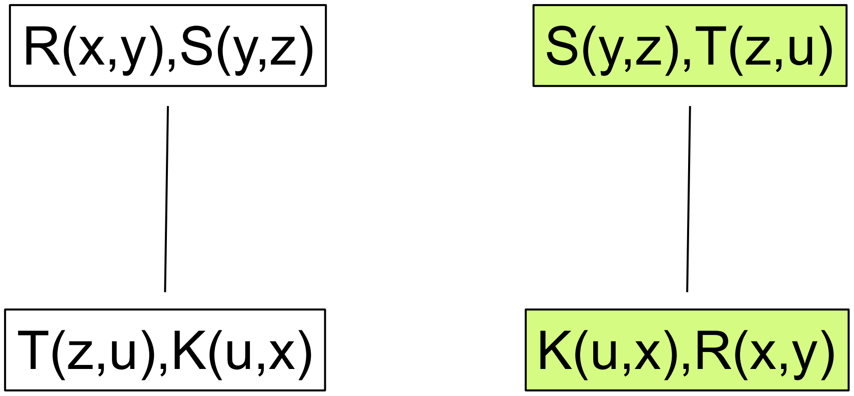



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\min_{tree} \max_{node\ t} \max_{D}$

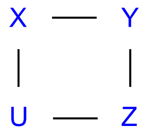
Tree decompositions




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

(suboptimal)

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

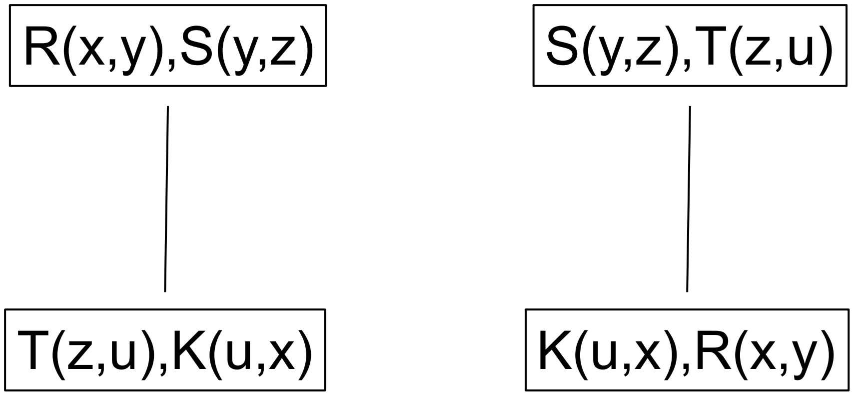



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

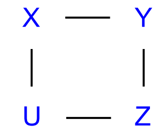
---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions



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



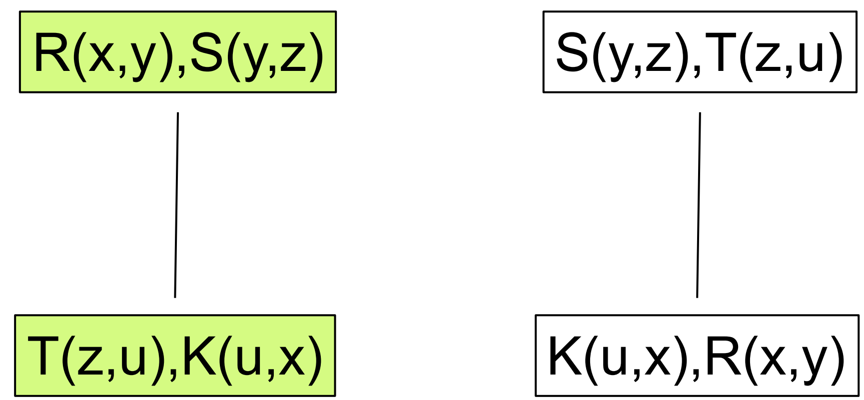

$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

$\max_D \min_{tree} \max_{node\ t}$

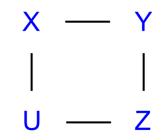
Tree decompositions

T1

max(h(xyz),h(zux))



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




$O(N^{3/2})$ algorithm [Alon, Yuster, Zwick'97]

---

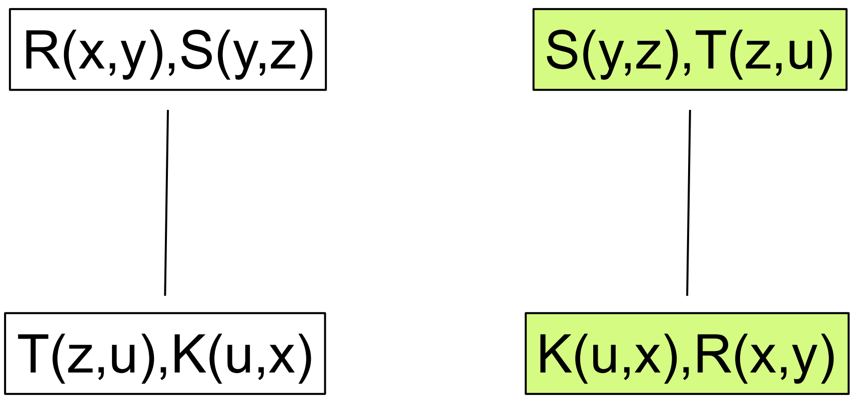
$\max_D \min_{tree} \max_{node\ t}$

T1

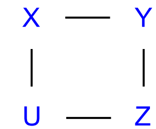
max(h(xyz),h(zux)),
max(h(yzu),h(uxy))

T2

Tree decompositions



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

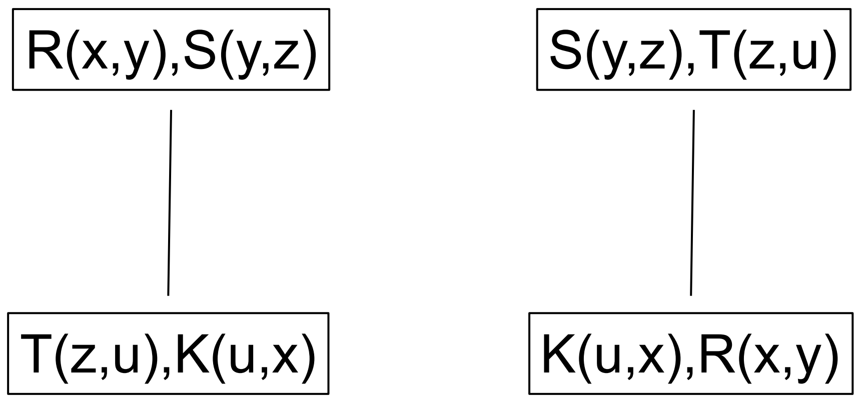



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

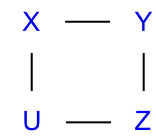



T1

min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

T2

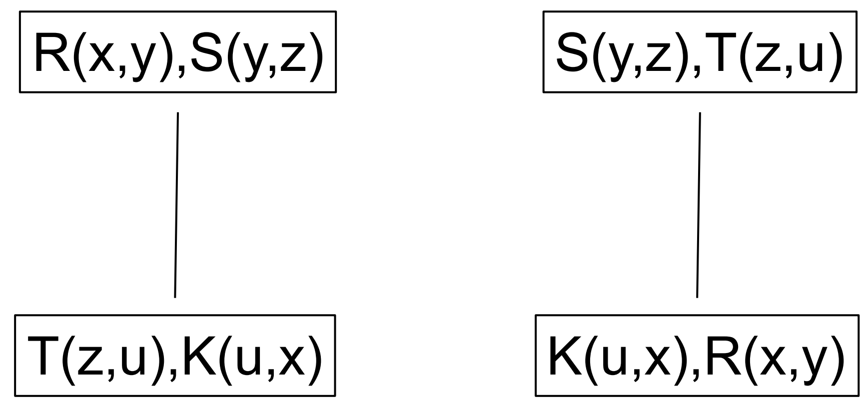
$Q() = \exists x \exists y \exists z \exists u\ R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,x)$




$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions



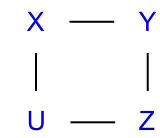

T1

min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

T2

= max(min(h(xyz),h(yzu)),
min(h(xyz),h(uxy)),
min(h(zux),h(yzu)),
min(h(zux),h(uxy)))

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

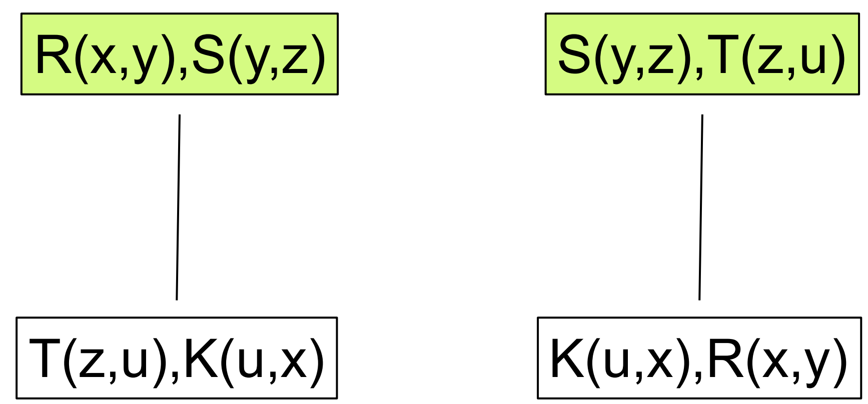



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




T1

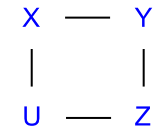
min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

T2

$3 \log N \geq h(xy) + h(yz) + h(zu)$

= max(**min(h(xyz),h(yzu))**,
min(h(xyz),h(uxy)),
min(h(zux),h(yzu)),
min(h(zux),h(uxy)))

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

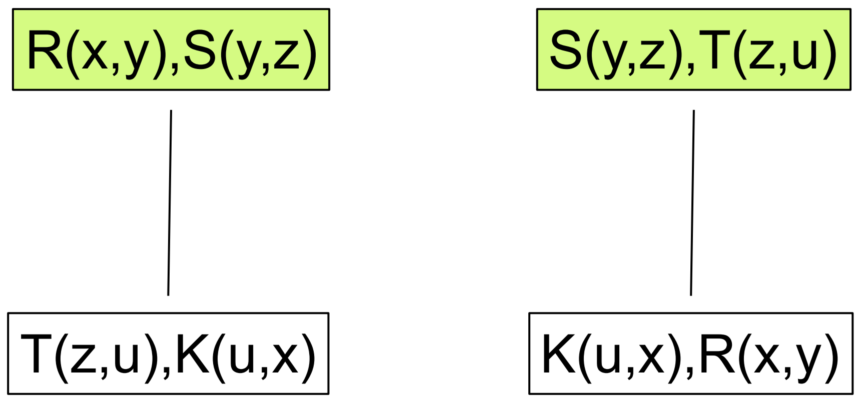



$O(N^{3/2})$ algorithm [Alon, Yuster, Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




T1

min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

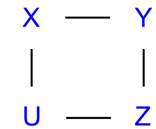
T2

= max(**min(h(xyz),h(yzu))**,
min(h(xyz),h(uxy)),
min(h(zux),h(yzu)),
min(h(zux),h(uxy)))

$3 \log N \geq \underline{h(xy) + h(yz)} + h(zu)$

$\geq h(xyz) + h(y) + h(zu)$

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

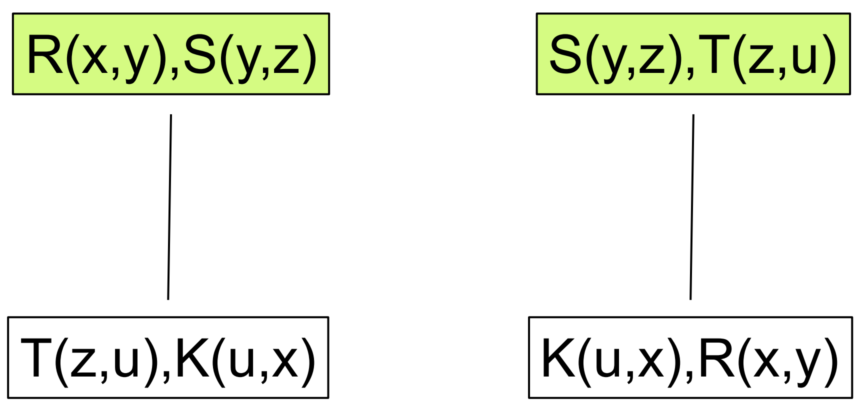



$O(N^{3/2})$ algorithm [Alon, Yuster, Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




T1

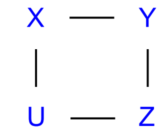
$\min(\max(h(xyz),h(zux)),$
$\max(h(yzu),h(uxy))) =$

T2

$= \max(\mathbf{\min(h(xyz),h(yzu))},$
$\min(h(xyz),h(uxy)),$
$\min(h(zux),h(yzu)),$
$\min(h(zux),h(uxy)))$

$3 \log N \geq h(xy) + h(yz) + h(zu)$
$\geq h(xyz) + h(y) + h(zu)$
$\geq h(xyz) + h(yzu)$

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

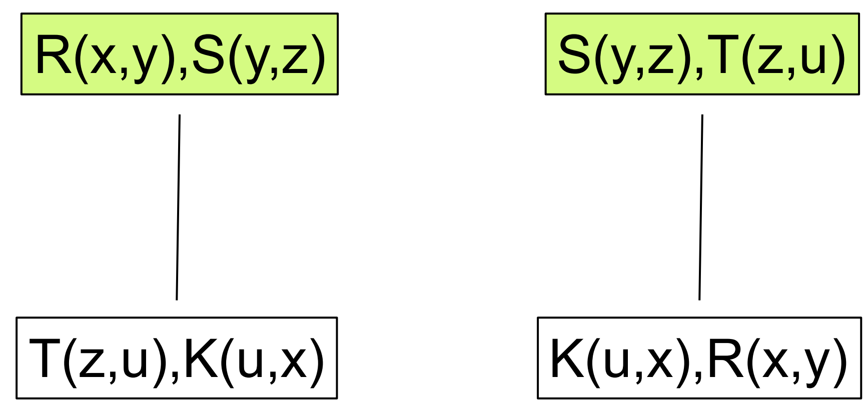



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




T1

min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

T2

= max(**min(h(xyz),h(yzu))**,
min(h(xyz),h(uxy)),
min(h(zux),h(yzu)),
min(h(zux),h(uxy)))

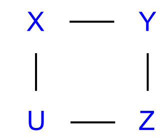
$3 \log N \geq \underline{h(xy) + h(yz)} + h(zu)$

$\geq h(xyz) + \underline{h(y) + h(zu)}$

$\geq h(xyz) + h(yzu)$

$\geq 2 \min(h(xyz),h(yzu))$

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

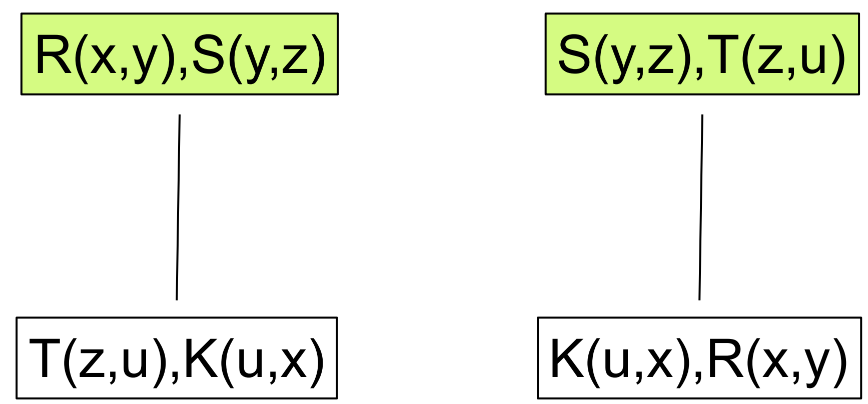



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




subw(Q) = 3/2 log N

T1

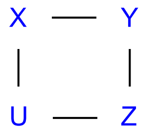
min( max(h(xyz),h(zux)),
max(h(yzu),h(uxy))) =

T2

= max(**min(h(xyz),h(yzu)**),
min(h(xyz),h(uxy)),
min(h(zux),h(yzu)),
min(h(zux),h(uxy)))

$3 \log N \geq h(xy) + h(yz) + h(zu)$
$\geq h(xyz) + h(y) + h(zu)$
$\geq h(xyz) + h(yzu)$
$\geq 2 \min(h(xyz), h(yzu))$

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

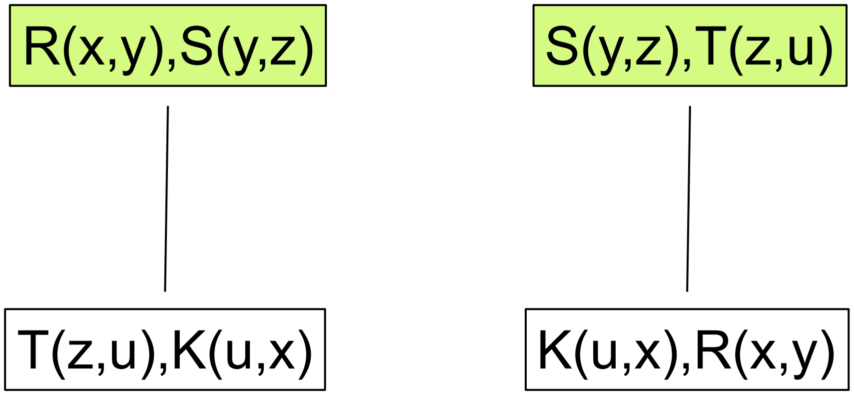



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




subw(Q) = 3/2 log N

$3 \log N \geq h(xy) + h(yz) + h(zu)$

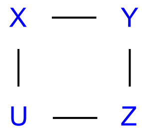
$\geq h(xyz) + h(y) + h(zu)$

$\geq h(xyz) + h(yzu)$

$\geq 2 \min(h(xyz), h(yzu))$

Next: ***proof*** to ***algorithm***

Q() = ∃x∃y∃z∃u R(x,y)∧S(y,z)∧T(z,u)∧K(u,x)

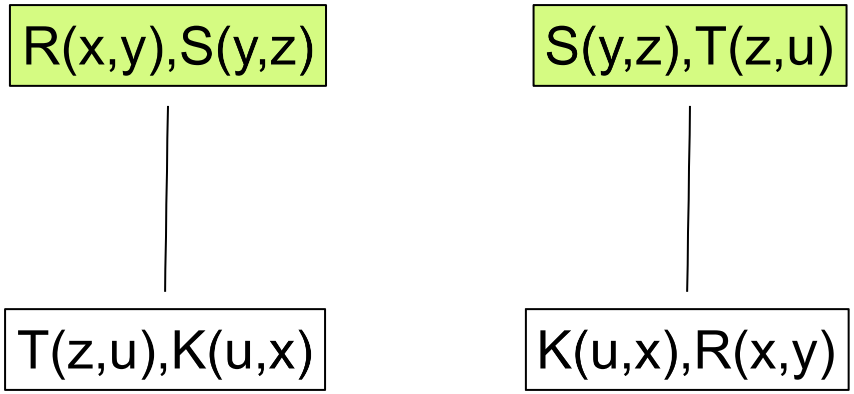



O($N^{3/2}$) algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




**subw(Q) = 3/2 log N**

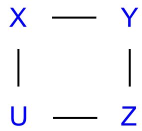
3 log N ≥ h(xy) + h(yz) + h(zu)

≥ h(xyz) + h(y) + h(zu)

≥ h(xyz) + h(yzu)

≥ 2 min(h(xyz),h(yzu))

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



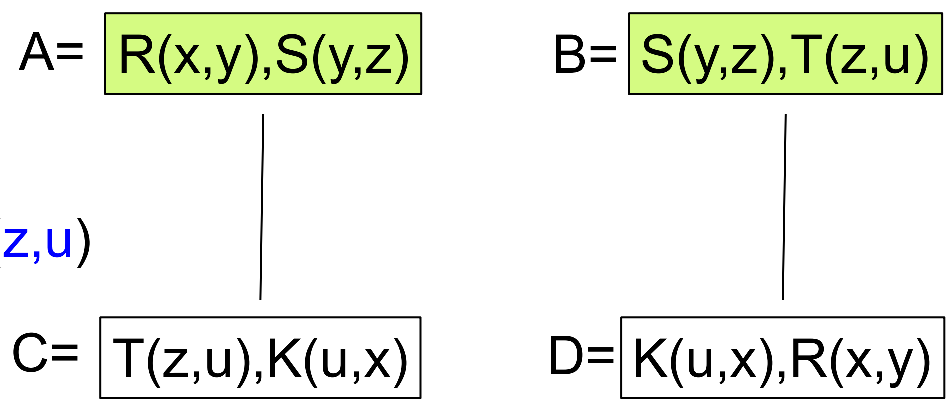

$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

subw(Q) = 3/2 log N




$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

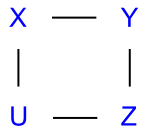
$3 \log N \geq h(xy) + h(yz) + h(zu)$

$\geq h(xyz) + \underline{h(y) + h(zu)}$

$\geq h(xyz) + h(yzu)$

$\geq 2 \min(h(xyz), h(yzu))$

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

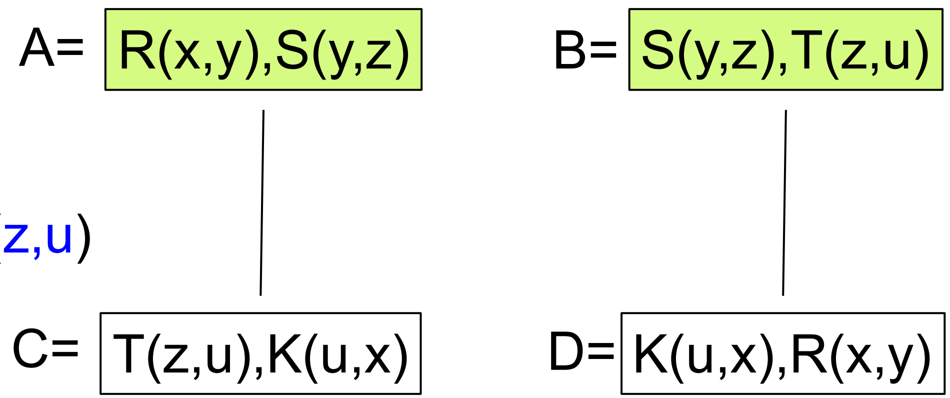



$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions




**subw(Q) = 3/2 log N**

$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

Partition $S$ into $S_{light} \cup S_{heavy}$

$A(x,y,z) \leftarrow R(x,y) \bowtie S_{light}(y,z)$

$3 \log N \geq h(xy) + h(yz) + h(zu)$

$\geq h(xyz) + \underline{h(y) + h(zu)}$

$\geq h(xyz) + h(yzu)$

$\geq 2 \min(h(xyz), h(yzu))$

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

x — y
| |
u — z

$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

$subw(Q) = 3/2 \log N$

A= R(x,y),S(y,z) — C= T(z,u),K(u,x)

B= S(y,z),T(z,u) — D= K(u,x),R(x,y)

$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

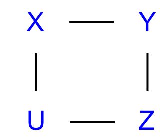
Partition S into $S_{light} \cup S_{heavy}$

$A(x,y,z) \leftarrow R(x,y) \bowtie S_{light}(y,z)$

$B(y,z,u) \leftarrow S_{heavy}(y) \bowtie T(z,u)$

$$3 \log N \geq h(xy) + h(yz) + h(zu)$$
$$\geq h(xyz) + \underline{h(y) + h(zu)}$$
$$\geq h(xyz) + h(yzu)$$
$$\geq 2 \min(h(xyz), h(yzu))$$

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



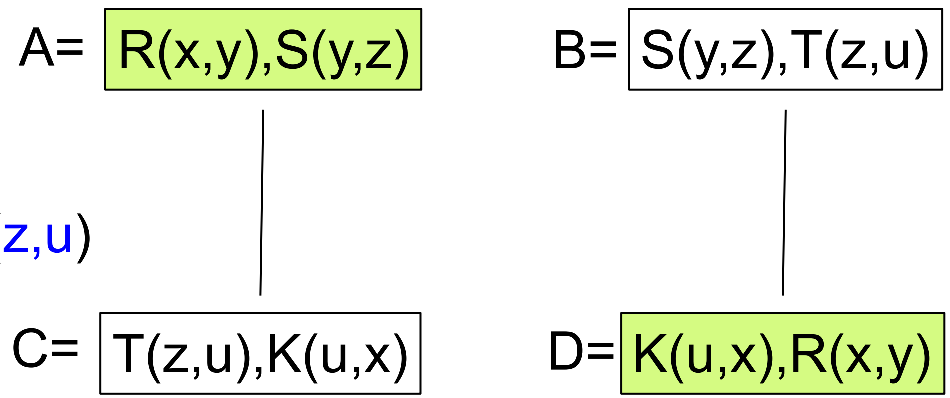

$O(N^{3/2})$ algorithm [Alon, Yuster, Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

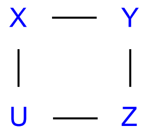
subw(Q) = 3/2 log N




$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

$A(x,y,z) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

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




$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

$subw(Q) = 3/2 \log N$

Tree decompositions

A= R(x,y),S(y,z)

C= T(z,u),K(u,x)

B= S(y,z),T(z,u)

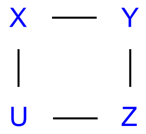
D= K(u,x),R(x,y)

$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

$A(x,y,z) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

$C(x,z,u) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

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




$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

**$subw(Q) = 3/2 \log N$**

A= R(x,y),S(y,z)

C= T(z,u),K(u,x)

B= S(y,z),T(z,u)

D= K(u,x),R(x,y)

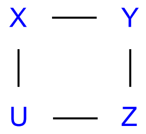
$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

$A(x,y,z) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

$C(x,z,u) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

$C(x,z,u) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

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



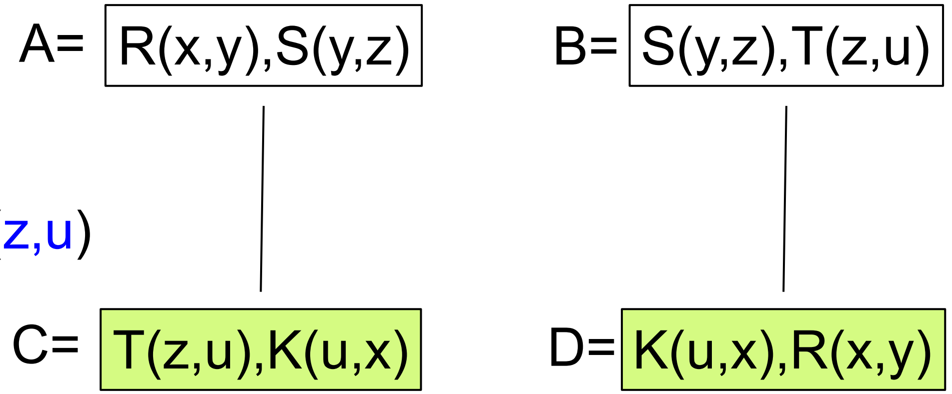

$O(N^{3/2})$ algorithm [Alon,Yuster,Zwick'97]

---

$\max_D \min_{tree} \max_{node\ t}$

Tree decompositions

$subw(Q) = 3/2 \log N$




$A(x,y,z) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

$A(x,y,z) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

$C(x,z,u) \vee B(y,z,u) \leftarrow R(x,y) \wedge S(y,z) \wedge T(z,u)$

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

$C(x,z,u) \vee D(x,y,u) \leftarrow R(x,y) \wedge S(y,z) \wedge K(u,x)$

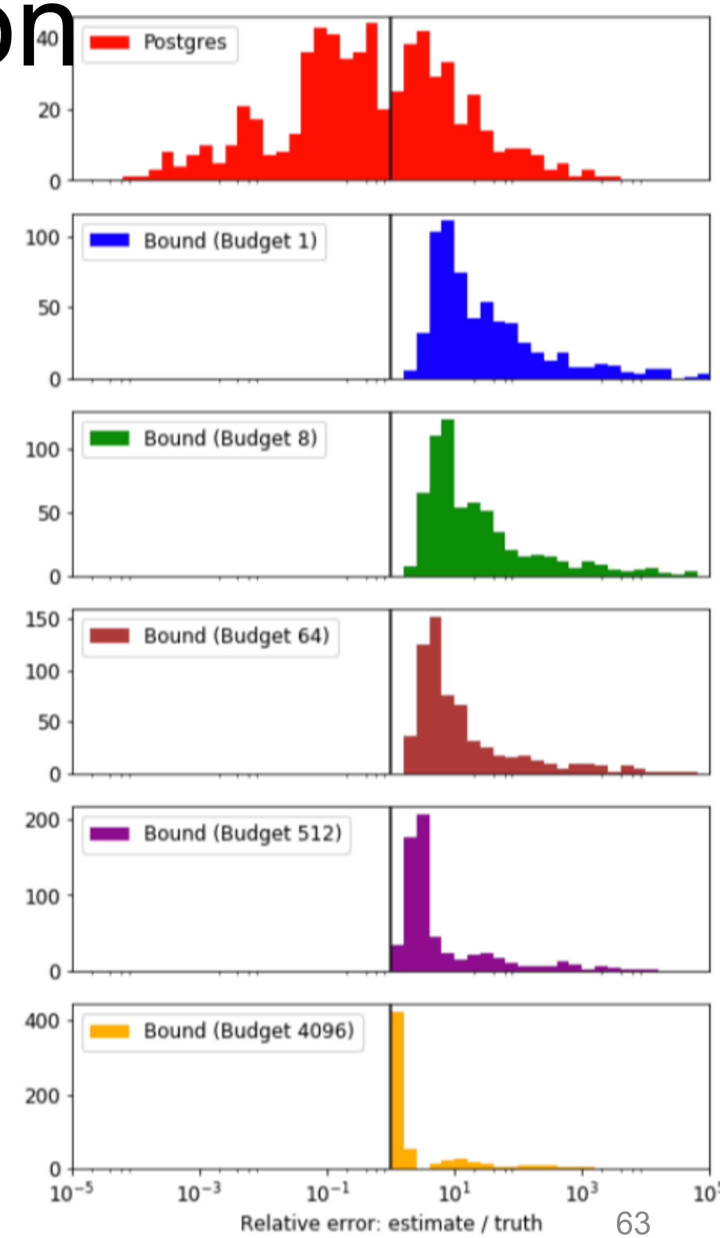
# Discussion

Query evaluation summary:

- *Proof* → *Algorithm*
- Cardinality estimation

Open problem:

- Better proofs → better algorithms



# Part 2: Relaxing Constraints

# Relaxation Problem

FDs and MVDs as hard constraints

- Exact Implication $\Gamma \models \tau$.

FDs and MVDs as soft constraints

- Approximate implication $\sum_{\sigma \in \Gamma} \sigma \geq \tau$

Relaxation problem

- When can we convert EI to AI?

# FD and MVD

Functional Dependency (FD)

- A→B

(Embedded) Multivalued Dependency:

- EMVD: A→(B|C) if $\Pi_{ABC}(R) = \Pi_{AB}(R) \bowtie \Pi_{AC}(R)$

| A | B | C |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 2 |
| 1 | 2 | 1 |
| 1 | 2 | 2 |
| 2 | 2 | 2 |

→(B|C)
A→(B|C)

# Conditional Independence

- $X \perp Y \mid Z$ if $P(X,Y \mid Z) = P(X \mid Z)\,P(Y \mid Z)$
- Graphoid axioms [Pearl&Paz]

MVD: $A \twoheadrightarrow B$ iff $B \perp C | A$
Fails for EMVD

$B \perp C | A$
$\neg(B \perp C)$

| A | B | C | |
|---|---|---|---|
| 1 | 1 | 1 | 1/5 |
| 1 | 1 | 2 | 1/5 |
| 1 | 2 | 1 | 1/5 |
| 1 | 2 | 2 | 1/5 |
| 2 | 2 | 2 | 1/5 |

$H(Y|X) = H(XY) – H(X)$ $I(X;Y|Z) = H(XZ) + H(YZ) – H(XYZ) – H(Z)$

# Soft Constraints

$X \rightarrow Y$ iff $H(Y|X) = 0$

$X \perp Y \mid Z$ iff $I(X;Y|Z) = 0$

# Relaxation Holds for FDs+MVDs

$\Gamma \models \tau$

**Theorem** [Kenig&S] If $\Gamma$ consists of FDs+MVDs

- Then: $n^2/4 \sum_{\sigma\in\Gamma} \sigma \geq \tau$
- If $\tau$ is an FD then: $\sum_{\sigma\in\Gamma} \sigma \geq \tau$

**Example**: AB → C, AD → E, CE → F $\models$ ABD → F

H(C|AB) + H(E|AD) + H(F|CE) ≥ H(F|ABD)

# Relaxation Fails in General

[Kaced&Romashchenko], [Kenig&S]

**Theorem** $(C \perp D | A)$, $(C \perp D | B)$, $(A \perp B)$, $(B \perp C | D)$ $\models$ $C \perp D$

$\forall \lambda \geq 0$: $\lambda\,(I(C;D|A) + I(C;D|B) + I(A;B) + I(B;C|D)) < I(C;D)$

However, we can relax "in the limit"

**Theorem** [Kenig&S] If $\Gamma \models \sigma$, then forall $\varepsilon > 0$ exists $\lambda \geq 0$:

$$\lambda \sum_{\sigma \in \Gamma} \sigma + \varepsilon\, H(V) \geq \tau$$

V = all variables

# CI's Restricted to Shannon

**Theorem** (folklore) A Shannon implication $\Gamma \models \sigma$ relaxes to $\lambda \sum \sigma \geq \tau$ for some $\lambda > 0$

**Theorem** [Kenig&S]

(1) $\lambda \leq (2^n)!$

(2) there exists implications where $\lambda \geq 3$

Example: *Contraction Axiom* in semi-graphoids:

$$X \perp Y|Z \quad \& \quad X \perp W \mid YZ \quad \models \quad X \perp YW \mid Z$$

Relaxes to:

$$I(X;Y|Z) + I(X;W|YZ) \geq I(X;YW \mid Z)$$

# Summary of Part 2

- The *relaxation problem:* when can we convert exact implications to approximate implications

- Great practical importance: real data satisfies constraints only approximatively, need to relax

- Open problems: bounds on $\lambda$

# Conclusions

- Information theory is routinely used in ML
- Applications to data management: query processing and constraints
- Connections to difficult results in IT