Worst-case Optimal Join Algorithms

Techniques, Results, and Open Problems

Hung Q. Ngo

RelationalAI Inc.

Based on Joint Works with
Mahmoud Abo Khamis, Ely Porat, Chris Ré, Atri Rudra, Dan Suciu

PODS 2018

Formulation sketch

Paper with the same title on ArXiV contains much more details and references

Three Fundamental Problems in RDBMS

1. Query plans
   - Based on variable elimination and (equivalently) tree decompositions
     - Very general query types (Boolean, aggregations, max, min, CSP, PGM, etc.)
   - Refs:
     - [**FDB:** Olteanu, Závodný TODS'13]
     - [**FAQ**: Abo Khamis, Ngo, Rudra PODS'16 / SIGMOD Records 16]
     - [**TD**: Gottlob, Greco, Leone, Scarcello PODS'16]

2. Output size bound

3. Worst-case optimal join algorithms
   - To evaluate intermediate queries

Notations

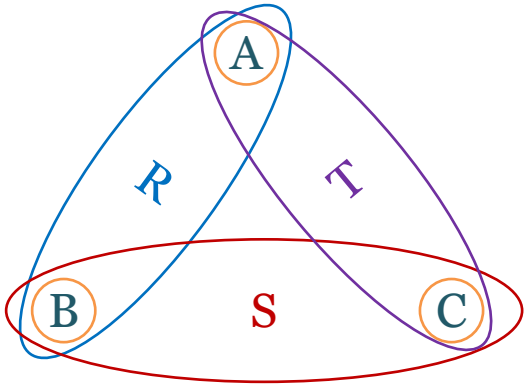
- $\mathbf{1}_E = \begin{cases} 1 & \text{if E holds} \\ 0 & \text{otherwise} \end{cases}$, e.g. $\mathbf{1}_{R(a,b)} = \begin{cases} 1 & \text{if } (\mathrm{a}, \mathrm{b}) \in R \\ 0 & \text{if } (\mathrm{a}, b) \notin R \end{cases}$

- (Query's) Hypergraph $\mathcal{H} = (V, \mathcal{E})$

- For each $F \in \mathcal{E}$, there's an input relation R_F on attributes F

$$Q(V) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(F)$$

- E.g. $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$



Degree Constraints

$R(A, B, C)$

A	B	C
1	1	1
1	1	2
1	2	3
2	1	4
2	1	5
2	1	6
3	1	7
3	2	8

- $\deg(A|BC) = \deg(ABC|BC) \leq 1$ Functional dependency
- $\deg(B|A) = \deg(AB|A) \leq 2$ Degree constraint
- $|R| = \deg(ABC|\emptyset) \leq 8$ Cardinality constraint

- Degree constraints
  - $(X, Y, N_{Y|X})$ where $\emptyset \subseteq X \subset Y \subseteq V$, with **guard** R
  - $\deg(Y|X) \coloneqq \max_{\boldsymbol{t}} |\pi_Y \sigma_{X=\boldsymbol{t}}(R)| \leq N_{Y|X}$
  - Generalizes both cardinality bounds and FDs

Twin Problems

Full conjunctive query Q

$$Q_\Delta(A,B,C) \leftarrow R(A,B), S(B,C), T(A,C)$$

Degree constraints DC

$$(\emptyset, AB, |R|), \quad (\emptyset, BC, |S|), \quad (\emptyset, AC, |T|), \quad (A, AB, K)$$

1. Worst-case output size bound

$$\sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|$$

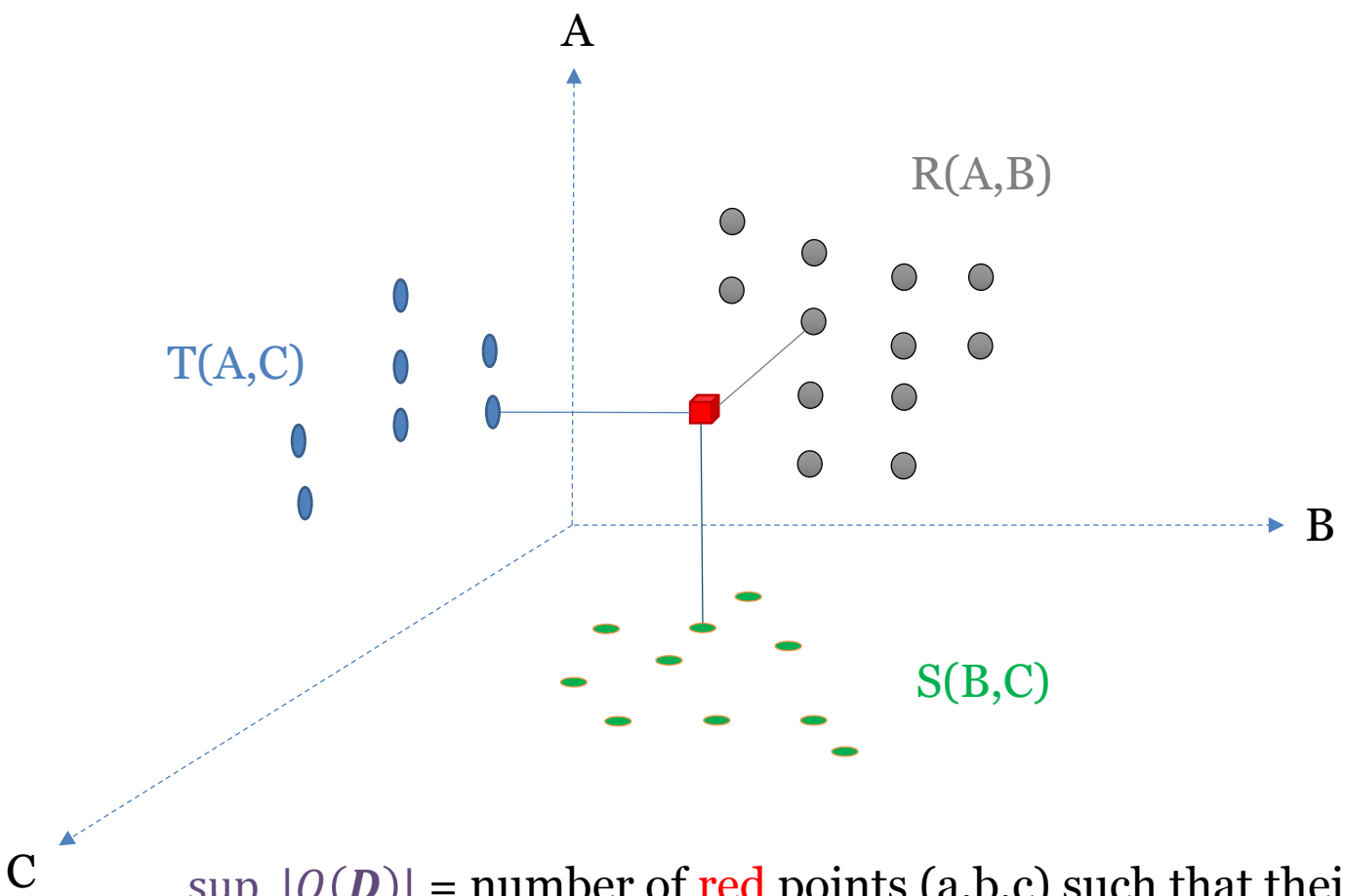
2. Worst-case optimal join (WCOJ) algorithm: evaluate Q in

$$O\left(\left[|\boldsymbol{D}| + \sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|\right] \mathrm{poly}(|Q|) \log |\boldsymbol{D}|\right) = \tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|\right)$$

$$\sup_{\boldsymbol{D} \models DC} |Q(\boldsymbol{D})|$$

- Geometry
  - [Loomis-Whitney 1949, isoperimetric inequalities]
  - [Bollobas-Thomason 1995] the Holder argument!
- Extremal combinatorics, graph theory
  - [Alon 81]
  - [Chung, Graham, Frankl, Shearer, 86] the entropy argument!
  - [Friedgut-Kahn 98]
- Constraint satisfaction
  - [Grohe & Marx SODA'06]
- Databases
  - [Atserias, Grohe, Marx FOCS'08]
  - [Gottlob, Lee, Valiant, Valiant JACM'2012]
  - [Ngo, Porat, Re, Rudra PODS'12]
  - [Abo Khamis, Ngo, Suciu PODS'16, PODS'17]

Geometric Connection



$\sup_{D \vDash DC} |Q(D)|$ = number of red points (a,b,c) such that their

shadow on (A,B) is covered by R
shadow on (A,C) is covered by T
shadow on (B,C) is covered by S

$\tilde{O}(|\boldsymbol{D}| + \sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|)$-Time Join Algorithms

- DC contains only cardinality constraints
  - [Grohe & Marx SODA'06] Join-Project plan $\tilde{O}(|\boldsymbol{D}| \times \sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|)$
  - [Todd Veldhuizen, LogicBlox 2009] LFTJ
  - [Ngo, Porat, Re, Rudra PODS'12] NPRR $\tilde{O}(\sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|)$
  - [Todd Veldhuizen, ICDT'14] LFTJ $\tilde{O}(\sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|)$
  - [Ngo, Re, Rudra, Sigmod Records 13] Generic Join $\tilde{O}(\sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})|)$
- DC contains FDs and degree constraints too
  - [Abo Khamis, Ngo, Suciu PODS'16, PODS'17] PANDA

Takeaways

- Proof of output size bound $\Longrightarrow$ efficient algorithm design
- Analyses crucially based on linear programming duality
- Connection to information theory and group theory
- **Practical**! "Just in time" for modern ML workload
- Many open problems!

Triangle Query Dissected

Illustration for $n = 3$, then let $3 \to \infty$

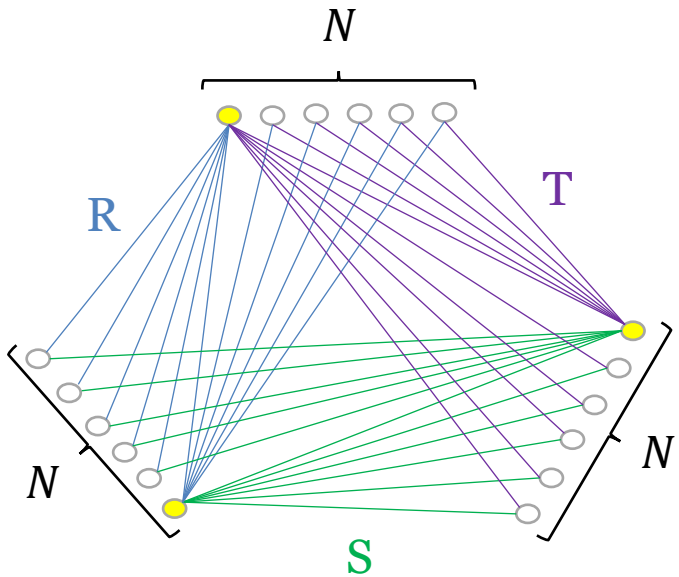
$Q_\Delta(A,B,C) \leftarrow R(A,B), S(B,C), T(A,C)$

- $\mathrm{DC} = \{(\emptyset, AB, N_R),\ (\emptyset, BC, N_S), (\emptyset, AC, N_T)\}$

- $\sup\limits_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})| = \max |R(A,B) \bowtie S(B,C) \bowtie T(A,C)|$

$$\text{subject to } |R| \leq N_R$$
$$|S| \leq N_S$$
$$|T| \leq N_T$$

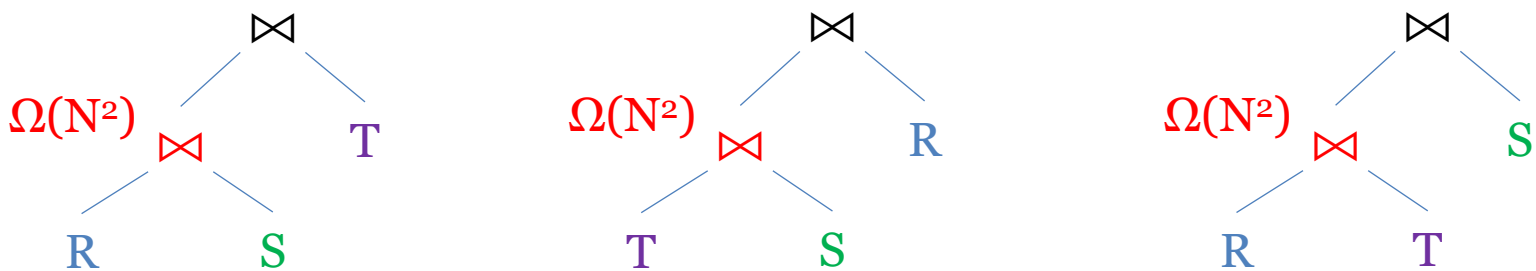
Any Pairwise Join Plan is Sub-Optimal



$$Q_\Delta(A,B,C) \leftarrow R(A,B), S(B,C), T(A,C)$$

$$|R| = |S| = |T| = 2N - 1 = O(N)$$

$$\sup_{\boldsymbol{D} \vDash DC} |Q_\Delta(\boldsymbol{D})| \leq \sqrt{|R| \cdot |S| \cdot |T|} = O(N^{1.5})$$



The Holder Argument

$$|Q_\Delta| = \sum_a \sum_b \sum_c \mathbf{1}_{R(a,b)} \mathbf{1}_{S(b,c)} \mathbf{1}_{T(a,c)} = \sum_a \sum_b \mathbf{1}_{R(a,b)} \sum_c \mathbf{1}_{S(b,c)} \mathbf{1}_{T(a,c)}$$

$$\leq \sum_a \sum_b \mathbf{1}_{R(a,b)} \sqrt{\sum_c \mathbf{1}_{S(b,c)}} \sqrt{\sum_c \mathbf{1}_{T(a,c)}}$$

Cauchy-Schwarz

$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \sqrt{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}$

$$= \sum_a \sum_b \mathbf{1}_{R(a,b)} \sqrt{|\sigma_{B=b} S|} \sqrt{|\sigma_{A=a} T|} = \sum_a \sqrt{|\sigma_{A=a} T|} \sum_b \mathbf{1}_{R(a,b)} \sqrt{|\sigma_{B=b} S|}$$

$$\leq \sum_a \sqrt{|\sigma_{A=a} T|} \sqrt{\sum_b \mathbf{1}_{R(a,b)}} \sqrt{\sum_b |\sigma_{B=b} S|}$$

Cauchy-Schwarz

$$= \sum_a \sqrt{|\sigma_{A=a} T|} \sqrt{|\sigma_{A=a} R|} \sqrt{|S|} = \sqrt{|S|} \sum_a \sqrt{|\sigma_{A=a} T|} \sqrt{|\sigma_{A=a} R|}$$

$$|Q_\Delta| \leq \sqrt{|R| \cdot |S| \cdot |T|}$$

E.g. number of triangles in a graph is $O(m^{1.5})$ [Alon, 1981]

Algorithm from Holder argument

$Q_\Delta(A,B,C) \leftarrow R(A,B), S(B,C), T(A,C)$

$$|Q_\Delta| = \sum_a \sum_b \sum_c \mathbf{1}_{R(a,b)} \mathbf{1}_{S(b,c)} \mathbf{1}_{T(a,c)}$$

$$\leq \sum_a \sum_b \mathbf{1}_{R(a,b)} \sqrt{\sum_c \mathbf{1}_{S(b,c)}} \sqrt{\sum_c \mathbf{1}_{T(a,c)}}$$

$$\boxed{= \sum_a \sum_b \mathbf{1}_{R(a,b)} \sqrt{|\sigma_{B=b} S|} \sqrt{|\sigma_{A=a} T|}}$$

$$\leq \sum_a \sqrt{|\sigma_{A=a} T|} \sqrt{\sum_b \mathbf{1}_{R(a,b)}} \sqrt{\sum_b |\sigma_{B=b} S|}$$

$$= \sum_a \sqrt{|\sigma_{A=a} T|} \sqrt{|\sigma_{A=a} R|} \sqrt{|S|}$$

$$\leq \sqrt{|R| \cdot |S| \cdot |T|}$$

for $a \in \pi_A R \cap \pi_A T$ **do**
  for $b \in \pi_B \sigma_{A=a} R \cap \pi_B S$ **do**
    for $c \in \pi_C \sigma_{B=b} S \cap \pi_C \sigma_{A=a} T$ **do**
      report (a, b, c)

$$\min\{|\pi_C \sigma_{B=b} S|, |\pi_C \sigma_{A=a} T|\}$$
$$\leq \sqrt{|\pi_C \sigma_{B=b} S| \cdot |\pi_C \sigma_{A=a} T|}$$
$$\leq \sqrt{|\sigma_{B=b} S| \cdot |\sigma_{A=a} T|}$$

Total runtime:

$$\sum_a \sum_b \mathbf{1}_{R(a,b)} \sqrt{|\sigma_{B=b} S| \cdot |\sigma_{A=a} T|}$$

The Entropy View (Shearer)

R

A	B	
1	1	1/3
1	2	1/3
2	2	1/3
2	5	¼
2	6	0

5

$H[A,B] \leq \log_2 |R|$

S

B	C	
1	1	1/6
1	2	1/6
2	1	1/3
2	2	1/3
3	1	0
3	3	0

6

$H[B,C] \leq \log_2 |S|$

T

A	C	
1	1	1/3
1	2	1/3
2	1	1/6
2	2	1/6

4

$H[A,C] \leq \log_2 |T|$

Q_Δ

A	B	C	
1	1	1	1/6
1	1	2	1/6
1	2	1	1/6
1	2	2	1/6
2	2	1	1/6
2	2	2	1/6

$H[A,B,C] = \log_2 |Q_\Delta|$

$H[B|A] \leq \log_2 3$

$H[A|B] \leq \log_2 2$

H satisfies degree constraints

H is entropic

Information Theory Detour

- Fix a joint distribution on n variables $V = \{X_1, \dots, X_n\}$
- The entropy function associated w/ the distribution is

$$H : 2^V \to \mathbb{R}^+$$

  - $H[F]$ = the entropy of the marginal distribution on F
  - $H[F]$ measures the amount of uncertainly on F
- $H[S|T] := H[S \cup T] - H[T]$: the conditional entropy
  - $H[S|T]$ = amount of uncertainty in S if we know T
- H is said to be *entropic*
  - $H[\emptyset] = 0$
  - $H[S] \leq H[T]$ if $S \subseteq T$ — monotonicity
  - $H[S] \geq 0 \; \forall S \subseteq V$ — non-negativity
  - $H[S|T] \geq H[S|T \cup Y], \forall S, T, Y \subseteq V$ — sub-modularity

  Shannon-type inequalities
- Γ_n^* = set of all entropic functions on n variables

The Entropy Argument

There is a joint distribution on (A,B,C) for which h is the entropy function of

$$\log|Q_\Delta| = H[A,B,C] \le \max h(A,B,C)$$

s.t. h is entropic

& h satisfies degree constraints

$$h(A,B) \le \log_2 |R|$$
$$h(B,C) \le \log_2 |S|$$
$$h(A,C) \le \log_2 |T|$$

$$h(A,B) + h(B,C) + h(A,C)$$

$$= h(A) + h(B|A) + h(B,C) + h(A,C)$$ Defn. of conditional entropy

$$= [h(A) + h(B,C)] \quad + [h(B|A) + h(A,C)]$$

$$\ge [h(A|B,C) + h(B,C)] + [h(B|A,C) + h(A,C)]$$ Submodularity, twice

$$= h(A,B,C) \quad + \quad h(A,B,C)$$

$$\log|Q_\Delta| \le \max_{h \in \Gamma_3^* \cap DC} h(A,B,C) \le \max_{h \in \Gamma_3^* \cap DC} \frac{h(A,B) + h(B,C) + h(A,C)}{2}$$

$$\le \frac{\log_2|R| + \log_2|S| + \log_2|T|}{2}$$

$$= \log_2 \sqrt{|R| \cdot |S| \cdot |T|} \qquad \Rightarrow \quad |Q_\Delta| \le \sqrt{|R| \cdot |S| \cdot |T|}$$

Algorithm from Entropy Argument

$$Q_\Delta(A,B,C) \leftarrow R(A,B), S(B,C), T(A,C)$$

$$\overset{R}{h(A,B)} + \overset{S}{h(B,C)} + \overset{T}{h(A,C)}$$

$$= \overset{R^h}{h(A)} + \overset{R^l}{h(B|A)} + h(B,C) + h(A,C)$$

$$= [h(A) + h(B,C)] \qquad + [h(B|A) \ + h(A,C)]$$

$$\geq [\overset{R^h}{h(A|B,C)} + \overset{S}{h(B,C)}] + [\overset{R^l}{h(B|A,C)} \ + \overset{T}{h(A,C)}]$$

$$= h(A,B,C) \qquad\qquad + \ \ h(A,B,C)$$

$$R^{heavy} = \{(a,b) \in R : |\sigma_{A=a} R| > \theta\}$$

$$R^{light} = \{(a,b) \in R : |\sigma_{A=a} R| \leq \theta\}$$

$$R(A,B) = R^{heavy}(A,B) \cup R^{light}(A,B)$$

$$O_1(A,B,C) = R^{heavy}(A,B) \bowtie S(B,C)$$

$$O_2(A,B,C) = R^{light}(A,B) \bowtie T(A,C)$$

$$\boxed{Q_\Delta = O_1 \cup O_2}$$

$$|R^{heavy}| \leq \frac{|R|}{\theta} \implies |O_1| \leq \frac{|R|}{\theta} \cdot |S|$$

$$|O_2| \leq |T| \cdot \theta$$

$$\theta = \sqrt{\frac{|R| \cdot |S|}{|T|}} \implies |O_1|, |O_2| \leq \sqrt{|R| \cdot |S| \cdot |T|}$$

Holder-based Bound

Holder wasn’t the first to prove the inequality

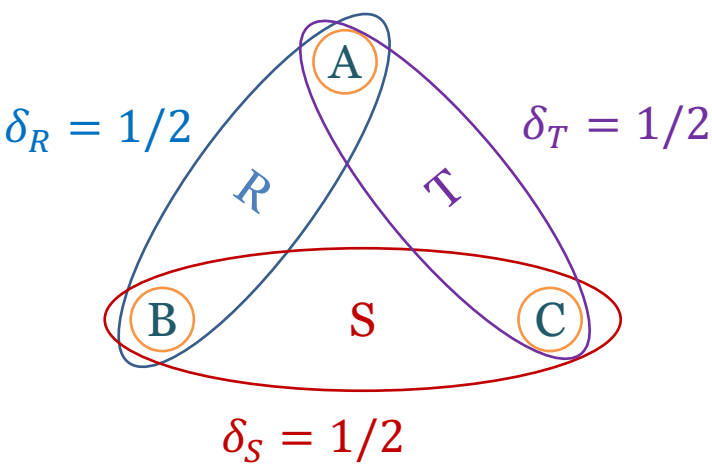
Fractional Edge Cover

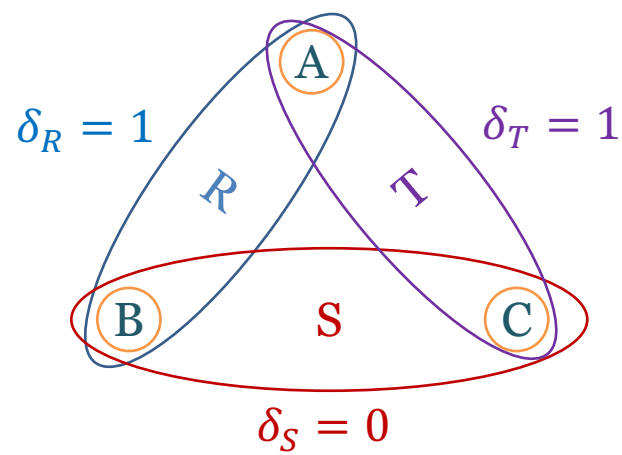
- Hypergraph $\mathcal{H} = (V, \mathcal{E})$ and query $Q(V) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(F)$

- *Fractional edge cover* $\delta: \mathcal{E} \to \mathbb{R}^+$

$$\sum_{\substack{F \in \mathcal{E} \\ v \in F}} \delta_F \geq 1, \qquad \forall v \in V$$

- E.g. $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$



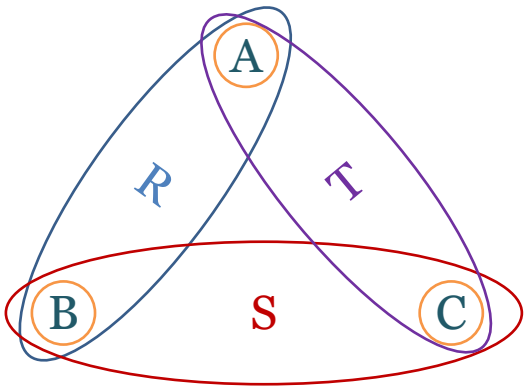


Friedgut Inequality (2004)

- Hypergraph $\mathcal{H} = (V, \mathcal{E})$ and query $Q(V) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(F)$
  - Fractional edge cover $\delta: \mathcal{E} \to \mathbb{R}^+$
  - Weight function $w_F: \prod_{X \in F} \text{Domain}(X) \to \mathbb{R}^+$

$$\sum_{\boldsymbol{x} \in Q} \prod_{F \in \mathcal{E}} w_F(\pi_F \boldsymbol{x})^{\delta_F} \leq \prod_{F \in \mathcal{E}} \left(\sum_{\boldsymbol{t} \in R_F} w_F(\boldsymbol{t}) \right)^{\delta_F}$$

- E.g. $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$
  - $w_R: \text{Domain}(A) \times \text{Domain}(B) \to \mathbb{R}^+$
  - $w_S: \text{Domain}(B) \times \text{Domain}(C) \to \mathbb{R}^+$
  - $w_T: \text{Domain}(A) \times \text{Domain}(C) \to \mathbb{R}^+$



$$\sum_{(a,b,c) \in Q_\Delta} w_R(a,b)^{\delta_R} w_S(b,c)^{\delta_S} w_T(a,c)^{\delta_T}$$

$$\leq \left(\sum_{(a,b) \in R} w_R(a,b) \right)^{\delta_R} \left(\sum_{(b,c) \in S} w_S(b,c) \right)^{\delta_S} \left(\sum_{(a,c) \in T} w_T(a,c) \right)^{\delta_T}$$

Proof for $n = 3$, then let $3 \to \infty$

$$\sum_{(a,b,c)\in Q_\Delta} w_R(a,b)^{\delta_R} w_S(b,c)^{\delta_S} w_T(a,c)^{\delta_T} = \sum_a \sum_b \sum_c w_R(a,b)^{\delta_R} w_S(b,c)^{\delta_S} w_T(a,c)^{\delta_T} \mathbf{1}_{R(a,b)} \mathbf{1}_{S(b,c)} \mathbf{1}_{T(a,c)}$$

$$= \sum_a \sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R} \sum_c [w_S(b,c)\mathbf{1}_{S(b,c)}]^{\delta_S} [w_T(a,c)\mathbf{1}_{T(a,c)}]^{\delta_T}$$

$$\leq \sum_a \sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R} \left[\sum_c w_S(b,c)\mathbf{1}_{S(b,c)}\right]^{\delta_S} \left[\sum_c w_T(a,c)\mathbf{1}_{T(a,c)}\right]^{\delta_T}$$

$$= \sum_a \left[\sum_c w_T(a,c)\mathbf{1}_{T(a,c)}\right]^{\delta_T} \sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R} \left[\sum_c w_S(b,c)\mathbf{1}_{S(b,c)}\right]^{\delta_S}$$

$$= \sum_a \left[\sum_c w_T(a,c)\mathbf{1}_{T(a,c)}\right]^{\delta_T} \sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R} \left[\sum_c w_S(b,c)\mathbf{1}_{S(b,c)}\right]^{\delta_S}$$

$$\leq \sum_a \left[\sum_c w_T(a,c)\mathbf{1}_{T(a,c)}\right]^{\delta_T} \left[\sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R}\right]^{\delta_R} \left[\sum_b \sum_c w_S(b,c)\mathbf{1}_{S(b,c)}\right]^{\delta_S}$$

$$\leq \left[\sum_a \sum_c w_T(a,c)\mathbf{1}_{T(a,c)}\right]^{\delta_T} \left[\sum_a \sum_b [w_R(a,b)\mathbf{1}_{R(a,b)}]^{\delta_R}\right]^{\delta_R} \left[\sum_b \sum_c w_S(b,c)\mathbf{1}_{S(b,c)}\right]^{\delta_S}$$

$$= \left(\sum_{(a,c)\in T} w_T(a,c)\right)^{\delta_T} \left(\sum_{(a,b)\in R} w_R(a,b)\right)^{\delta_R} \left(\sum_{(b,c)\in S} w_S(b,c)\right)^{\delta_S}$$

Holder inequality

$\alpha + \beta \geq 1, \alpha, \beta \in \mathbb{R}^+$

$$\sum_i x_i^\alpha y_i^\beta \leq \left(\sum_i x_i\right)^\alpha \left(\sum_i y_i\right)^\beta$$

$\delta_S + \delta_T \geq 1$

Friedgut Inequality ⇒ AGM-Bound

- Hypergraph $\mathcal{H} = (V, \mathcal{E})$ and query $Q(V) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(F)$
  - Fractional edge cover $\delta: \mathcal{E} \to \mathbb{R}^+$
  - Weight function $w_F: \prod_{X \in F} \text{Domain}(X) \to \mathbb{R}^+$

$$\sum_{\boldsymbol{x} \in Q} \prod_{F \in \mathcal{E}} w_F(\pi_F \boldsymbol{x})^{\delta_F} \leq \prod_{F \in \mathcal{E}} \left(\sum_{\boldsymbol{t} \in R_F} w_F(\boldsymbol{t}) \right)^{\delta_F}$$

- **AGM Bound**, set $w_F \equiv 1, \forall F$

$$|Q(\boldsymbol{D})| \leq \prod_F |R_F|^{\delta_F} \leq |\boldsymbol{D}|^{\sum_{F \in \mathcal{E}} \delta_F}$$

- $\rho^*(\mathcal{H}) \coloneqq \min_{\boldsymbol{\delta}} \sum_{F \in \mathcal{E}} \delta_F$: *fractional edge cover number*

$$|Q(D)| \leq |D|^{\rho^*(\mathcal{H})}$$

Entropy-based Bound

Lots of open problems

Information Theory Reminder

- Fix a joint distribution on n variables $V = \{X_1, \dots, X_n\}$
- The entropy function associated w/ the distribution is

$$H : 2^V \to \mathbb{R}^+$$

  - $H[F]$ = the entropy of the marginal distribution on F
  - $H[F]$ measures the amount of uncertainly on F
- $H[S|T] := H[S \cup T] - H[T]$: the conditional entropy
  - $H[S|T]$ = amount of uncertainty in S if we know T
- H is said to be *entropic*
  - $H[\emptyset] = 0$
  - $H[S] \leq H[T]$ if $S \subseteq T$ — monotonicity
  - $H[S] \geq 0 \; \forall S \subseteq V$ — non-negativity
  - $H[S|T] \geq H[S|T \cup Y], \forall S, T, Y \subseteq V$ — sub-modularity

  Shannon-type inequalities
- Γ_n^* = set of all entropic functions on n variables

The Entropy Argument

Hypergraph $\mathcal{H} = (V, \mathcal{E})$ and query $Q(V) \leftarrow \bigwedge_{F \in \mathcal{E}} R_F(F)$

h is the entropy function of some joint distribution on n random variables

$\log_2 |Q(\boldsymbol{D})| = H[V] \leq \sup h(V)$

s.t. $h \in \Gamma_n^*$ (i.e. h is entropic)

& $h \in HDC$ (i.e. h satisfies degree constraints)

$$h(Y) - h(X) := h(Y|X) \leq \log_2 N_{Y|X}$$
$$\forall (X, Y, N_{Y|X}) \in \mathrm{DC}$$

The entropic bound

$$\sup_{\boldsymbol{D} \vDash DC} \log_2 |Q(\boldsymbol{D})| = \sup \{h(V) \mid h \in \Gamma_n^* \cap \mathrm{HDC}\}$$

[ANS'16]

- Is a cone
- Not topologically closed
- Not polyhedral
- Membership not known to be decidable

Relaxations $\mathrm{M}_n \subset \Gamma_n^* \subset \bar{\Gamma}_n^* \subset \Gamma_n \subset \mathrm{SA}_n$

Non-negative set functions $h : 2^V \to \mathbb{R}^+$

SA_n

- Monotonicity $h(S) \leq h(T), S \subseteq T$
- Subadditivity $h(S \cup T) \leq h(S) + h(T), \forall S, T \subseteq [V]$

Polymatroids, Γ_n

- Submodularity $h(S \cup T) + h(S \cap T) \leq h(S) + h(T)$

Topological closure, $\bar{\Gamma}_n^*$

Entropic, Γ_n^*

Modular, M_n

$$h(X) = \sum_x h(x), \forall X \subseteq V$$

Hierarchy $\mathrm{M}_n \subset \Gamma_n^* \subset \bar{\Gamma}_n^* \subset \Gamma_n \subset \mathrm{SA}_n$

$\max\{h(V) \mid h \in \mathrm{M}_n \cap \mathrm{HDC}\} \leq$

- Modular bound
- Efficiently computable

$$\sup_{\boldsymbol{D} \vDash DC} \log_2 |Q(\boldsymbol{D})| = \sup\{h(V) \mid h \in \Gamma_n^* \cap \mathrm{HDC}\}$$

$$= \max\{h(V) \mid h \in \bar{\Gamma}_n^* \cap \mathrm{HDC}\}$$

- Entropic bound
- Not known to be computable

$$\leq \max\{h(V) \mid h \in \Gamma_n \cap \mathrm{HDC}\}$$

- Polymatroid bound
- Computable
- Not (known to be) efficiently computable
- Not tight
- Becomes fractional edge cover # under cardinality constraints

$$\leq \max\{h(V) \mid h \in \mathrm{SA}_n \cap \mathrm{HDC}\}$$

- Becomes **integral** edge cover # under cardinality constraints

Natural Follow-Up Questions

- Entropic bound
- Not known to be computable

- Polymatroid bound
- Computable, not tight

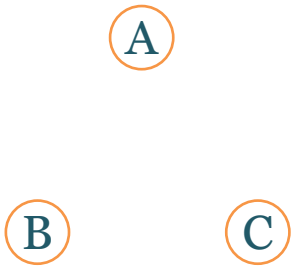
$$\sup_{\boldsymbol{D} \vDash DC} \log_2 |Q(\boldsymbol{D})| = \max\{h(V) \mid h \in \overline{\Gamma}_n^* \cap \mathrm{HDC}\} \quad \leq \max\{h(V) \mid h \in \Gamma_n \cap \mathrm{HDC}\}$$

When do these two bounds collide?

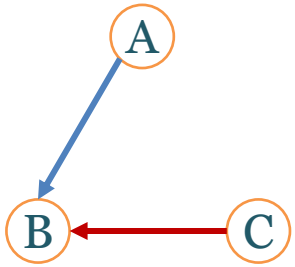
When they do, is there a matching algorithm?

Degree-Constraint Dependency Graph

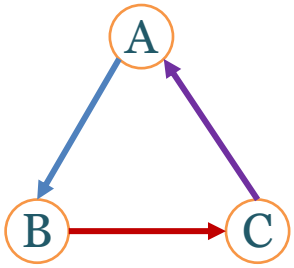
- $G_{\mathrm{DC}} = (V, E_{\mathrm{DC}})$
  - V = set of all variables -- same as V in $\mathcal{H} = (V, \mathcal{E})$
  - $E_{\mathrm{DC}} = \{X \times (Y - X) : (X, Y, N) \in \mathrm{DC}\}$



- Ex 1: $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$
  - $(\emptyset, AB, |R|), (\emptyset, BC, |S|), (\emptyset, AC, |T|)$
  - $G_{\mathrm{DC}} = (\{A, B, C, \}, \emptyset)$



- Ex 2: $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$
  - $(\emptyset, AB, |R|), (A\ AB, 1), (\emptyset, BC, |S|), (C, BC, N_2), (\emptyset, AC, |T|)$



- Ex 3: $Q(A, B, C) \leftarrow R(A, B), S(B, C), T(A, C)$
  - $(\emptyset, AB, |R|), (A\ AB, 1), (\emptyset, BC, |S|), (B, BC, N_2),$
  - $(\emptyset, AC, |T|), (C, AC, N_3)$

Acyclic Degree Constraints & Properties

- DC is *acyclic* iff G_{DC} is a directed acyclic graph (DAG)
  - DC is acyclic if it contains only cardinality constraints
  - DC is acyclic if it has cardinality constraints and key-foreign key FDs

Theorem: given Q and acyclic DC, then

$$\max_{h \in \mathrm{M}_n \cap \mathrm{HDC}} h(V) = \sup_{\boldsymbol{D} \vDash DC} \log_2 |Q(\boldsymbol{D})| = \max_{h \in \overline{\Gamma}_n^* \cap \mathrm{HDC}} h(V) = \max_{h \in \Gamma_n \cap \mathrm{HDC}} h(V)$$

- Recall: the modular bound is efficiently computable!

$$M_n = \{h: 2^V \to \mathbb{R}^+ \mid h(X) = \sum_x h(x), \forall X \subseteq V\}$$

$\max_{h \in \mathrm{M}_n \cap \mathrm{HDC}} h(V)$ and Linear Programming Duality

$$\max \sum_{v \in V} h(v)$$

$$\sum_{v \in Y - X} h(v) \le \log N_{Y|X}, \left(X, Y, N_{Y|X}\right) \in DC$$

$$h(v) \ge 0, v \in V$$

Fractional vertex packing

$$\min \sum_{(X,Y,N_{Y|X}) \in \mathrm{DC}} \delta_{Y|X} \log N_{Y|X}$$

$$\sum_{\substack{(X,Y,N_{Y|X}) \in DC \\ v \in Y - X}} \delta_{Y|X} \ge 1, \qquad v \in \mathrm{V}$$

$$\delta_{Y|X} \ge 0, \quad \left(X, Y, N_{Y|X}\right) \in \mathrm{DC}$$

Fractional edge covering

Exactly the AGM bound if all constraints are of the form $(\emptyset, Y, N_Y)$

h^* : primal optimal, and δ^* : dual optimal, then

$$\sup_{\boldsymbol{D} \vDash DC} |Q(\boldsymbol{D})| = 2^{h^*(V)} = \prod_{(X,Y,N_{Y|X}) \in \mathrm{DC}} N_{Y|X}^{\delta^*_{Y|X}}$$

Natural Follow-Up Questions

- Entropic bound
- Not known to be computable

- Polymatroid bound
- Computable, not tight

$$\sup_{\boldsymbol{D} \vDash DC} \log_2 |Q(\boldsymbol{D})| = \max\{h(V) \mid h \in \bar{\Gamma}_n^* \cap \mathrm{HDC}\} \quad \leq \max\{h(V) \mid h \in \Gamma_n \cap \mathrm{HDC}\}$$

When do these two bounds collide?

Acyclic DC is sufficient

When they do, is there a matching algorithm?

Yes, a Holder-based algorithm

But what if DC is not acyclic?

- Run the entropy-based algorithm (PANDA)

- **Theorem**: when DC is not acyclic, we can construct DC' from DC such that
  - DC' is acyclic
  - $\boldsymbol{D} \vDash \mathrm{DC} \Longrightarrow \boldsymbol{D} \vDash \mathrm{DC}'$

- Run Holder-based algorithm on DC'

Holder-based Algorithm

Backtracking search, along the line of LFTJ & Generic Join

Backtracking Search for Acylic DC

BTS

Input: $\mathcal{H} = (V, \mathcal{E})$ Relations R_F for $F \in \mathcal{E}$
Acyclic DC set, each $(X, Y, N_{Y|X})$ has a guard R

Pick $A \in V$ s.t. $\text{indegree}_{G_{\text{DC}}}(A) = 0$

$$I = \bigcap_{\substack{(\emptyset, Y, N_{Y|\emptyset}) \in DC \\ A \in Y, R \text{ guards } it}} \pi_A R$$

for each $a \in I$ do

$\mathcal{H}^a = (V^a, \mathcal{E}^a)$; relations R_F^a for $F \in \mathcal{E}^a$, constraints DC^a

$Q^a = \textbf{BTS}(\mathcal{H}^a, R_F^a, DC^a)$

$Q \leftarrow Q \cup Q^a$

Return Q

The Sub-Problems

… // variable A selected

for each $a \in I$ do

$\mathcal{H}^a = (V^a, \mathcal{E}^a)$; relations R_F^a for $F \in \mathcal{E}^a$, constraints DC^a

$Q^a = \mathbf{BTS}(\mathcal{H}^a, R_F^a, DC^a)$

$Q \leftarrow Q \cup Q^a$

…

$$V^a = V - \{A\} \qquad \mathcal{E}^a = \{F \setminus \{A\} : F \in \mathcal{E}\}$$

$$R_F^a = \begin{cases} R_F, & A \notin F \\ \pi_{F-\{A\}}\sigma_{A=a}R_F, & A \in F \end{cases}$$

$$\mathrm{DC}^a = \{ \underbrace{(X \setminus \{A\}, Y \setminus \{A\}, N_{Y\setminus\{A\}|\setminus\{A\}})}_{R^a \text{ guards this}} \mid \underbrace{(X, Y, N_{Y|X})}_{R \text{ guards this}} \in \mathrm{DC}\}$$

$\Leftarrow$

$R(A, B, C)$

A	B	C
a	1	1
a	1	2
a	2	3
a’	1	1
a’	1	2
a’	1	3
a”	1	1
a”	2	2

R^a

Analysis

$$Time(Q) \leq \sum_{a \in I} Time(Q^a)$$

$$\leq \sum_{a \in I} \prod_{\substack{(\emptyset, Y, N_{Y|\emptyset}) \in DC \\ A \in Y \\ R \text{ is the guard}}} |\sigma_{A=a} R|^{\delta_{Y|\emptyset}} \prod_{\substack{(X, Y, N_{Y|X}) \in DC \\ A \notin Y}} N_{Y|X}^{\delta_{Y|X}}$$

Induction hypothesis

$$\leq \prod_{\substack{(\emptyset, Y, N_{Y|\emptyset}) \in DC \\ A \in Y \\ R \text{ is the guard}}} \left(\sum_{a \in I} |\sigma_{A=a} R| \right)^{\delta_{Y|\emptyset}} \prod_{\substack{(X, Y, N_{Y|X}) \in DC \\ A \notin Y}} N_{Y|X}^{\delta_{Y|X}}$$

Friedgut inequality

$$\leq \prod_{\substack{(\emptyset, Y, N_{Y|\emptyset}) \in DC \\ A \in Y \\ R \text{ is the guard}}} (N_{Y|\emptyset})^{\delta_{Y|\emptyset}} \prod_{\substack{(X, Y, N_{Y|X}) \in DC \\ A \notin Y}} N_{Y|X}^{\delta_{Y|X}}$$

R guards $(\emptyset, Y, N_{Y|\emptyset})$

$$\leq \prod_{(X, Y, N_{Y|X}) \in DC} N_{Y|X}^{\delta_{Y|X}}$$

Entropy-based Algorithm

Where Shannon-type inequalities are interpreted as relational operators

So we have to settle for less

- Entropic bound
- Not known to be computable

- Polymatroid bound
- Computable, not tight

$$\sup_{\boldsymbol{D} \models DC} \log_2 |Q(\boldsymbol{D})| = \max\{h(V) \mid h \in \bar{\Gamma}_n^* \cap \mathrm{HDC}\} \quad \leq \max\{h(V) \mid h \in \Gamma_n \cap \mathrm{HDC}\}$$

- PANDA runs in time $\tilde{O}\left(|\boldsymbol{D}| + 2^{\text{polymatroid bound}}\right)$
  - $\tilde{O}$ hides humongous factors in query size and polylog in data size
  - Works for (a type of) disjunctive datalog rule ⊃ conjunctive queries

- Three main steps
  1. $\max\{h(V) \mid h \in \Gamma_n \cap \mathrm{HDC}\}$ and its dual gives Shannon-flow inequality
  2. Show that there's a "*proof sequence*" for the SFI
  3. Convert each step in the proof sequence into a relational operator

Example

- $Q \leftarrow R(A,B), S(B,C), T(C,D), W(A,C,D), V(A,B,D)$

Constraint	Guard
$(\emptyset, AB, N_1)$	R
$(\emptyset, BC, N_2)$	S
$(\emptyset, CD, N_3)$	T
(AC, ACD, N_4)	W
(BD, ABD, N_5)	V

- $2^{\text{polymatroid bound}} = \sqrt{N_1 N_2 N_3 N_4 N_5}$
  - (under mild assumptions above relative magnitutes of N_i)
  - E.g. $N_1 = N_2 = N_3 = N, N_4 = N_5 = 1$ then we want $O(N^{1.5})$-time

Corresponding Shannon-flow inequality

- The following inequality holds for *all* polymatroids

$$h(ABCD) \leq \frac{1}{2}(h(AB) + h(BC) + h(CD) + h(ACD|AC) + h(ABD|BD))$$

- Implies $|Q| \leq \sqrt{N_1 N_2 N_3 N_4 N_5}$

Proof sequence

$$
\begin{aligned}
& h(AB) + h(BC) + h(CD) + h(ACD|AC) + h(ABD|BD) \\
&= h(AB) + h(B) + h(BC|B) + h(CD) + h(ACD|AC) + hABD|BD) \\
&\geq h(AB) + h(B) + h(BC|B) + h(BCD|B) + h(ACD|AC) + h(ABD|BD) \\
&= h(AB) + h(BC|B) + h(BCD) + h(ACD|AC) + h(ABD|BD) \\
&\geq h(AB) + h(BC|B) + h(BCD) + h(ACD|AC) + h(ABCD|BCD) \\
&= h(AB) + h(BC|B) + h(ACD|AC) + h(ABCD) \\
&\geq h(AB) + h(ABC|AB) + h(ACD|AC) + h(ABCD) \\
&= h(ABC) + h(ACD|AC) + h(ABCD) \\
&\geq h(ABC) + h(ABCD|ABC) + h(ABCD) \\
&= h(ABCD) + h(ABCD)
\end{aligned}
$$

PANDA

Proof step	Operation	Action
$h(BC) \rightarrow h(B) + h(BC\|B)$	Partition	$S \rightarrow S^{heavy} \cup S^{light}$ threshold θ
$h(CD) \rightarrow h(BCD\|D)$	NOOP	T "affiliated" with $h(BCD\|B\|)$
$h(B) + h(BCD\|B) \rightarrow h(BCD)$	Join	$I_1(B,C,D) \leftarrow S^{heavy}(B,C) \bowtie T(C,D)$
$h(ABD\|BD) \rightarrow h(ABCD\|BCD)$	NOOP	V "affiliated" with $h(ABCD\|BCD)$
$h(ABCD\|BCD) + h(BCD) \rightarrow h(ABCD)$	Join	$O_1 \leftarrow V(A,B,D) \bowtie I_1(B,C,D)$
$h(BC\|B) \rightarrow \mathrm{h}(ABC\|AB)$	NOOP	S^{light} "affiliated" with $h(ABC\|AB)$
$h(AB) + h(ABC\|AB) \rightarrow h(ABC)$	Join	$I_2(A,B,C) \leftarrow R(A,B) \bowtie S^{light}(B,C)$
$h(ACD\|AC) \rightarrow h(ABCD\|ACD)$	NOOP	W "affiliated with $h(ABCD\|ACD)$
$h(ABC) + h(ABCD\|ABC) \rightarrow h(ABCD)$	Join	$O_2(A,B,C,D) \leftarrow I_2(A,B,C) \bowtie W(A,C,D)$

Open Problems

Some are **much** harder than others

1. Is the entropic bound computable?

2. What is the computational (query) complexity of computing the polymatroid bound?

3. For which class of DC is the best acyclic modification tight w.r.t. DC?

4. For which class of DC is the polymatroid bound tight?

5. Improve the (humongous) hidden factors in PANDA

6. Theory and algorithm for average case output size bound

7. Theory and practical algorithms for instance-optimal query evaluation

8. Theory and practical algorithms for optimizers of multi-way join operators

Spectrum of Problems and Solutions

	Simple		Complex
Constraints	Cardinality only		General degree constraints ($\supset$ FDs & cardinality constraints)
Output size bound	Tight & poly-time computable in query complexity		Tight & not (known to be) computable
			Not tight & computable (but not known to be in P)
Argument	Holder argument		Entropy argument
Best known algorithms	NPRR, LFTJ, Generic Join		PANDA
	Simple		Complicated
	Worst-case optimal		Not worst-case optimal
	No hidden factor in $\tilde{O}$		Large hidden factor in $\tilde{O}$
	$O(1)$ memory footprint		Requires intermediates
	Based on Holder-style proof		Based on Shannon-type inequality style proof
	Practical (LogicBlox, RelationalAI, etc.)		…

See writeup

Thank you!