RelationalAI

Logic & Algorithms in DB & AI – 2023

Variable Elimination and Tensor Decomposition

Hung Q. Ngo

Outline

Worst-Case Optimality is Insufficient

Variable Elimination

Tensor Decomposition

Open Problems – Research Directions

Worst-Case Optimality is Insufficient for Queries with Aggregations

4-cycle detection

- $Q() \leftarrow \exists A, B, C, D\ E(A,B) \wedge E(B,C) \wedge E(C,D) \wedge E(D,A)$
- Worst-case size bound on $|Q|$ is 1, on the conjunction is $|E|^2$
- Answerable in $O(|E|^{\frac{3}{2}})$-time [AYZ 97]

k-cycle detection

- Worst-case size bound on the conjunction is $O(|E|^{\lceil k/2 \rceil})$
- k-cycle detection can be done in $O(|E|^{2-\frac{1}{\lceil k/2 \rceil}})$-time [AYZ 97]

Worst-Case Optimality is Insufficient for Queries with Aggregations

k-walk counting (or any sub-trees in general)

- $Q = \text{count}[X_1, \ldots, X_{k+1} : E_1(X_1, X_2) \wedge \cdots \wedge E(X_k, X_{k+1})]$
- Worst-case size bound on $|Q|$ is 1, on the conjunction is $|E|^{\lceil k/2 \rceil}$
- But, we can answer this in $\tilde{O}(k|E|)$-time.

Conjunctive query with negation

- $Q() \leftarrow R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge X \neq U$
- Worst-case size bound on $|Q|$ is 1, on the conjunction is $O(N^2)$
- But, we can answer this in $\tilde{O}(N)$-time.

Need Two New Ideas in Our Framework

Idea 1: an abstraction to capture aggregation queries.

- $\max$, $\min$, $\sum$, $\exists$, etc.

Idea 2: a better notion of "optimality"

- Fine-grained complexity, parameterized complexity.

Given a semiring $(D, \oplus, \otimes, \mathbf{0}, \mathbf{1})$, a Sum-Product-query is

$$\varphi(\boldsymbol{X}_F) := \sum_{\boldsymbol{X}_{V-F}} \prod_{S \in \mathcal{E}} \psi_S(\boldsymbol{X}_S)$$

Example 1: $(\{\text{true}, \text{false}\}, \vee, \wedge, \text{false}, \text{true})$ is the Boolean semiring.

$$Q() = \bigvee_{A,B,C} E(A,B) \wedge E(A,C) \wedge E(B,C)$$

Example 2: $(\mathbb{R}, +, \times, 0, 1)$ is the sum-product semiring.

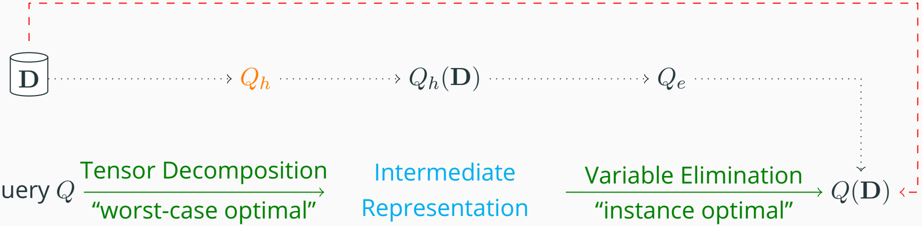
$$Q(A) = \sum_{B,C,D} E(A,B) \times E(B,C) \times E(C,D) \times \mathbf{1}_{A \neq D}$$

Idea 2: Better Notion of Optimality

$$\tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D}'\models s(\boldsymbol{D})} |Q(\boldsymbol{D}')|\right) \implies \tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D}'\models s(\boldsymbol{D})} f(Q,\boldsymbol{D}') + |Q(\boldsymbol{D})|\right)$$

Accompanied by a meta algorithm

$Q \in$ { database, CSP, PGM, logic, graph, linear algebra } queries



Outline

Worst-Case Optimality is Insufficient

Variable Elimination

Tensor Decomposition

Open Problems – Research Directions

Overloading notation

$$R(a, b) := \mathbf{1}_{(a,b)\in R} \qquad S(b, c) := \mathbf{1}_{(b,c)\in S} \qquad T(c, d) := \mathbf{1}_{(c,d)\in T}$$

Variable elimination [ZP 94]

Works for any semi-ring!

$$\begin{aligned} Q() &= \sum_a \sum_b \sum_c \sum_d R(a, b) \cdot S(b, c) \cdot T(c, d) \\ &= \sum_a \sum_b \sum_c R(a, b) \cdot S(b, c) \cdot \sum_d T(c, d) = \sum_a \sum_b \sum_c R(a, b) \cdot S(b, c) \cdot W(c) \\ &= \sum_a \sum_b R(a, b) \cdot \sum_c S(b, c) \cdot W(c) = \sum_a \sum_b R(a, b) \cdot V(b) \end{aligned}$$

Database jargon: (aggregation | projection | predicate) pushdowns

Indicator Projection, and Fractional Hypertree Width

[GM 06] [ANR 16]

$$Q = \sum_{a,b,c,d,e,f} R(a,b)S(a,c)T(b,c,d,e)W(e,f)V(d,f)$$

$$= \sum_{a,b,c,d,e} R(a,b)S(a,c)T(b,c,d,e) \underbrace{\sum_{f} \pi_{de}T(d,e)\, W(e,f)V(d,f)}_{\tilde{O}(N^{3/2}),\ \text{WCOJ}}$$

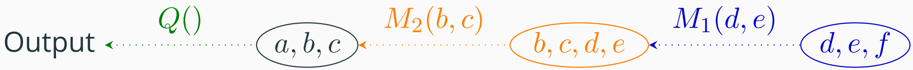
$$= \sum_{a,b,c,d,e} R(a,b)S(a,c)T(b,c,d,e)M_1(d,e)$$

$$= \sum_{a,b,c} R(a,b)S(a,c) \sum_{d,e} \pi_b R(b) \pi_S(c)\, T(b,c,d,e)M_1(d,e)$$

$$= \sum_{a,b,c} R(a,b)S(a,c)M_2(b,c) \qquad = \underbrace{\sum_{a,b,c} R(a,b)S(a,c)M_2(b,c)}_{\tilde{O}(N^{3/2}),\ \text{WCOJ}}$$

Variable Elimination, Message Passing, Belief Propagation, Yannakakis

$$Q = \sum_{a,b,c,d,e,f} R(a,b)S(a,c)T(b,c,d,e)W(e,f)V(d,f)$$



- Yannakakis algorithm [Yannakakis 81]
- Belief propagation [Pearl 82]

Indicator Projection, and Fractional Hypertree Width

$$Q = \sum_{a,b,c,d,e,f} R(a,b)S(a,c)T(b,c,d,e)W(e,f)V(d,f)$$

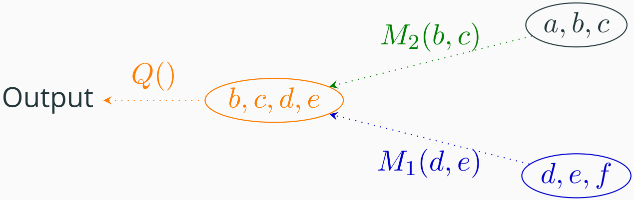
$$= \sum_{a,b,c,d,e} R(a,b)S(a,c)T(b,c,d,e) \underbrace{\sum_{f} \pi_{de}T(d,e)\, W(e,f)V(d,f)}_{M_1(d,e) \text{ in } \tilde{O}(N^{3/2}), \text{ WCOJ}}$$

$$= \sum_{b,c,d,e} T(b,c,d,e)M_1(d,e) \underbrace{\sum_{a} \pi_{bc}T(b,c)R(a,b)S(a,c)}_{M_2(b,c) \text{ in } O(N^{3/2}), \text{ WCOJ}}$$

$$= \sum_{b,c,d,e} M_1(d,e)T(b,c,d,e)M_2(b,c)$$

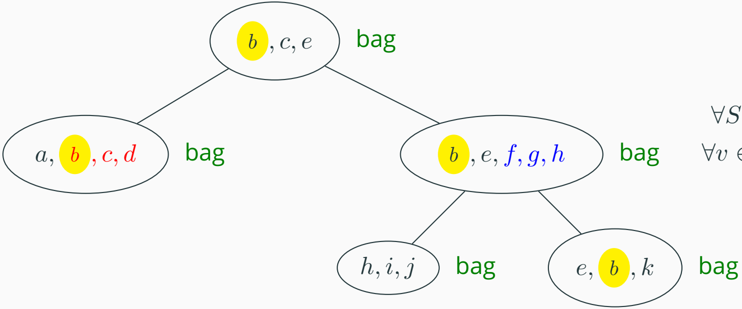
Variable Elimination, Message Passing, Belief Propagation, Yannakakis

$$Q = \sum_{a,b,c,d,e,f} R(a,b)S(a,c)T(b,c,d,e)W(e,f)V(d,f)$$



Detour: Tree Decompositions

$$S() \leftarrow R(a,b,d) \wedge c < d \wedge T(c,b,d) \wedge U(b,e) \wedge V(c,e)$$
$$\wedge\, b+e=f \wedge W(b,e,g) \wedge g/f=h \wedge X(i,j,h) \wedge e-b=k.$$



Hypergraph $\mathcal{H} = (V, \mathcal{E})$

Tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$

$\forall S \in \mathcal{E}, \exists t \in V(\mathcal{T})$ s.t. $S \subseteq \chi(t)$

$\forall v \in V, \{t \mid v \in \chi(t)\}$ is a subtree

Dynamic Programming over a Single Tree Decomposition

Fix $(\mathcal{T}, \chi)$

Multiple Conjunctive Rules

$\mathbf{D}$ ⋯→ Q_h ⋯→ $Q_h(\mathbf{D})$ ⋯→ Q_e ⋯→ $Q(\mathbf{D})$

$$\bigwedge_{F \in \mathcal{E}} R_F \xrightarrow[\text{WCOJ algo}]{\tilde{O}\left(\min_{(\mathcal{T},\chi)} \max_{\mathbf{D}' \models \mathsf{DC}} \max_{t \in V(\mathcal{T})} |P_t(\mathbf{D}')|\right)} \bigwedge_{t \in V(\mathcal{T})} T_{\chi(t)} \xrightarrow[\text{Variable Elimination}]{\tilde{O}(|Q_h(\mathbf{D})| + |Q(\mathbf{D})|)} Q(\mathbf{D})$$

$$P_t \;:\; T_{\chi(t)} \leftarrow \bigwedge_{F \in \mathcal{E}} R_F$$

Recall the polymatroid bound:

$$\log \min_{(\mathcal{T},\chi)} \max_{\mathbf{D}' \models \mathsf{DC}} \max_{t \in V(\mathcal{T})} |P_t(\mathbf{D}')| \leq \min_{(\mathcal{T},\chi)} \max_{t \in V(\mathcal{T})} \max_{h \in \Gamma_n \cap \mathsf{DC}} h(\chi(t))$$

$$\tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} f(Q, \boldsymbol{D}') + |Q(\boldsymbol{D})|\right)$$

Becomes:

$$\tilde{O}\left(|\boldsymbol{D}| + 2^{\min_{(\mathcal{T},\chi)} \max_{t \in V(\mathcal{T})} \max_{h \in \Gamma_n \cap \mathsf{DC}} h(\chi(t))} + |Q(\boldsymbol{D})|\right)$$

Corrolaries when the input has only cardinality constraints:

- $\tilde{O}\left(|\boldsymbol{D}| + |Q(\boldsymbol{D})|\right)$ if Q is acyclic [Y 81]
- $\tilde{O}\left(|\boldsymbol{D}| + |\boldsymbol{D}|^{\mathsf{fhtw}(Q)} + |Q(\boldsymbol{D})|\right)$ fhtw = *fractional hypertree width*, [GM 06]

Other Variations

- Using algorithms other than WCOJ, we get *(generalized) hypertree width* bounds or *tree-width* bounds. See [GGLS 16] for more.

$$\text{subw} \leq \text{fhtw} \leq \text{ghtw} \leq \text{tw} + 1.$$

- Sum-Product-Queries With Free Variables $F \subset V$ faqw

$$\tilde{O}\left(|\boldsymbol{D}| + 2^{\min_{(\mathcal{T},\chi) \in \mathsf{TD}(F)} \max_{t \in V(\mathcal{T})} \max_{h \in \Gamma_n \cap \mathsf{DC}} h(\chi(t))} + |Q(\boldsymbol{D})| \right)$$

Outline

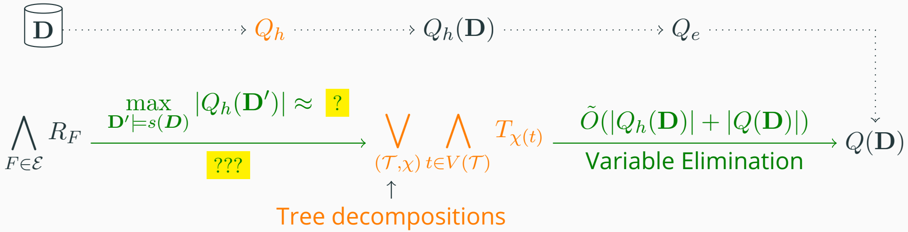
Worst-Case Optimality is Insufficient

Variable Elimination

Tensor Decomposition

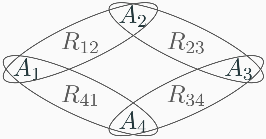
Open Problems – Research Directions

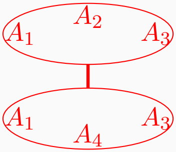
Multiple Tree Decompositions!

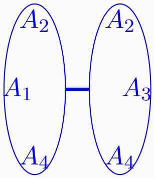


Assuming Boolean conjunctive query for now

Example: $Q : S() \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$







$$(T_{123} \wedge T_{134}) \vee (T_{124} \wedge T_{234}) \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$

By distributivity, rewrite the head:

$$(T_{123} \vee T_{124}) \wedge (T_{123} \vee T_{234}) \wedge (T_{134} \vee T_{124}) \wedge (T_{134} \vee T_{234}) \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$

(Each “clause” has one bag per TD)

Example: $Q \; : \; S() \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$

$$(T_{123} \vee T_{124}) \wedge (T_{123} \vee T_{234}) \wedge (T_{134} \vee T_{124}) \wedge (T_{134} \vee T_{234}) \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$

Is the same as evaluating 4 disjunctive datalog rules:

$$T_{123} \vee T_{124} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$
$$T_{123} \vee T_{234} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$
$$T_{134} \vee T_{124} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$
$$T_{134} \vee T_{234} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$

Semantics of a Disjunctive Datalog Query

$$P: \qquad T_{123} \vee T_{124} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}$$

The model contains two tables $T_{123}(A_1, A_2, A_3)$ and $T_{124}(A_1, A_2, A_4)$, such that:

- if $R_{12}(a_1, a_2)$, $R_{23}(a_2, a_3)$, $R_{34}(a_3, a_4)$, and $R_{41}(a_4, a_1)$
- then $T_{123}(a_1, a_2, a_3)$ OR $T_{124}(a_1, a_2, a_4)$.

Output size

- Write $\boldsymbol{T} \models P$ if $\boldsymbol{T}$ is a model $\qquad \boldsymbol{T} = (T_{123}, T_{124})$.
- Define $|\boldsymbol{T}| = \max\{|T_{123}|, |T_{124}|\}$

Multiple Tree Decompositions

Multiple Disjunctive Datalog Rules!

$$\mathbf{D} \dashrightarrow Q_h \dashrightarrow Q_h(\mathbf{D}) \dashrightarrow Q_e \dashrightarrow Q(\mathbf{D})$$

$$\bigwedge_{F \in \mathcal{E}} R_F \xrightarrow[\text{how?}]{\max_{\mathbf{D}' \models s(\boldsymbol{D})} \max_{\mathcal{B}} |P_{\mathcal{B}}(\mathbf{D}')|} \bigwedge_{\mathcal{B} \in \boldsymbol{B}} \bigvee_{B \in \mathcal{B}} T_B \xrightarrow[\text{Variable elimination}]{\tilde{O}(|Q_h(\mathbf{D})| + |Q(\mathbf{D})|)} Q(\mathbf{D})$$

$$P_{\mathcal{B}} \; : \; \bigvee_{B \in \mathcal{B}} T_B \leftarrow \bigwedge_{F \in \mathcal{E}} R_F \qquad\qquad |P_{\mathcal{B}}| \stackrel{\text{def}}{=} \min_{\boldsymbol{T} : \boldsymbol{T} \models P_{\mathcal{B}}} \max_{B \in \mathcal{B}} |T_B|$$

Each $\mathcal{B} \in \boldsymbol{B}$ is a collection of bags, one per TD.

New Problem: Polymatroid Bound for Disjunctive Datalog

$$P_{\mathcal{B}} : \bigvee_{B \in \mathcal{B}} T_B \leftarrow \bigwedge_{F \in \mathcal{E}} R_F \qquad \text{what is } \max_{\mathbf{D}' \models s(\boldsymbol{D})} |P_{\mathcal{B}}(\mathbf{D}')|?$$

Theorem (ANS 17)

Define $DC \stackrel{\text{def}}{=} \{h \mid \quad h(Y|X) \leq \log N \qquad \forall (X, Y, N) \in s(\boldsymbol{D})\}$, *then*

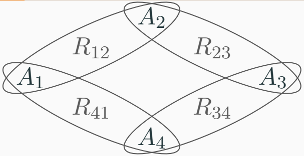
$$\log \sup_{\mathbf{D}' \models s(\boldsymbol{D})} |P_{\mathcal{B}}(\boldsymbol{D}')| \leq \max_{h \in \overline{\Gamma}^*_n \cap DC} \min_{B \in \mathcal{B}} h(B) \qquad \textit{Entropic Bound}$$

$$\leq \max_{h \in \Gamma_n \cap DC} \min_{B \in \mathcal{B}} h(B) \qquad \textit{Polymatroid Bound}$$

Generalize the entropy argument for conjunctive datalog (fun exercise!)

4-Cycle Example

$$P: \bigvee_{B \in \mathcal{B}} T_B \leftarrow \bigwedge_{F \in \mathcal{E}} R_F \qquad |P(\boldsymbol{D})| \stackrel{\text{def}}{=} \min_{\boldsymbol{T}:\boldsymbol{T} \models P} \max_{B \in \mathcal{B}} |T_B|$$



$$\text{DC}: \quad |R_{12}| \leq N, \quad |R_{23}| \leq N, \quad |R_{34}| \leq N, \quad |R_{41}| \leq N.$$

$$P_{123,234}: \ T_{123} \vee T_{234} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41} \qquad \mathcal{B} = \{123, 234\}$$

$$\begin{aligned}
\sup_{\mathbf{D}' \models \text{DC}} \log |P_{123,234}(\boldsymbol{D}')| &\leq \max_{h \in \Gamma_n \cap \text{DC}} \min\{h(A_1A_2A_3), h(A_2A_3A_4)\} \\
&\leq \max_{h \in \Gamma_n \cap \text{DC}} \frac{1}{2}[h(A_1A_2A_3) + h(A_2A_3A_4)] \\
\text{(Shannon-inequality)} &\leq \max_{h \in \Gamma_n \cap \text{DC}} \frac{1}{2}[h(A_1A_2) + h(A_2A_3) + h(A_3A_4)] \\
&\leq \frac{3}{2} \log N.
\end{aligned}$$

Realization of Idea 2

$$\tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} f(Q, \boldsymbol{D}') + |Q(\boldsymbol{D})|\right)$$

Becomes:

$$\tilde{O}\left(|\boldsymbol{D}| + 2^{\max_{\mathcal{B} \in \boldsymbol{B}} \max_{h \in \Gamma_n \cap \mathsf{DC}} \min_{B \in \mathcal{B}} h(B)} + |Q(\boldsymbol{D})|\right)$$

Another problem: WCOJ for disjunctive datalog!

Another Problem : WCOJ for Disjunctive Datalog

$$P_{\mathcal{B}} : \bigvee_{B \in \mathcal{B}} T_B \leftarrow \bigwedge_{F \in \mathcal{E}} R_F$$

compute a model in worst-case optimal time

Theorem (ANS 17)

Given a collection $s(\boldsymbol{D})$ of degree constraints, and a disjunctive datalog program $P_{\mathcal{B}}$, PANDA computes a model of $P_{\mathcal{B}}$ in time

$$\left(|\boldsymbol{D}| + 2^{\max_{h \in \Gamma_n \cap DC} \min_{B \in \mathcal{B}} h(B)} \right)$$

But where is the Shannon-flow inequality?

Evaluating a Disjunctive Datalog Program within $\max_{h \in \Gamma_n \cap \mathsf{HDC}} \min_{B \in \mathcal{B}} h(B)$

There exists non-negative $\boldsymbol{\lambda} = (\lambda_B)_{B \in \mathcal{B}}$, with $\|\boldsymbol{\lambda}\|_1 = 1$, s.t.

$$\max_{h \in \Gamma_n \cap \mathsf{HDC}} \min_{B \in \mathcal{B}} h(B) = \max_{h \in \Gamma_n \cap \mathsf{HDC}} \sum_{B \in \mathcal{B}} \lambda_B \, h(B)$$

Shannon-flow inequality: There exists $\boldsymbol{\delta} \geq \mathbf{0}$ s.t. (Farkas's lemma)

$$2^{\max_{h \in \Gamma_n \cap \mathsf{HDC}} \min_{B \in \mathcal{B}} h(B)} = \prod_{(X,Y,N) \in s(\boldsymbol{D})} N^{\delta_{Y|X}}$$

$$\sum_{B \in \mathcal{B}} \lambda_B \cdot h(B) \leq \sum_{(X,Y,N) \in s(\boldsymbol{D})} \delta_{Y|X} \cdot h(Y|X), \quad \forall h \in \Gamma_n$$

Proof sequence

Shannon-flow inequality: $h(Y|X) \stackrel{\text{def}}{=} h(Y) - h(X)$, $X \subseteq Y$

$$\sum_{B \in \mathcal{B}} \lambda_B \cdot h(B) \leq \sum_{(X,Y,N)} \delta_{Y|X} \cdot h(Y|X)$$

Proof sequence, convert RHS to LHS using following steps

(In)equality	Steps ($X \subseteq Y$)
$h(X) + h(Y|X) = h(Y)$	$h(X) + h(Y|X) \to h(Y)$
$h(Y) = h(X) + h(Y|X)$	$h(Y) \to h(X) + h(Y|X)$
$h(Y) \geq h(X)$	$h(Y) \to h(X)$
$h(Y|X) \geq h(Y \cup Z|X \cup Z)$	$h(Y|X) \to h(Y \cup Z|X \cup Z)$

Theorem

There is a proof sequence for every Shannon-flow inequality.

$$\tilde{O}\left(|\boldsymbol{D}| + \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} f(Q, \boldsymbol{D}') + |Q(\boldsymbol{D})|\right)$$

Becomes:

$$\tilde{O}\left(|\boldsymbol{D}| + 2^{\max_{\mathcal{B} \in \boldsymbol{B}} \max_{h \in \Gamma_n \cap \mathsf{DC}} \min_{B \in \mathcal{B}} h(B)} + |Q(\boldsymbol{D})|\right)$$

Corrolaries when the input has only cardinality constraints:

- $\tilde{O}\left(|\boldsymbol{D}| + |\boldsymbol{D}|^{\mathsf{subw}(Q)} + |Q(\boldsymbol{D})|\right)$ subw $=$ *submodular-width*, [Marx 13]
- Detecting k-cycle can be done in $O(N^{2-\frac{1}{\lceil k/2 \rceil}})$-time [AYZ 97]

Example: $P: \quad T_{123} \vee T_{234} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}.$

$$|R_{12}|, |R_{23}|, |R_{34}|, |R_{41}| \leq N \quad \Rightarrow \quad |P| \leq N^{3/2}$$

$$\begin{aligned}
\log |P| &\leq \min(h(A_1A_2A_3), h(A_2A_3A_4)) && \text{(polymatroid bound)} \\
&\leq \frac{1}{2}\left(h(A_1A_2A_3) + h(A_2A_3A_4)\right) && \text{(linearize)} \\
&\leq \frac{1}{2}\left(h(A_1A_2) + h(A_2A_3) + h(A_3A_4)\right) && \text{(Shannon-flow)} \\
&\leq \frac{3}{2}\log N && \text{(Cardinality constraints)}
\end{aligned}$$

Proof sequence	Proof Step
$h(A_1A_2) + h(A_2A_3) + h(A_3A_4)$	$\left(h(A_3A_4) \to h(A_4 \mid A_3) + h(A_3)\right)$
$h(A_1A_2) + h(A_2A_3) + h(A_4 \mid A_3) + h(A_3)$	$\left(h(A_4 \mid A_3) \to h(A_4 \mid A_2A_3)\right)$
$h(A_1A_2) + h(A_2A_3) + h(A_4 \mid A_2A_3) + h(A_3)$	$\left(h(A_2A_3) + h(A_4 \mid A_2A_3) \to h(A_2A_3A_4)\right)$
$h(A_1A_2) + h(A_2A_3A_4) + h(A_3)$	$\left(h(A_1A_2) \to h(A_1A_2 \mid A_3)\right)$
$h(A_1A_2 \mid A_3) + h(A_2A_3A_4) + h(A_3)$	$\left(h(A_1A_2 \mid A_3) + h(A_3) \to h(A_1A_2A_3)\right)$
$h(A_1A_2A_3) + h(A_2A_3A_4)$	

Example: $P: \quad T_{123} \vee T_{234} \leftarrow R_{12} \wedge R_{23} \wedge R_{34} \wedge R_{41}.$

$$|R_{12}|, |R_{23}|, |R_{34}|, |R_{41}| \leq N \quad \Rightarrow \quad |P| \leq N^{3/2}$$

$h(A_3 A_4) \to h(A_4 \mid A_3) + h(A_3)$	$R_{34}(A_3, A_4) \to R_{34}^{(\ell)}(A_3, A_4), R_3^{(h)}(A_3)$
$h(A_4 \mid A_3) \to h(A_4 \mid A_2 A_3)$	$R_{34}^{(\ell)}(A_3, A_4) \to R_{34}^{(\ell)}(A_3, A_4)$
$h(A_2 A_3) + h(A_4 \mid A_2 A_3) \to h(A_2 A_3 A_4)$	$R_{23}(A_2, A_3) \bowtie R_{34}^{(\ell)}(A_3, A_4) \to T_{234}(A_2, A_3, A_4)$
$h(A_1 A_2) \to h(A_1 A_2 \mid A_3)$	$R_{12}(A_1, A_2) \to R_{12}(A_1, A_2)$
$h(A_1 A_2 \mid A_3) + h(A_3) \to h(A_1 A_2 A_3)$	$R_{12}(A_1, A_2) \bowtie R_3^{(h)}(A_3) \to T_{123}(A_1, A_2, A_3)$

Outline

Worst-Case Optimality is Insufficient

Variable Elimination

Tensor Decomposition

Open Problems – Research Directions

Research Questions

- Reducing the number of tree decompositions to range over
- Sum-Product queries do not capture all problems
  - Extend this abstraction further, with the corresponding meta-algorithm
  - *Functional aggregate queries* KNR 16
  - Queries with arithmetic circuits, top-k queries, rank-enumeration
- General theory for tensor decompositions
- Removing the poly-log factor from PANDA
- Dealing with operators such as $\neq$, $<$

Some References

Yannakakis 81 Mihalis Yannakakis: Algorithms for Acyclic Database Schemes. VLDB 1981: 82-94

Pearl 82 Pearl, Judea. "Reverend Bayes on inference engines: A distributed hierarchical approach"). AAAI-82

ZP 1994 Zhang, N.L., Poole, D.:A Simple Approach to Bayesian Network Computations.In: 7th Canadian Conference on Artificial Intelligence,pp. 171–178

AYZ 95 Noga Alon, Raphael Yuster, Uri Zwick: Color-Coding. J. ACM 42(4): 844-856 (1995)

AYZ 97 Noga Alon, Raphael Yuster, Uri Zwick: Finding and Counting Given Length Cycles. Algorithmica 17(3): 209-223 (1997)

OZ 12 Dan Olteanu, Jakub Zavodny: Factorised representations of query results: size bounds and readability. ICDT 2012: 285-298

Marx 13 D. Marx. Tractable hypergraph properties for constraint satisfaction and conjunctive queries. J. ACM, 60(6):Art. 42, 51, 2013.

GGLS 16 Georg Gottlob, Gianluigi Greco, Nicola Leone, Francesco Scarcello: Hypertree Decompositions: Questions and Answers. PODS 2016: 57-74

ANR 16 Mahmoud Abo Khamis, Hung Q. Ngo, Atri Rudra: FAQ: Questions Asked Frequently. PODS 2016: 13-28

GM 06 Martin Grohe, Dániel Marx: Constraint solving via fractional edge covers. SODA 2006: 289-298

ANOS 19 Mahmoud Abo Khamis, Hung Q. Ngo, Dan Olteanu, Dan Suciu: Boolean Tensor Decomposition for Conjunctive Queries with Negation. ICDT 2019: 21:1-21:19

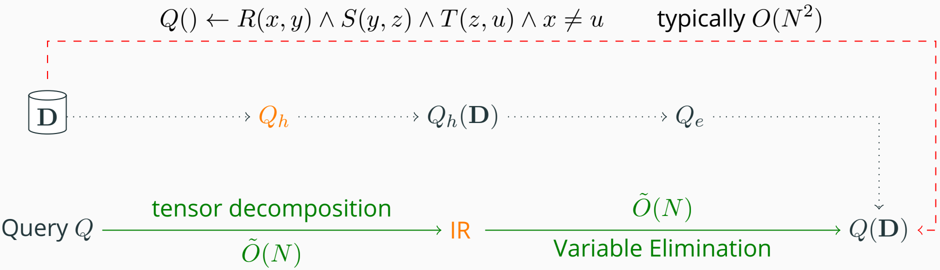
- Mahmoud Abo Khamis, Ryan R. Curtin, Benjamin Moseley, Hung Q. Ngo, XuanLong Nguyen, Dan Olteanu, Maximilian Schleich: On Functional Aggregate Queries with Additive Inequalities. PODS 2019: 414-431

Many Thanks!

Outline

Appendix

Example: Conjunctive Query with Negation



$$\text{IR} = \bigvee_{x,y,z,u} R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge \Big(\bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} x_i = b \wedge u_i = \neg b \Big)$$

where x_i, u_i denote the ith bit of x and u

Variable Elimination

$$\begin{aligned}
& \bigvee_{x,y,z,u} \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge x_i = b \wedge u_i = \neg b \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x,y,z} R(x,y) \wedge S(y,z) \wedge x_i = b \wedge \bigvee_{u} T(z,u) \wedge u_i = \neg b \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x,y,z} R(x,y) \wedge S(y,z) \wedge x_i = b \wedge \mu_1(z,i,b) \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x,y} R(x,y) \wedge x_i = b \wedge \bigvee_{z} S(y,z) \wedge \mu_1(z,i,b) \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x,y} R(x,y) \wedge x_i = b \wedge \mu_2(y,i,b) \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x} x_i = b \wedge \bigvee_{y} R(x,y) \wedge \mu_2(y,i,b) \\
= & \bigvee_{b\in\{0,1\}} \bigvee_{i\in[\log N]} \bigvee_{x} x_i = b \wedge \mu_3(x,i,b)
\end{aligned}$$