The Fine Grained Data Complexity of Conjunctive Query Evaluation

Dan Suciu

University of Washington

Problem Definition

Boolean conjunctive query $Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given input database $\boldsymbol{D}$, what is the time complexity to compute $Q(\boldsymbol{D})$?

Time expressed in terms of statistics on $\boldsymbol{D}$.

Problem Definition

Boolean conjunctive query $Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given input database $\boldsymbol{D}$, what is the time complexity to compute $Q(\boldsymbol{D})$?

Time expressed in terms of statistics on $\boldsymbol{D}$.

In this talk:
Combinatorial Algorithms: time $O(|\boldsymbol{D}|^{\text{subw}(Q)})$ [Khamis et al., 2016a, Khamis et al., 2024c].
Some lower bound techniques [Fan et al., 2023].

Not in this talk:
Using Fast Matrix Multiplication: time $O(|D|^{\omega\text{-subw}(Q)})$ [Khamis et al., 2024a].

“Proofs to Algorithms” Paradigm

Assume the algorithm for $Q(\boldsymbol{D})$ computes several intermediate results.

"Proofs to Algorithms" Paradigm

Assume the algorithm for $Q(\boldsymbol{D})$ computes several intermediate results.

- If these intermediate outputs have size B, then the runtime is $\Omega(B)$
- Thus, any algorithm with runtime $O(B)$ is a proof that outputs $\leq B$.

"Proofs to Algorithms" Paradigm

Assume the algorithm for $Q(\boldsymbol{D})$ computes several intermediate results.

- If these intermediate outputs have size B, then the runtime is $\Omega(B)$
- Thus, any algorithm with runtime $O(B)$ is a proof that outputs $\leq B$.

"Proofs to algorithms" paradigm is the converse:

(1) Prove: if $\boldsymbol{D}$ satisfies the given stats, then all outputs have size $\leq B$.
(2) Convert this proof into an algorithm for $Q(\boldsymbol{D})$, with runtime $\tilde{O}(B)$

"Proofs to Algorithms" Paradigm

Assume the algorithm for $Q(\boldsymbol{D})$ computes several intermediate results.

- If these intermediate outputs have size B, then the runtime is $\Omega(B)$
- Thus, any algorithm with runtime $O(B)$ is a proof that outputs $\leq B$.

"Proofs to algorithms" paradigm is the converse:

(1) Prove: if $\boldsymbol{D}$ satisfies the given stats, then all outputs have size $\leq B$.
(2) Convert this proof into an algorithm for $Q(\boldsymbol{D})$, with runtime $\tilde{O}(B)$

To prove bounds on the output size: Information Inequalities

Information Inequalities

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

The entropic vector of n r.v. $X_1, \ldots, X_n$ is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

The entropic vector of n r.v. $X_1, \ldots, X_n$ is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y
a	p
a	q
b	q
a	m

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

The entropic vector of n r.v. $X_1, \ldots, X_n$ is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y	p
a	p	$1/4$
a	q	$1/4$
b	q	$1/4$
a	m	$1/4$

$h(XY) = \log 4$

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

The entropic vector of n r.v. $X_1, \ldots, X_n$ is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y	p
a	p	$1/4$
a	q	$1/4$
b	q	$1/4$
a	m	$1/4$

$h(XY) = \log 4$

X	p
a	$3/4$
b	$1/4$

$h(X) \leq \log 2$

Y	p
p	$1/4$
q	$2/4$
m	$1/4$

$h(Y) \leq \log 3$

$\emptyset$	p
	1

$h(\emptyset) = 0$

Entropy, Entropic Vectors

Definition ([Shannon, 1948])

The entropy of a r.v. X with domain D is $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

The entropic vector of n r.v. $X_1, \ldots, X_n$ is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y	p
a	p	$1/4$
a	q	$1/4$
b	q	$1/4$
a	m	$1/4$

$h(XY) = \log 4$

X	p
a	$3/4$
b	$1/4$

$h(X) \leq \log 2$

Y	p
p	$1/4$
q	$2/4$
m	$1/4$

$h(Y) \leq \log 3$

$\emptyset$	p
	1

$h(\emptyset) = 0$

$\Gamma_n^* =$ set of entropic vectors (a.k.a. entropic functions)

Basic Shannon Inequalities

For all sets $U, V \subseteq \{X_1, \ldots, X_n\}$:

$$h(\emptyset) = 0$$

$$h(U \cup V) \geq h(U) \qquad \text{Monotonicity}$$

$$h(U) + h(V) \geq h(U \cup V) + h(U \cap V) \qquad \text{Submodularity}$$

Basic Shannon Inequalities

For all sets $U, V \subseteq \{X_1, \ldots, X_n\}$:

$$h(\emptyset) = 0$$

$$h(U \cup V) \geq h(U) \qquad \text{Monotonicity}$$

$$h(U) + h(V) \geq h(U \cup V) + h(U \cap V) \qquad \text{Submodularity}$$

Setting $h(V|U) \stackrel{\text{def}}{=} h(UV) - h(U)$, the inequalities become:

$$\boxed{h(V|U) \geq 0} \qquad \boxed{h(V|U) \geq h(V|UW)}$$

Basic Shannon Inequalities

For all sets $U, V \subseteq \{X_1, \ldots, X_n\}$:

$$h(\emptyset) = 0$$

$$h(U \cup V) \geq h(U) \qquad \text{Monotonicity}$$

$$h(U) + h(V) \geq h(U \cup V) + h(U \cap V) \qquad \text{Submodularity}$$

Setting $h(V|U) \stackrel{\text{def}}{=} h(UV) - h(U)$, the inequalities become:

$$\boxed{h(V|U) \geq 0} \qquad \boxed{h(V|U) \geq h(V|UW)}$$

A Shannon inequality is a consequence of the basic inequalities.

A Polymatroid is a vector $h \in \mathbb{R}_+^{2^n}$ that satisfies all Shannon inequalities

$$\Gamma_n = \text{set of polymatroids} \qquad \Gamma_n^* \subseteq \Gamma_n$$

Example: A Shannon Inequality

Example (Shearer)

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

Example: A Shannon Inequality

Example (Shearer)

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

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

Example: A Shannon Inequality

Example (Shearer)

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

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

Example: A Shannon Inequality

Example (Shearer)

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

$$\begin{aligned} &\underline{h(XY) + h(YZ)} + h(XZ) \\ &\qquad \geq h(XYZ) + h(Y) + h(XZ) \end{aligned}$$

Example: A Shannon Inequality

Example (Shearer)

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

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

Example: A Shannon Inequality

Example (Shearer)

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

$$\begin{aligned} \underline{h(XY) + h(YZ)} + h(XZ) & \\ & \geq h(XYZ) + \underline{h(Y) + h(XZ)} \\ & \geq 2h(XYZ) + h(\emptyset) \end{aligned}$$

Example: A Shannon Inequality

Example (Shearer)

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

$$\begin{aligned}
\underline{h(XY) + h(YZ)} + h(XZ) & \\
&\geq h(XYZ) + \underline{h(Y) + h(XZ)} \\
&\geq 2h(XYZ) + h(\emptyset) \\
&= 2h(XYZ)
\end{aligned}$$

History of Information Inequalities

- Shannon inequality: one satisfied by all polymatroids Γ_n
  Information inequality:[1] one satisfied by all entropic vectors Γ_n^*
- Pippenger [Pippenger, 1986]: Are Shannon inequalities complete?
  Breakthrough [Zhang and Yeung, 1998]: NO for $n \geq 4$. It follows $\Gamma_n^* \subsetneq \Gamma_n$.
- Γ_n^* is not convex; $\bar{\Gamma}_n^*$ is convex [Yeung, 2008], but is not a polytope [Matús, 2007].
- The characterization of $\bar{\Gamma}_n^*$ is open to date.

[1] A.k.a. entropic inequality

Output Size Bounds

Upper Bounds on the Output Size

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given statistics Stats, compute $B \stackrel{\text{def}}{=} \max_{\boldsymbol{D}:\boldsymbol{D} \models Stats} Q(|\boldsymbol{D}|)$.

2

Upper Bounds on the Output Size

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given statistics Stats, compute $B \stackrel{\text{def}}{=} \max_{\boldsymbol{D}:\boldsymbol{D} \models \mathit{Stats}} Q(|\boldsymbol{D}|)$.

- Cardinality constraints [2] $|R| \leq N$: $\boxed{h(XYZ) \leq \log N}$

[2] Assume R is ternary, $R(XYZ)$.

Upper Bounds on the Output Size

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given statistics Stats, compute $B \stackrel{\text{def}}{=} \max_{\boldsymbol{D}:\boldsymbol{D}\models Stats} Q(|\boldsymbol{D}|)$.

- Cardinality constraints [2] $|R| \leq N$: $\boxed{h(XYZ) \leq \log N}$
- Functional dependency $X \to Y$: $\boxed{h(Y|X) = 0}$
- Degree constraint $||\deg_R(YZ|X)||_\infty \leq P$: $\boxed{h(YZ|X) \leq \log P}$

[2] Assume R is ternary, $R(XYZ)$.

Upper Bounds on the Output Size

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$

Given statistics Stats, compute $B \stackrel{\text{def}}{=} \max_{\boldsymbol{D}:\boldsymbol{D} \models Stats} Q(|\boldsymbol{D}|)$.

- Cardinality constraints [2] $|R| \leq N$: $\boxed{h(XYZ) \leq \log N}$
- Functional dependency $X \rightarrow Y$: $\boxed{h(Y|X) = 0}$
- Degree constraint $||\deg_R(YZ|X)||_\infty \leq P$: $\boxed{h(YZ|X) \leq \log P}$
- ℓ_p-constraint $||\deg_R(YZ|X)||_p \leq S$: $\boxed{\frac{1}{p}h(X) + h(YZ|X) \leq \log S}$

[2] Assume R is ternary, $R(XYZ)$.

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$$

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$$

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$$

Let $p : Q_\Delta(\boldsymbol{D}) \to [0, 1]$ be uniform and $\boldsymbol{h}$ its entropy.

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$

Let $p : Q_\Delta(\boldsymbol{D}) \to [0, 1]$ be uniform and $\boldsymbol{h}$ its entropy.

$$\log N_1 + \log N_2 + \log N_3 \geq h(XY) + h(YZ) + h(XZ)$$

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$

Let $p : Q_\Delta(\boldsymbol{D}) \to [0, 1]$ be uniform and $\boldsymbol{h}$ its entropy.

$$\begin{aligned}\log N_1 + \log N_2 + \log N_3 \geq & h(XY) + h(YZ) + h(XZ) \\ \geq & 2h(XYZ)\end{aligned}$$

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$

$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$

Let $p : Q_\Delta(\boldsymbol{D}) \to [0, 1]$ be uniform and $\boldsymbol{h}$ its entropy.

$$\log N_1 + \log N_2 + \log N_3 \geq h(XY) + h(YZ) + h(XZ)$$
$$\geq 2h(XYZ) = 2 \log |Q(\boldsymbol{D})|$$

Example

Full CQ: $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$
$|R| \leq N_1, |S| \leq N_2, |T| \leq N_3.$

Example ([Chung et al., 1986] [Grohe and Marx, 2014] [Atserias et al., 2013])

$|Q_\Delta(\boldsymbol{D})| \leq (N_1 \cdot N_2 \cdot N_3)^{\frac{1}{2}}$

Let $p : Q_\Delta(\boldsymbol{D}) \to [0, 1]$ be uniform and $\boldsymbol{h}$ its entropy.

$$\log N_1 + \log N_2 + \log N_3 \geq h(XY) + h(YZ) + h(XZ)$$
$$\geq 2h(XYZ) = 2\log|Q(\boldsymbol{D})|$$

Other inequalities exists, e.g. $|Q_\Delta| \leq N_1 \cdot N_2$.

The Polymatroid Bound

Full CQ $Q(\boldsymbol{X})$, statistics Stats.

The polymatroid bound is: $b(Q, \texttt{Stats}) \stackrel{\text{def}}{=} \max_{\boldsymbol{h} \in \Gamma_n, \boldsymbol{h} \models \text{stats}} h(\boldsymbol{X})$

Fact: if $\boldsymbol{D} \models \mathit{Stats}$, then $|Q(\boldsymbol{D})| \leq 2^{b(Q, \mathit{Stats})}$.

Equal to the best bound one can derive using inequalities (by strong duality).

Example Revisited

$$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$$

$b(Q, \texttt{Stats})$ can be computed in one of two ways

Example Revisited

$$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$$

$b(Q, \texttt{Stats})$ can be computed in one of two ways

Primal program:

Maximize $h(XYZ)$ where

$$\begin{cases} h(XY) \leq \log |N_1| \\ h(YZ) \leq \log |N_2| \\ h(XZ) \leq \log |N_3| \\ \boldsymbol{h} \in \Gamma_3 \end{cases}$$

Example Revisited

$$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$$

$b(Q, \texttt{Stats})$ can be computed in one of two ways

Primal program:

Maximize $h(XYZ)$ where

$$\begin{cases} h(XY) \leq \log |N_1| \\ h(YZ) \leq \log |N_2| \\ h(XZ) \leq \log |N_3| \\ \boldsymbol{h} \in \Gamma_3 \end{cases}$$

Dual program:

Minimize

$$w_1 \log |N_1| + w_2 \log |N_2| + w_3 \log |N_3|$$

where $w_1, w_2, w_3 \geq 0$ s.t. $\forall \boldsymbol{h} \in \Gamma_3$:

$$w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$$

Short History of Output Size Bounds

AGM bound [Atserias et al., 2013]

Add FDs [Gottlob et al., 2012, Khamis et al., 2016b, Gogacz and Torunczyk, 2017, Ngo, 2022]

Add max degrees [Khamis et al., 2017].

Add degree sequences [Deeds et al., 2023a, Deeds et al., 2023b, Khamis et al., 2024b, Zhang et al., 2025]

Short History of Output Size Bounds

AGM bound [Atserias et al., 2013]

$$|Q_\Delta| \leq (|R| \cdot |S| \cdot |T|)^{1/2}$$

Add FDs [Gottlob et al., 2012, Khamis et al., 2016b, Gogacz and Torunczyk, 2017, Ngo, 2022]

Add max degrees [Khamis et al., 2017].

Add degree sequences [Deeds et al., 2023a, Deeds et al., 2023b, Khamis et al., 2024b, Zhang et al., 2025]

Short History of Output Size Bounds

AGM bound [Atserias et al., 2013]

$$|Q_\Delta| \leq (|R| \cdot |S| \cdot |T|)^{1/2}$$

Add FDs [Gottlob et al., 2012, Khamis et al., 2016b, Gogacz and Torunczyk, 2017, Ngo, 2022]

$$|Q_\Delta| \leq |R| \quad \text{when } S.Y \text{ is a key}$$

Add max degrees [Khamis et al., 2017].

$$|Q_\Delta| \leq |R| \cdot ||\deg_S(Z|Y)||_\infty$$

Add degree sequences [Deeds et al., 2023a, Deeds et al., 2023b, Khamis et al., 2024b, Zhang et al., 2025]

Short History of Output Size Bounds

AGM bound [Atserias et al., 2013]

$$|Q_\Delta| \leq \left(|R| \cdot |S| \cdot |T|\right)^{1/2}$$

Add FDs [Gottlob et al., 2012, Khamis et al., 2016b, Gogacz and Torunczyk, 2017, Ngo, 2022]

$$|Q_\Delta| \leq |R| \quad \text{when } S.Y \text{ is a key}$$

Add max degrees [Khamis et al., 2017].

$$|Q_\Delta| \leq |R| \cdot ||\deg_S(Z|Y)||_\infty$$

Add degree sequences [Deeds et al., 2023a, Deeds et al., 2023b, Khamis et al., 2024b, Zhang et al., 2025]

$$|Q_\Delta| \leq \left(||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2\right)^{1/3}$$

Short History of Output Size Bounds

AGM bound [Atserias et al., 2013]

$$|Q_\Delta| \leq (|R| \cdot |S| \cdot |T|)^{1/2}$$

Add FDs [Gottlob et al., 2012, Khamis et al., 2016b, Gogacz and Torunczyk, 2017, Ngo, 2022]

$$|Q_\Delta| \leq |R| \quad \text{when } S.Y \text{ is a key}$$

Add max degrees [Khamis et al., 2017].

$$|Q_\Delta| \leq |R| \cdot ||\deg_S(Z|Y)||_\infty$$

Add degree sequences [Deeds et al., 2023a, Deeds et al., 2023b, Khamis et al., 2024b, Zhang et al., 2025]

$$|Q_\Delta| \leq \left(||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2\right)^{1/3}$$

$$|Q_\Delta| \leq \left(||\deg_R(Y|X)||_3^3 \cdot ||\deg_S(Y|Z)||_3^3 \cdot |T|^5\right)^{1/6}$$

From Proofs to Algorithms

Data Partitioning: Example

Example

$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX), \qquad |Q_\Delta| \leq (N_1 N_2 N_3)^{\frac{1}{2}}$

Data Partitioning: Example

Example

$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$, $\quad |Q_\Delta| \leq (N_1 N_2 N_3)^{\frac{1}{2}}$

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

$$+$$

$$h(XY) + h(Z|Y) \qquad h(Y) + h(XZ)$$

$$\geq h(XYZ) \qquad \geq h(XYZ)$$

Data Partitioning: Example

Example

$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$, $\quad |Q_\Delta| \leq (N_1 N_2 N_3)^{\frac{1}{2}}$

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

$+$

- $h(XY) + h(Z|Y) \geq h(XYZ)$
- $h(Y) + h(XZ) \geq h(XYZ)$

$$R(XY) \bowtie S(YZ) \bowtie T(XZ)$$

$\bigcup$

- $R(XY) \bowtie S_{\text{light}}(Z|Y) = Q_{\text{light}}(XYZ)$
- $S_{\text{heavy}}(Y) \bowtie T(XZ) = Q_{\text{heavy}}(XYZ)$

Data Partitioning: Example

Example

$Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$, $\quad |Q_\Delta| \leq (N_1 N_2 N_3)^{\frac{1}{2}}$

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

$+$

$h(XY) + h(Z|Y) \quad\quad h(Y) + h(XZ)$

$\geq h(XYZ) \quad\quad \geq h(XYZ)$

$R(XY) \bowtie S(YZ) \bowtie T(XZ)$

$\bigcup$

$R(XY) \bowtie S_{\text{light}}(Z|Y) \quad\quad S_{\text{heavy}}(Y) \bowtie T(XZ)$

$= Q_{\text{light}}(XYZ) \quad\quad = Q_{\text{heavy}}(XYZ)$

The light/heavy thresholds are given by $\boldsymbol{h}^*$:

Light: $\quad ||\deg_{S_{\text{light}}}(Z|Y)||_\infty \leq 2^{h^*(Z|Y)}$

Heavy: $\quad |S_{\text{heavy}}(Y)| \leq 2^{h^*(Y)}$

Data Partitioning

Theorem ([Khamis et al., 2024c])

$Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{b(Q,Stats)})$ by data partitioning & joins.

- Light/heavy with threshold $h^*(Z|Y)$:

  We can only use this partitioning when inequality is divergent:
  i.e. a sum of two inequalities each with $h(Y)$ and $h(Z|Y)$ respectively.

Data Partitioning

Theorem ([Khamis et al., 2024c])

$Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{b(Q,Stats)})$ by data partitioning & joins.

- Light/heavy with threshold $h^*(Z|Y)$:

  We can only use this partitioning when inequality is divergent:
  i.e. a sum of two inequalities each with $h(Y)$ and $h(Z|Y)$ respectively.

- Uniformization partition into $\log N$ buckets
  Bucket i holds tuples with $\deg_R(Z|y) \in [2^{i-1}, 2^i]$.

  Repeated uniformizations lead to poly-log factor.

PANDA

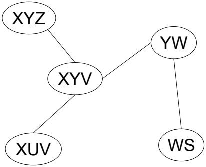
Proof-**A**ssisted e**N**tropic **D**egree-**A**ware

The Fractional Tree Width

Boolean CQ $Q() = \bigwedge_j R(\boldsymbol{U}_j)$

Denote a tree decomposition by $\boldsymbol{T} = (T, \chi)$

Fractional tree width $\boxed{\text{ftw}(Q) \stackrel{\text{def}}{=} \min_{\boldsymbol{T}} \max_{\boldsymbol{h} \models \text{Stats}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))}$



Theorem ([Grohe and Marx, 2014, Khamis et al., 2016a])

$Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{\text{ftw}(Q, \text{Stats})})$

The Submodular Width

$$\boxed{\text{ftw}(Q, \text{Stats}) \stackrel{\text{def}}{=} \min_{\boldsymbol{T}} \max_{\boldsymbol{h} \models \text{Stats}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))}$$

Marx [Marx, 2013] defines the submodular width using all tree decompositions:

$$\boxed{\text{subw}(Q, \text{Stats}) \stackrel{\text{def}}{=} \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))}$$

The Submodular Width

$$\boxed{\text{ftw}(Q, \text{Stats}) \stackrel{\text{def}}{=} \min_{\boldsymbol{T}} \max_{\boldsymbol{h} \models \text{Stats}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))}$$

Marx [Marx, 2013] defines the submodular width using all tree decompositions:

$$\boxed{\text{subw}(Q, \text{Stats}) \stackrel{\text{def}}{=} \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))}$$

Theorem ([Khamis et al., 2016a, Khamis et al., 2024c])

PANDA computes a Boolean query Q in time $\tilde{O}(2^{subw(Q, Stats)})$

Key technique: disjunctive datalog rules & data partitioning

Disjunctive Datalog Rules

A Disjunctive Datalog Rule, DDR, is:

$$\Sigma : \boxed{Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$$

Disjunctive Datalog Rules

A Disjunctive Datalog Rule, DDR, is:

$$\Sigma : \boxed{Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$$

A model is a tuple of relations $\Sigma = (Q_1, \ldots, Q_k)$ that satisfies the rule.

Its size is $\|\Sigma\| = \max_i |Q_i|$.

Disjunctive Datalog Rules

A Disjunctive Datalog Rule, DDR, is:

$$\Sigma : \boxed{Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$$

A model is a tuple of relations $\Sigma = (Q_1, \ldots, Q_k)$ that satisfies the rule.

Its size is $\|\Sigma\| = \max_i |Q_i|$.

Examples:

$$Q(XZ) \leftarrow R(XY) \wedge S(YZ)$$

Disjunctive Datalog Rules

A Disjunctive Datalog Rule, DDR, is:

$$\Sigma : \boxed{Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$$

A model is a tuple of relations $\Sigma = (Q_1, \ldots, Q_k)$ that satisfies the rule.

Its size is $\|\Sigma\| = \max_i |Q_i|$.

Examples:

$$Q(XZ) \leftarrow R(XY) \wedge S(YZ)$$
$$A(X) \vee B(Y) \leftarrow R(XY)$$

Disjunctive Datalog Rules

A Disjunctive Datalog Rule, DDR, is:

$$\Sigma : \boxed{Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$$

A model is a tuple of relations $\Sigma = (Q_1, \ldots, Q_k)$ that satisfies the rule.

Its size is $\|\Sigma\| = \max_i |Q_i|$.

Examples:

$$\begin{aligned} Q(XZ) &\leftarrow R(XY) \wedge S(YZ) \\ A(X) \vee B(Y) &\leftarrow R(XY) \\ A(XYZ) \vee B(YZU) &\leftarrow R(XY) \wedge S(YZ) \wedge T(ZU) \end{aligned}$$

Upper Bound and Algorithm for a DDR

$$\Sigma : \quad Q_1(\boldsymbol{V}_1) \vee \cdots \vee Q_k(\boldsymbol{V}_k) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$$

Polymatroid bound $\boxed{b(\Sigma, Stats) \stackrel{\text{def}}{=} \max_{\boldsymbol{h} \in \Gamma_n, \boldsymbol{h} \models \text{Stats}} \min(h(\boldsymbol{V}_1), \ldots, h(\boldsymbol{V}_k))}$

A linear optimization problem which has a matching Shannon inequality.

Theorem ([Khamis et al., 2024c])

$\exists$model Σ of size $\leq 2^{b(\Sigma, Stats)}$; it can be computed in time $\tilde{O}(2^{b(\Sigma, Stats)})$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

$$\text{subw}(Q, \text{Stats}) = \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T} \in \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k\}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))$$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

$$\text{subw}(Q, \text{Stats}) = \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T} \in \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k\}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))$$

$$= \max_{t_1 \in \texttt{Nodes}(T_1), \ldots, t_k \in \texttt{Nodes}(T_k)} \max_{\boldsymbol{h} \models \text{Stats}} (\min(h(\chi_1(t_1)), \ldots, h(\chi_k(t_k))))$$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

$$\text{subw}(Q, \text{Stats}) = \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T} \in \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k\}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))$$

$$= \max_{t_1 \in \texttt{Nodes}(T_1), \ldots, t_k \in \texttt{Nodes}(T_k)} \underbrace{\max_{\boldsymbol{h} \models \text{Stats}} (\min(h(\chi_1(t_1)), \ldots, h(\chi_k(t_k))))}_{b(\Sigma, \text{Stats})}$$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

$$\text{subw}(Q, \text{Stats}) = \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T} \in \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k\}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))$$

$$= \max_{t_1 \in \texttt{Nodes}(T_1), \ldots, t_k \in \texttt{Nodes}(T_k)} \underbrace{\max_{\boldsymbol{h} \models \text{Stats}} (\min(h(\chi_1(t_1)), \ldots, h(\chi_k(t_k))))}_{b(\Sigma, \text{Stats})}$$

DDR: $\Sigma : \boxed{B_1(\chi(t_1)) \vee \cdots \vee B_k(\chi(t_k)) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

PANDA

Boolean CQ $\boxed{Q() = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

$$\text{subw}(Q, \text{Stats}) = \max_{\boldsymbol{h} \models \text{Stats}} \min_{\boldsymbol{T} \in \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k\}} \max_{t \in \texttt{Nodes}(T)} h(\chi(t))$$

$$= \max_{t_1 \in \texttt{Nodes}(T_1), \ldots, t_k \in \texttt{Nodes}(T_k)} \underbrace{\max_{\boldsymbol{h} \models \text{Stats}} (\min(h(\chi_1(t_1)), \ldots, h(\chi_k(t_k))))}_{b(\Sigma, \text{Stats})}$$

DDR: $\Sigma : \boxed{B_1(\chi(t_1)) \vee \cdots \vee B_k(\chi(t_k)) \leftarrow R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)}$

Compute bags $B_1, \ldots, B_k$ in time $\tilde{O}(2^{b(\Sigma, \text{Stats})})$, repeat for all $(t_1, \ldots, t_k)$.

Run Yannakakis' algorithm on each tree T_i. Runtime: $\tilde{O}(2^{\text{subw}(Q, \text{Stats})})$

PANDA: Discussion

Positives:

- Runtime $\tilde{O}(2^{\mathsf{subw}})$
- Steps of the algorithm entirely guided by solving an LP system.
- Generalizes to ω-subw [Khamis et al., 2024a].

PANDA: Discussion

Positives:

- Runtime $\tilde{O}(2^{\text{subw}})$
- Steps of the algorithm entirely guided by solving an LP system.
- Generalizes to ω-subw [Khamis et al., 2024a].

Negatives:

- Polylog factor. But is it needed???
- Computing subw is really hard!!
- Counting queries, sum-products: require restriction [Khamis et al., 2019]

Lower Bound Techniques

Overview

Is $O(2^{\text{subw}})$ optimal? Two questions:

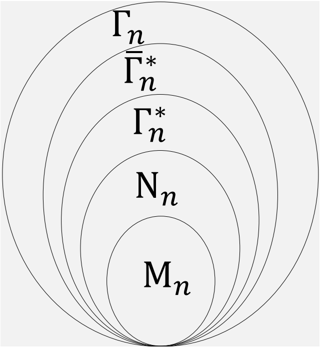
- Is the polymatroid bound of a full CQ tight?
- Do some accepted hypothesis imply a runtime of $\Omega(2^{\text{subw}})$?

Neither question has a definitive answer, but we have some insights.

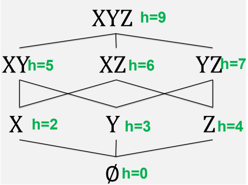
Types of Polymatroids

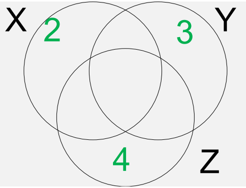
Polymatroids with n variables $\boldsymbol{X}$:

- Polymatroids Γ_n: Shannon inequalities
- Almost entropic $\bar{\Gamma}_n^*$: closure of Γ_n^*
- Entropic Γ_n^*: joint random variables
- Normal N_n: $h(\boldsymbol{U}) = \mu(f(\boldsymbol{U}))$ where
  Set function: $f : \boldsymbol{X} \rightarrow 2^\Omega$
  Set measure: $\mu : \Omega \rightarrow \mathbb{R}_p$
- Modular M_n: $\Omega = \boldsymbol{X}$, $f =$ identity.



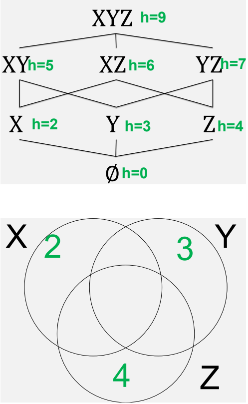
Examples



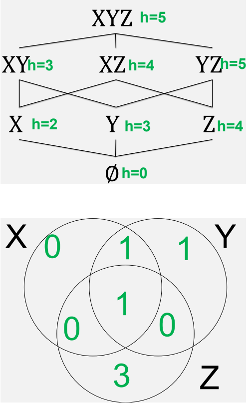


Modular M_3

Examples

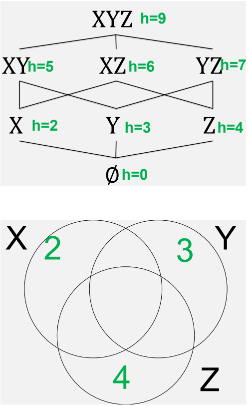


Modular M_3

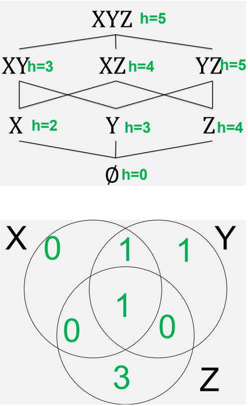


Normal N_3

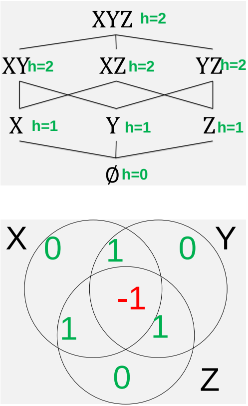
Examples



Modular M_3



Normal N_3



Entropic Γ_3^*

Tightness of the Polymatroid Bound for Full CQ

- Cardinality stats: $\boldsymbol{h}^* \in M_n$ and the AGM bound is tight.
  $\exists \boldsymbol{D}$ s.t. $|Q(\boldsymbol{D})| \geq \frac{1}{2^n} AGM(Q)$. [Atserias et al., 2013]

- Add FDs: there is a query for which $\boldsymbol{h}^* \in \Gamma_n - \bar{\Gamma}_n^*$. Not tight.
  [Khamis et al., 2017, Gogacz and Torunczyk, 2017, Suciu, 2023]

- Add "simple" degree constraints: $\boldsymbol{h}^* \in N_n$: bound is tight.
  $\exists \boldsymbol{D}$, $|Q(\boldsymbol{D})| \geq \frac{1}{2^{2^n-1}} 2^{h^*(\boldsymbol{X})}$ [Khamis et al., 2021, Suciu, 2023]:

Lower Bounds [Fan et al., 2023]

Conjecture (Boolean Clique Conjecture)

$\forall \varepsilon > 0$, no algorithm can check for a k-clique in time $O(n^{1-\varepsilon})$.

[3] μ is a constant function and (1) $\forall \omega \in \Omega$, $\{Z \mid \omega \in f(Z)\}$ is connected in Q, (2) $\forall \omega_1, \omega_2 \in \Omega$, $\exists j$, $\exists Y_1, Y_2 \in \boldsymbol{U}_j$ s.t. $\omega_i \in f(Y_i)$, $i = 1, 2$.

Lower Bounds [Fan et al., 2023]

Conjecture (Boolean Clique Conjecture)

$\forall \varepsilon > 0$, no algorithm can check for a k-clique in time $O(n^{1-\varepsilon})$.

Boolean CQ $Q() = \bigwedge_j R_j(\boldsymbol{U}_j)$

An embedding of Q is $h = \mu \circ f$, satisfying some syntactic conditions[3]

Embedding power: $\boxed{\text{emb}(Q) \stackrel{\text{def}}{=} \text{subw}(Q)\text{-restricted to an embedding}}$

[3] μ is a constant function and (1) $\forall \omega \in \Omega$, $\{Z \mid \omega \in f(Z)\}$ is connected in Q, (2) $\forall \omega_1, \omega_2 \in \Omega$, $\exists j$, $\exists Y_1, Y_2 \in \boldsymbol{U}_j$ s.t. $\omega_i \in f(Y_i)$, $i = 1, 2$.

Lower Bounds [Fan et al., 2023]

Conjecture (Boolean Clique Conjecture)

$\forall \varepsilon > 0$, no algorithm can check for a k-clique in time $O(n^{1-\varepsilon})$.

Boolean CQ $Q() = \bigwedge_j R_j(\boldsymbol{U}_j)$

An embedding of Q is $h = \mu \circ f$, satisfying some syntactic conditions[3]

Embedding power: $\boxed{\text{emb}(Q) \stackrel{\text{def}}{=} \text{subw}(Q)\text{-restricted to an embedding}}$

Theorem ([Fan et al., 2023])

$\forall \varepsilon > 0$, *no algorithm can compute* $Q(\boldsymbol{D})$ *in time* $O(2^{emb(Q)(1-\varepsilon)})$.

[3] μ is a constant function and (1) $\forall \omega \in \Omega$, $\{Z \mid \omega \in f(Z)\}$ is connected in Q, (2) $\forall \omega_1, \omega_2 \in \Omega$, $\exists j$, $\exists Y_1, Y_2 \in \boldsymbol{U}_j$ s.t. $\omega_i \in f(Y_i)$, $i = 1, 2$.

Discussion

Combinatorial Algorithms: $Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{\text{subw}})$

- Proofs to algorithms, Disjunctive Datalog Rules

  Open problem Characterize divergent inequalities.

- Lower bounds: embeddings [Fan et al., 2023]; circuit lower bounds [Fan et al., 2024].

Discussion

Combinatorial Algorithms: $Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{\text{subw}})$

- Proofs to algorithms, Disjunctive Datalog Rules

  Open problem Characterize divergent inequalities.

- Lower bounds: embeddings [Fan et al., 2023]; circuit lower bounds [Fan et al., 2024].

Using FMM: $Q(\boldsymbol{D})$ can be computed in time $\tilde{O}(2^{\omega\text{-subw}})$ [Khamis et al., 2024a].

- Triangle: $O(N^{\frac{2\omega}{\omega+1}})$
- 4-Clique: $O(N^{\frac{\omega+1}{2}})$
- 4-Cycle: $O(N^{\frac{4\omega-1}{2\omega+1}})$
- k-pyramid: $\tilde{O}(N^{2-\frac{2}{\omega(k-1)-k+3}})$ new result

Alon, N., Yuster, R., and Zwick, U. (1997).
Finding and counting given length cycles.
Algorithmica, 17(3):209–223.

Atserias, A., Grohe, M., and Marx, D. (2013).
Size bounds and query plans for relational joins.
SIAM J. Comput., 42(4):1737–1767.

Chung, F. R. K., Graham, R. L., Frankl, P., and Shearer, J. B. (1986).
Some intersection theorems for ordered sets and graphs.
J. Comb. Theory, Ser. A, 43(1):23–37.

Deeds, K., Suciu, D., Balazinska, M., and Cai, W. (2023a).
Degree sequence bound for join cardinality estimation.
In Geerts, F. and Vandevoort, B., editors, *26th International Conference on Database Theory, ICDT 2023, March 28-31, 2023, Ioannina, Greece*, volume 255 of *LIPIcs*, pages 8:1–8:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Deeds, K. B., Suciu, D., and Balazinska, M. (2023b).
Safebound: A practical system for generating cardinality bounds.
Proc. ACM Manag. Data, 1(1):53:1–53:26.

Fan, A. Z., Koutris, P., and Zhao, H. (2023).
The fine-grained complexity of boolean conjunctive queries and sum-product problems.
In Etessami, K., Feige, U., and Puppis, G., editors, *50th International Colloquium on Automata, Languages, and Programming, ICALP 2023, July 10-14, 2023, Paderborn, Germany*, volume 261 of *LIPIcs*, pages 127:1–127:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Fan, A. Z., Koutris, P., and Zhao, H. (2024).
Tight bounds of circuits for sum-product queries.
Proc. ACM Manag. Data, 2(2):87.

Gogacz, T. and Torunczyk, S. (2017).
Entropy bounds for conjunctive queries with functional dependencies.

In Benedikt, M. and Orsi, G., editors, *20th International Conference on Database Theory, ICDT 2017, March 21-24, 2017, Venice, Italy*, volume 68 of *LIPIcs*, pages 15:1–15:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Gottlob, G., Lee, S. T., Valiant, G., and Valiant, P. (2012).
Size and treewidth bounds for conjunctive queries.
J. ACM, 59(3):16:1–16:35.

Grohe, M. and Marx, D. (2014).
Constraint solving via fractional edge covers.
ACM Trans. Algorithms, 11(1):4:1–4:20.

Khamis, M. A., Curtin, R. R., Moseley, B., Ngo, H. Q., Nguyen, X., Olteanu, D., and Schleich, M. (2019).
On functional aggregate queries with additive inequalities.
In Suciu, D., Skritek, S., and Koch, C., editors, *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 414–431. ACM.

Khamis, M. A., Hu, X., and Suciu, D. (2024a).
Fast matrix multiplication meets the subdmodular width.
CoRR, abs/2412.06189.

Khamis, M. A., Kolaitis, P. G., Ngo, H. Q., and Suciu, D. (2021).
Bag query containment and information theory.
ACM Trans. Database Syst., 46(3):12:1–12:39.

Khamis, M. A., Nakos, V., Olteanu, D., and Suciu, D. (2024b).
Join size bounds using l_p-norms on degree sequences.
Proc. ACM Manag. Data, 2(2):96.

Khamis, M. A., Ngo, H. Q., and Rudra, A. (2016a).
FAQ: questions asked frequently.
In Milo, T. and Tan, W., editors, *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 13–28. ACM.

Khamis, M. A., Ngo, H. Q., and Suciu, D. (2016b).
Computing join queries with functional dependencies.
In Milo, T. and Tan, W., editors, *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 327–342. ACM.

Khamis, M. A., Ngo, H. Q., and Suciu, D. (2017).
What do shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another?
In Sallinger, E., den Bussche, J. V., and Geerts, F., editors, *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 429–444. ACM. Extended version available at http://arxiv.org/abs/1612.02503.

Khamis, M. A., Ngo, H. Q., and Suciu, D. (2024c).
PANDA: query evaluation in submodular width.
CoRR, abs/2402.02001.

Marx, D. (2013).
Tractable hypergraph properties for constraint satisfaction and conjunctive queries.
J. ACM, 60(6):42:1–42:51.

Matús, F. (2007).
Infinitely many information inequalities.
In *IEEE International Symposium on Information Theory, ISIT 2007, Nice, France, June 24-29, 2007*, pages 41–44. IEEE.

Ngo, H. Q. (2022).
On an information theoretic approach to cardinality estimation (invited talk).
In Olteanu, D. and Vortmeier, N., editors, *25th International Conference on Database Theory, ICDT 2022, March 29 to April 1, 2022, Edinburgh, UK (Virtual Conference)*, volume 220 of *LIPIcs*, pages 1:1–1:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Pippenger, N. (1986).
What are the laws of information theory.
In *1986 Special Problems on Communication and Computation Conference*, pages 3–5.

Shannon, C. E. (1948).

A mathematical theory of communication.
Bell Syst. Tech. J., 27(3):379–423.

Suciu, D. (2023).
Applications of information inequalities to database theory problems.
In *LICS*, pages 1–30.

Yeung, R. W. (2008).
Information Theory and Network Coding.
Springer Publishing Company, Incorporated, 1 edition.

Zhang, H., Mayer, C., Khamis, M. A., Olteanu, D., and Suciu, D. (2025).
Lpbound: Pessimistic cardinality estimation using $_p$-norms of degree sequences.
CoRR, abs/2502.05912.

Zhang, Z. and Yeung, R. W. (1998).
On characterization of entropy function via information inequalities.
IEEE Transactions on Information Theory, 44(4):1440–1452.

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$,

Statistics are cardinality constraints: $|R_j| \leq N_j$, $j = 1, m$.

Fix any fractional edge cover $w_1, \ldots, w_m$.

Theorem (The AGM Bound [Atserias et al., 2013])

If $\boldsymbol{D}$ satisfies the cardinality constraints, then $|Q(\boldsymbol{D})| \leq \prod_j |N_j|^{w_j}$.

We will prove the AGM bound, then derive an algorithm from its proof.

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$, frac. edge cover $w_1, \ldots, w_m$.

Shearer's Inequality: $\sum_j w_j h(\boldsymbol{U}_j) \geq h(\boldsymbol{X})$

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$, frac. edge cover $w_1, \ldots, w_m$.

Shearer's Inequality: $\sum_j w_j h(\boldsymbol{U}_j) \geq h(\boldsymbol{X})$

If $X_1 \in \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) = h(X_1) + h(\boldsymbol{U}_j | X_1)$; If $X_1 \notin \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) \geq h(\boldsymbol{U}_j | X_1)$.

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$, frac. edge cover $w_1, \ldots, w_m$.

Shearer's Inequality: $\sum_j w_j h(\boldsymbol{U}_j) \geq h(\boldsymbol{X})$

If $X_1 \in \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) = h(X_1) + h(\boldsymbol{U}_j|X_1)$; If $X_1 \notin \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) \geq h(\boldsymbol{U}_j|X_1)$.

$$\sum w_j h(\boldsymbol{U}_j) \geq (\sum_{j:X_1 \in \boldsymbol{U}_j} w_j) h(X_1) + \sum_j w_j h(\boldsymbol{U}_j|X_1) \geq h(X_1) + h(\boldsymbol{X}|X_1) = h(\boldsymbol{X}$$

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$, frac. edge cover $w_1, \ldots, w_m$.

Shearer's Inequality: $\sum_j w_j h(\boldsymbol{U}_j) \geq h(\boldsymbol{X})$

If $X_1 \in \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) = h(X_1) + h(\boldsymbol{U}_j|X_1)$; If $X_1 \notin \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) \geq h(\boldsymbol{U}_j|X_1)$.

$$\sum w_j h(\boldsymbol{U}_j) \geq (\sum_{j:X_1 \in \boldsymbol{U}_j} w_j) h(X_1) + \sum_j w_j h(\boldsymbol{U}_j|X_1) \geq h(X_1) + h(\boldsymbol{X}|X_1) = h(\boldsymbol{X})$$

Generic Join

for $x_1 \in \text{Dom}(X_1)$
  compute $Q[\boldsymbol{X}|X_1 := x_1]$

Runtime: $\tilde{O}(\prod_j |R_j|^{w_j})$

For-Loops and the AGM Bound

Full CQ: $Q(\boldsymbol{X}) = R_1(\boldsymbol{U}_1) \wedge \cdots \wedge R_m(\boldsymbol{U}_m)$, frac. edge cover $w_1, \ldots, w_m$.

Shearer's Inequality: $\sum_j w_j h(\boldsymbol{U}_j) \geq h(\boldsymbol{X})$

If $X_1 \in \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) = h(X_1) + h(\boldsymbol{U}_j|X_1)$; If $X_1 \notin \boldsymbol{U}_j$, $h(\boldsymbol{U}_j) \geq h(\boldsymbol{U}_j|X_1)$.

$$\sum w_j h(\boldsymbol{U}_j) \geq (\sum_{j:X_1 \in \boldsymbol{U}_j} w_j) h(X_1) + \sum_j w_j h(\boldsymbol{U}_j|X_1) \geq h(X_1) + h(\boldsymbol{X}|X_1) = h(\boldsymbol{X})$$

Generic Join

for $x_1 \in \mathrm{Dom}(X_1)$
  compute $Q[\boldsymbol{X}|X_1 := x_1]$

Runtime: $\tilde{O}(\prod_j |R_j|^{w_j})$

E.g. $Q_\Delta(XYZ) = R(XY) \wedge S(YZ) \wedge T(ZX)$:

for $x \in R[X] \cap T[X]$
 for $y \in R[Y|x] \cap S[Y]$
  for $z \in S[Z|y] \cap T[Z|x]$
   output(x, y, z)

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

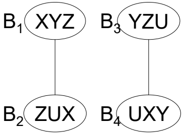
$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$

Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

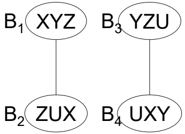
$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$

Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.



Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

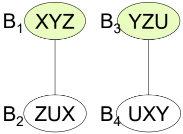
$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$



Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

Claim: $\boxed{\text{subw}(Q) = 3/2}$

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]



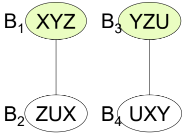
$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$

Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

Claim: $\boxed{\text{subw}(Q) = 3/2}$

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$



Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

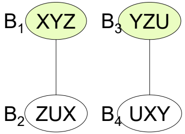
Claim: $\boxed{\text{subw}(Q) = 3/2}$

$$3 \geq$$

$$2\min(h(XYZ), h(YZU))$$

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$



Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

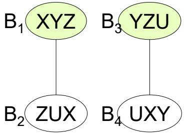
Claim: $\boxed{\text{subw}(Q) = 3/2}$

$$\begin{aligned} 3 \geq & h(XY) + h(YZ) + h(ZU) = h(XY) + h(Z|Y) + h(Y) + h(ZU) \\ \geq & h(XYZ) + h(YZU) \geq 2\min(h(XYZ), h(YZU)) \end{aligned}$$

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$

Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.



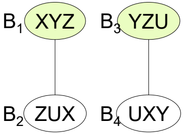
Claim: $\boxed{\text{subw}(Q) = 3/2}$

$$\begin{aligned}3 \geq & h(XY) + h(YZ) + h(ZU) = h(XY) + h(Z|Y) + h(Y) + h(ZU) \\ \geq & h(XYZ) + h(YZU) \geq 2\min(h(XYZ), h(YZU))\end{aligned}$$

$$\boxed{B_1(XYZ) \vee B_3(YZU) \leftarrow R(XY) \wedge S(YZ) \wedge T(ZU)}$$

Example: Detect 4-Cycle in Time $\tilde{O}(N^{3/2})$ [Alon et al., 1997]

$$\boxed{Q() = R(XY) \wedge S(YZ) \wedge T(ZU) \wedge K(UX)}$$



Cardinality constraints $|R|, |S|, |T|, |K| \leq N$.

Claim: $\boxed{\text{subw}(Q) = 3/2}$

$$\begin{aligned} 3 \geq & h(XY) + h(YZ) + h(ZU) = h(XY) + h(Z|Y) + h(Y) + h(ZU) \\ \geq & h(XYZ) + h(YZU) \geq 2\min(h(XYZ), h(YZU)) \end{aligned}$$

$$\boxed{B_1(XYZ) \vee B_3(YZU) \leftarrow R(XY) \wedge S(YZ) \wedge T(ZU)}$$

Compute in time $\tilde{O}(N^{3/2})$ (light/heavy), repeat for 4 DDRs.