RelationalAI

Logic & Algorithms ∈ DB & AI – Simons 2023

Output Cardinality Bounds and Information Theory

Hung Q. Ngo

Outline

Listing Triangles

In English

Given a (directed/undirected) graph G, find all triangles in G.

In Logic

$$Q(a,b,c) \leftarrow E(a,b) \wedge E(b,c) \wedge E(c,a)$$

$$Q(a,b,c) \leftarrow E(a,b) \wedge E(b,c) \wedge E(c,a) \wedge a < b \wedge b < c$$

Or, more generally:

$$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(c,a)$$

4-Cycle Detection

In English

Given a graph G, does it have a 4-cycle?

In Logic

$$Q() \leftarrow \exists a, b, c, d \quad E(a, b) \wedge E(b, c) \wedge E(c, d) \wedge E(d, a)$$

Counting 3-Walks

In English

Given a graph G, how many 3-walks are there in G?

In Sum-Product Form

$$Q() = \sum_{a,b,c,d} E(a,b) \cdot E(b,c) \cdot E(c,d)$$

Counting 3-Paths

In English

Given a graph G, how many 3-paths are there in G?

In Sum-Product Form

$$Q() = \sum_{a,b,c,d} E(a,b) \cdot E(b,c) \cdot E(c,d) \cdot \mathbf{1}_{a \neq b} \cdot \mathbf{1}_{a \neq c} \cdot \mathbf{1}_{a \neq d} \cdot \mathbf{1}_{b \neq c} \cdot \mathbf{1}_{b \neq d} \cdot \mathbf{1}_{c \neq d}$$

All-Pairs Shortest Paths (APSP)

In English

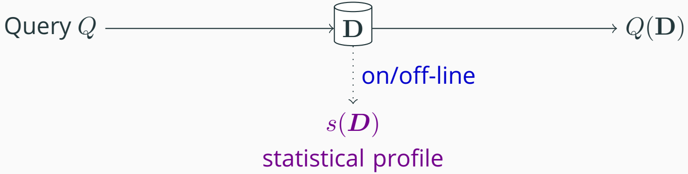
Given a graph G, compute the shortest path lengths between every pair of vertices.

In Datalogo (recursive query!)

$$Q[x,y] = \min\left(E[x,y],\ \min_z \{Q[x,z] + E[z,y]\}\right) \qquad \text{linear-form}$$

$$Q[x,y] = \min\left(E[x,y],\ \min_z \left\{Q[x,z] + Q[z,y]\right\}\right) \qquad \text{binary-form}$$

The Main Query Optimization / Evaluation Problem



Given Q and $\mathbf{D}$, compute $Q(\mathbf{D})$ in the most efficient (optimal!?) way possible.

Incremental View Maintenance (IVM) (a.k.a. Dynamic Algorithms)

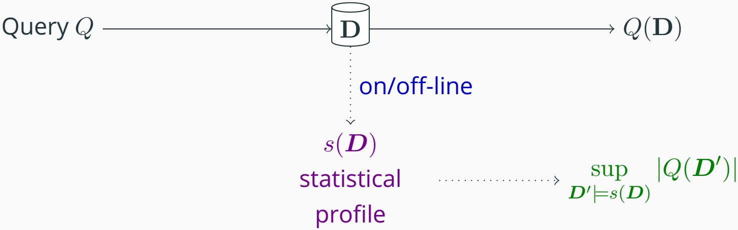
Given an update to $\boldsymbol{D}$, how do we update $Q(\boldsymbol{D})$ efficiently?

Precise Problem Formulation

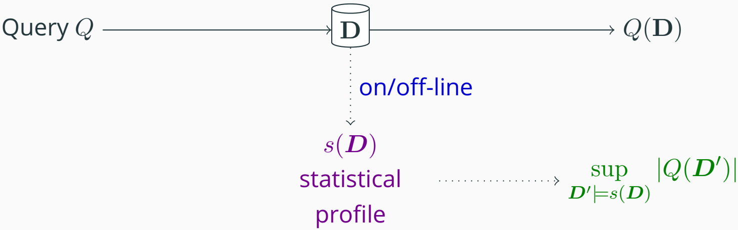
- What do you mean by "query"?
  - Full conjunctive queries
  - Sum-product queries
  - $\cdots$ First-Order, Second-Order, Rank-Enumeration
- What is in the statistical profile $s(\boldsymbol{D})$?
  - Degree constraints
  - $\cdots$ Frequency moment constraints, histograms, samples, ML models
- What do you mean by "optimality"?
  - Worst-Case Optimality
  - Instance Optimality
  - $\cdots$ Fine-grained complexity.
- Optimizer designed to work before seeing Q.
  - Optimizer = *Meta-Algorithm* (input: problem, output: algorithm)
  - Tutorial on 3 meta-algorithms: join, variable elimination, tensor decomposition

Outline

Worst-Case Cardinality Bound



Worst-Case Optimal Join (WCOJ) Algorithm



Definition

A “worst-case optimal” join algorithm is an algorithm computing $Q(\boldsymbol{D})$ in time

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

$\tilde{O}$ hides $\log$ and query dependent factors

(For Now) Q is a Full Conjunctive Queries

In a movie database

```
Q(director, actor, movie, actor_age, name) ←
     parent(director, actor)
     ∧ acted_in(actor, movie)
     ∧ director_of(director, movie)
     ∧ age(actor, actor_age) ∧ (20 < actor_age ∨ actor_age != 10)
     ∧ person_name(director, name) ∧ regex_match(".*spiel.*", name)
```

In a graph database with edge relation E,

```
Q(a, b, c) ← E(a, b) ∧ E(a, c) ∧ E(b, c)
```

(For Now) Q is a Full Conjunctive Queries

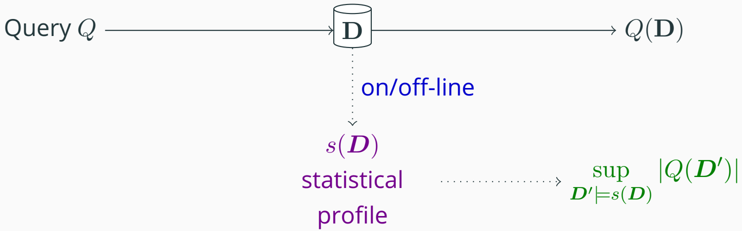
More generally, $\mathcal{H} = (V, \mathcal{E})$ is the hypergraph of a query:

$$Q(\boldsymbol{X}_V) \leftarrow \bigwedge_{S \in \mathcal{E}} R_S(\boldsymbol{X}_S)$$

For example `Q(a, b, c)` $\leftarrow$ `E(a, b)` $\wedge$ `E(a, c)` $\wedge$ `E(b, c)`

- $V = \{a, b, c\}$
- $\mathcal{H} = (V, \mathcal{E}) = (V, \{ab, ac, bc\})$
- $R_F = E$ for all $F \in \mathcal{E}$.

What is in the Statistical Profile $s(D)$?



Degree constraints

Degree Constraints

[ANS 16, ANS 17]

Relation R(actor, movie, role), imagine a frequency vector $\boldsymbol{d}_{\text{actor}}$ (billions of entries)

actor	movie	role
alice		
bob		
bob		
bob		
bob		
carol		
carol		

$d_{\text{actor}}(\text{alice}) = 1 \qquad d_{\text{actor}}(\text{bob}) = 4$

$d_{\text{actor}}(\text{carol}) = 2$

$d_{\text{actor}}(v) = 0 \qquad v \notin \{\text{alice}, \text{bob}, \text{carol}\}$

The profile $s(\boldsymbol{D})$ contains degree constraints:

- $\|\boldsymbol{d}_{\text{actor}}\|_\infty = 4$ (degree constraint!)
- $\|\boldsymbol{d}_{\emptyset}\|_\infty = 7 = |R|$ (cardinality constraint!)
- $\|\boldsymbol{d}_{\text{actor},\text{movie}}\|_\infty = 1$ (functional dependency)

General DC : (X, Y, N) in relation R means $|\pi_Y \sigma_{X=\boldsymbol{x}} R| \leq N, \ \forall \boldsymbol{x}$

Outline

Hierarchy of Set Functions

$h : 2^{[n]} \to \mathbb{R}_+$, non-negative, monotone, $h(\emptyset) = 0$, $h(X) \leq h(Y)$ if $X \subseteq Y$

SA$_n$ $:= \{h \mid h$ is sub-additive$\}$ $h(X \cup Y) \leq h(X) + h(Y)$

Γ_n $:= \{h \mid h$ is submodular$\}$ = polymatroids

$h(X \cup Y) + h(X \cap Y) \leq h(X) + h(Y)$

$\overline{\Gamma}^*_n$: topological closure of Γ^*_n, *almost entropic*

$\Gamma^*_n = \{h : h$ is entropic$\}$

N_n : Normal convex-hull of step functions

(weighted coverage functions)

(non-negative multivariate mutual information)

M_n : Modular $h(X) = \sum_{x \in X} h(x)$

$$
\begin{aligned}
\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| &= \text{combinatorial-bound}(Q, s) \\
&\leq \text{entropic-bound}(Q, s) \\
&\leq \boxed{\text{polymatroid-bound}(Q, s)} \\
&\leq \text{flow-bound}(Q, s, \sigma) \\
&\leq \text{chain-bound}(Q, s, \sigma) \\
&\leq \text{agm-bound}(Q, s) \\
&\leq \text{integral-edge-cover}(Q, s)
\end{aligned}
$$

computable but impractical

- $Q(a,b,c) = R(a,b) \wedge S(b,c) \wedge T(a,c)$
- $\boldsymbol{D} = \{R, S, T\}$
- $s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$

$|R|, |S|, |T|$ are integers, in the order of 10^9 or more

$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(a,c)$

$s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$

$$\begin{aligned}
\max \quad & \sum_{a,b,c} \mathbf{1}_{R(a,b)} \mathbf{1}_{S(a,c)} \mathbf{1}_{T(b,c)} \\
\text{s.t} \quad & \sum_{a,b} \mathbf{1}_{R(a,b)} \leq |R| \\
& \sum_{b,c} \mathbf{1}_{S(b,c)} \leq |S| \\
& \sum_{a,c} \mathbf{1}_{T(a,c)} \leq |T|.
\end{aligned}$$

$\mathbf{1}_X \in \{0,1\}$

Can turn this into a *linear integer program*, but does not help much.

Entropy Argument [CGFS 86]

$$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(a,c) \qquad s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$$

- Fix a *worst-case* input R, S, T. Select tuples $(a,b,c) \in Q$ uniformly at random
- Let h be the entropy function of this 3D-distribution, then

$$\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| = h(a,b,c)$$

$$h(a,b) \le \log |R|$$

$$h(b,c) \le \log |S|$$

$$h(a,c) \le \log |T|$$

$$h \in \Gamma_3^*$$

h is entropic

$$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(a,c)$$

$$s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$$

$$\begin{aligned}
\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \leq \max \quad & h(a,b,c) \\
\text{s.t.} \quad & h(a,b) \leq \log|R|, \\
& h(b,c) \leq \log|S|, \\
& h(a,c) \leq \log|T|, \\
& h \in \Gamma_3^*
\end{aligned}$$

h is entropic

$$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(a,c) \qquad s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$$

$$\begin{aligned}
\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \leq \max \quad & h(a,b,c) \\
\text{s.t.} \quad & h(a,b) \leq \log |R|, \\
& h(b,c) \leq \log |S|, \\
& h(a,c) \leq \log |T|, \\
& h \in \boxed{\Gamma_3} \qquad \qquad h \text{ is a polymatroid}
\end{aligned}$$

$$\max\{h(a,b,c) \mid h(a,b) \leq \log|R|, h(b,c) \leq \log|S|, h(a,c) \leq \log|T|, h \in \Gamma_3\}$$

Define a modular $g \in \mathrm{M}_3$ as follows, then g satisfies all constraints:

$$\begin{aligned}
g(a) &= h(a) & g(ab) &= g(a) + g(b) & &= h(ab) \\
g(b) &= h(b|a) & g(bc) &= g(b) + g(c) = h(abc) - h(a) & &\leq h(bc) \\
g(c) &= h(c|ab) & g(ac) &= g(a) + g(c) = h(abc) + h(a) - h(ab) & &\leq h(ac) \\
& & g(abc) &= h(abc) & &
\end{aligned}$$

Problem can be reformulated, optimized over *modular* functions g:

$$\begin{aligned}
\max \quad & g(a) + g(b) + g(c) \\
& g(a) + g(b) \leq \log|R|, \quad g(b) + g(c) \leq \log|S|, \quad g(a) + g(c) \leq \log|T|, \\
& g(a), g(b), g(c) \geq 0
\end{aligned}$$

Vertex packing LP

$$\begin{aligned} \max \quad & g(a) + g(b) + g(c) \\ & g(a) + g(b) \leq \log|R|, \qquad g(b) + g(c) \leq \log|S|, \qquad g(a) + g(c) \leq \log|T|, \\ & g(a), g(b), g(c) \geq 0 \end{aligned}$$

Fractional edge cover LP

[AGM 2008] took the dual of the above LP:

$$\begin{aligned} \min \quad & \lambda_{ab} \log|R| + \lambda_{bc} \log|S| + \lambda_{ac} \log|T| \\ & \lambda_{ab} + \lambda_{ac} \geq 1 \\ & \lambda_{ab} + \lambda_{bc} \geq 1 \\ & \lambda_{bc} + \lambda_{ac} \geq 1 \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

$$Q(a,b,c) \leftarrow R(a,b) \wedge S(b,c) \wedge T(a,c)$$

$$s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$$

$$\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \leq \log \min\{|R| \cdot |S|, |R| \cdot |T|, |S| \cdot |T|\}$$

We used the example to illustrate the following bounds / concepts:

$$
\begin{aligned}
\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| &= \text{combinatorial-bound}(Q, s) \quad \text{(computable but impractical)} \\
&\leq \text{entropic-bound}(Q, s) \\
&= \text{polymatroid-bound}(Q, s) \\
&= \text{modular-bound}(Q, s) \\
&= \text{agm-bound}(Q, s) \\
&\leq \text{integral-edge-cover}(Q, s).
\end{aligned}
$$

Let g be an optimal solution to the modular bound:

$$
\begin{aligned}
\max \quad & g(a) + g(b) + g(c) \\
& g(a) + g(b) \le \log|R|, \qquad g(b) + g(c) \le \log|S|, \qquad g(a) + g(c) \le \log|T|, \\
& g(a), g(b), g(c) \ge 0
\end{aligned}
$$

Then, $\sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \le$ modular-bound $= 2^{g(a)+g(b)+g(c)} = 2^{g(a)} \times 2^{g(b)} \times 2^{g(c)}$

Construct a database instance $\boldsymbol{D}'$ [AGM 08]

- $R = [\lfloor 2^{g(a)} \rfloor] \times [\lfloor 2^{g(b)} \rfloor]$, $S = [\lfloor 2^{g(b)} \rfloor] \times [\lfloor 2^{g(c)} \rfloor]$, $T = [\lfloor 2^{g(a)} \rfloor] \times [\lfloor 2^{g(c)} \rfloor]$
- Then $\boldsymbol{D}' \models s(\boldsymbol{D})$ *and* $|Q| \ge \frac{1}{8} 2^{g(a)+g(b)+g(c)} = \Omega(\sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')|)$

More Generally

- Given a set of *degree constraints* (DCs)
  - Triples (X, Y, N) which says, for each $\boldsymbol{x}$ there are at most N $\boldsymbol{y}$'s
  - If $X = \emptyset$, then this is a *cardinality constraint* (CC) (e.g. *distinct counts*)
  - If $N = 1$, then this is a *functional dependency* (FD) (very common)
- Entropy argument implies

$$\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \leq \text{entropic-bound}$$

The Entropic and Polymatroid Bounds

Theorem (ANS 17)

If $s(\boldsymbol{D})$ contains only degree constraints, then

$$\log \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \leq \sup_{h \in \overline{\Gamma}^*_n \cap DC} h(V) \leq \boxed{\max_{h \in \Gamma_n \cap DC} h(V)}$$

where DC is the set of linear constraints of the form

$$h(Y|X) \leq \log N$$

for each degree constraint (X, Y, N).

Example: the Triangle Query

- $R(a,b) \wedge S(b,c) \wedge T(a,c)$ $\quad\quad$ $\boldsymbol{D} = \{R, S, T\}$
- $s(\boldsymbol{D}) = \{|R|, |S|, |T|\}$
- Constraint set:

$$\mathsf{DC} = \{h \mid h(ab) \leq \log |R| \wedge h(bc) \leq \log |S| \wedge h(ac) \leq \log |T|\}$$

- Polymatroid bound:

$$\max\{h(abc) \mid h \in \Gamma_3 \cap \mathsf{DC}\} = \log \min\{|R| \cdot |S|, |S| \cdot |T|, |T| \cdot |R|, \sqrt{|R| \cdot |S| \cdot |T|}\}$$

e.g. if $|R|, |S|, |T| = N$, then $|Q| \leq N^{3/2}$ (Loomis-Whitney inequality!)

Example: the Triangle Query with Extra FD Information

- $R(a, b) \wedge S(b, c) \wedge T(a, c)$ $\qquad$ $\boldsymbol{D} = \{R, S, T\}$
- $s(\boldsymbol{D}) = \{|R|, |S|, |T|, b \to c\}$ (b is a key in S)
- Constraint set:

$$\text{DC} = \{h \mid h(ab) \leq \log|R| \wedge h(bc) \leq \log|S| \wedge h(ac) \leq \log|T| \wedge h(c|b) = 0\}$$

- Polymatroid bound:

$$\max\{h(abc) \mid h \in \Gamma_3 \cap \text{DC}\} = \log \min\{|R|, |S| \cdot |T|\}$$

e.g. $|R|, |S|, |T| = N$, then $|Q| \leq N$

Example: Builtins and FDs

- $R(a) \wedge S(b) \wedge a + b = 5$ $\quad\quad \boldsymbol{D} = \{R, S\}$
- $s(\boldsymbol{D}) = \{|R|, |S|, a \rightarrow b, b \rightarrow a\}$
- Constraint set:

$$\text{DC} = \{h(a) \leq \log |R| \wedge h(b) \leq \log |S| \wedge h(a|b) = h(b|a) = 0\}$$

- Polymatroid bound:

$$\max\{h(ab) \mid h \in \Gamma_2 \cap \text{DC}\} = \log \min\{|R|, |S|\}$$

Example: A Non-Trivial Bound

- $R(a, b) \wedge S(b, c) \wedge T(c, d) \wedge f_1(a, c) = d \wedge f_2(b, d) = a$ $\quad$ f_1, f_2 are UDFs
- $s(\boldsymbol{D}) = \{|R|, |S|, |T|, ac \to d, bd \to a\}$
- Constraint set:

$$\text{DC} = \{h \mid h(ab) \leq \log |R| \wedge h(bc) \leq \log |S| \wedge h(cd) \leq \log |T| \wedge h(d|ac) = h(a|bd) = 0\}$$

- Polymatroid bound:

$$\max\{h(abcd) \mid h \in \Gamma_4 \cap \text{DC}\} = \log \min\{|R| \cdot |S|, |S| \cdot |T|, |T| \cdot |R|, \sqrt{|R| \cdot |S| \cdot |T|}\}$$

Outline

Main Classes of (Mostly) Open Problems

1. Is the entropic bound computable?
2. Is the polymatroid bound computable in PTIME?
3. Find classes of inputs where the polymatroid bound is computable in PTIME.
4. When is which bound asymptotically tight?
5. Approximate the bounds efficiently. Hardness of approximation.
6. Going beyond conjunctive queries?
7. Dealing with more constraints
    - Conditional independence constraints
    - Frequency moment constraints
    - Histogram constraints
    - ...

Hierarchy of Set Functions

$h : 2^{[n]} \to \mathbb{R}_+$, non-negative, monotone, $h(\emptyset) = 0$, $h(X) \leq h(Y)$ if $X \subseteq Y$

SA$_n$ $:= \{h \mid h$ is sub-additive$\}$ $\quad h(X \cup Y) \leq h(X) + h(Y)$

$\Gamma_n := \{h \mid h$ is submodular$\}$ = polymatroids

$$h(X \cup Y) + h(X \cap Y) \leq h(X) + h(Y)$$

$\overline{\Gamma}^*_n$: topological closure of Γ^*_n, *almost entropic*

$\Gamma^*_n = \{h : h$ is entropic$\}$

N_n : Normal $\quad$ convex-hull of step functions

(weighted coverage functions)

(non-negative multivariate mutual information)

M_n : Modular $\quad h(X) = \sum_{x \in X} h(x)$

A Collection of Optimization Problems

Many bounds can be formulated with two parameters

$$\sup\{h(V) \mid h \in P \cap \mathsf{DC}\}$$

- DC $=$ set of constraints $h(Y|X) \leq \log N$, one for degree constraint (X, Y, N).
- P is a member in the aforementioned hierarchy of set functions
  - $P = \mathsf{M}_n$: *modular bound*
  - $P = \mathsf{N}_n$: *normal bound*
  - $P = \overline{\Gamma}_n^*$: *entropic bound*
  - $P = \Gamma_n$: *polymatroid bound*

Bound Hierarchy

ANS 2017, IMNPS 2023

Integral Edge Cover B.

↑

AGM B.

↑

Chain B.

↑

Flow B.

↑

Polymatroid B.

↑

Entropic B.

Normal B. - - - (asymptotic) - - → Combinatorial B.

↑

Modular B.

$\xrightarrow{\leq}$

$\overset{O(\cdot)}{\dashrightarrow}$
dimensionality = $O(1)$

Define $\boldsymbol{c} = (c_X)_{X \subseteq V}$, where $c_X = -1_{X=V}$, then – because C is linear –

$$\sup\{h(V) \mid h \in C \cap \overline{\Gamma}^*_n\} = \inf\{\langle \boldsymbol{c}, \boldsymbol{h}\rangle \mid \boldsymbol{A}\boldsymbol{h} \leq \boldsymbol{b} \wedge \boldsymbol{h} \in \overline{\Gamma}^*_n\}$$

Lagrangian: $\mathcal{L}(\boldsymbol{\delta}) = \inf_{\boldsymbol{h} \in \overline{\Gamma}^*_n} \langle \boldsymbol{c}, \boldsymbol{h}\rangle + \langle \boldsymbol{A}\boldsymbol{h} - \boldsymbol{b}, \boldsymbol{\delta}\rangle = -\langle \boldsymbol{b}, \boldsymbol{\delta}\rangle + \inf_{\boldsymbol{h} \in \overline{\Gamma}^*_n} \langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h}\rangle$

Lagrangian dual problem

$(\overline{\Gamma}^*_n)^*$ denotes the dual cone of $\overline{\Gamma}^*_n$

$$\sup\{\mathcal{L}(\boldsymbol{\delta}) \mid \boldsymbol{\delta} \geq \boldsymbol{0}\} = \inf\{\langle \boldsymbol{b}, \boldsymbol{\delta}\rangle \mid \boldsymbol{\delta} \geq \boldsymbol{0} \wedge \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta} \in (\overline{\Gamma}^*_n)^*\}$$

Checking whether $\boldsymbol{\delta} \geq \boldsymbol{0}$ is dual feasible is equivalent to verifying whether

$$h(V) \leq \sum_{(X,Y) \in \mathsf{DC}} \delta_{Y|X} h(Y|X) \qquad \forall h \in \overline{\Gamma}^*_n$$

2. Computational Complexity of Polymatroid Bound

NP-hard	integral-edge-cover-bound(Q, s)	
	$\uparrow$	
PTime	agm-bound(Q, s)	
	$\uparrow$	
PTime	chain-bound(Q, s, σ)	NP-hard to find best σ
	$\uparrow$	
PTime	flow-bound(Q, s, σ)	NP-hard to find best σ
	$\uparrow$	
No idea	polymatroid-bound(Q, s)	
	$\uparrow$	
Not computable?	entropic-bound(Q, s)	
NP-hard	normal-bound(Q, s) $\xrightarrow{\text{(asymptotic)}}$ combinatorial-bound(Q, s)	Impractical
	$\uparrow$	
PTime	modular-bound(Q, s)	

3. Parameterized Complexity of Polymatroid Bound

The polymatroid bound is computable in PTIME for some classes of inputs:

- Acyclic degree constraints
- Simple degree constraints
- Degree constraints with bounded SCCs

4. Tightness of Various Bounds

- For which class of queries does adding conditional independence constraints improves the bound?
- How close can we get to the entropic bound? (Ignore the combinatorial bound)
- How close we can get to combinatorial bound, under which condition?

Theorem

Except for the combinatorial $\rightarrow$ entropic edge, there is an asymptotic gap in between every adjacent bound (connected by $\rightarrow$ in the hierarchy), even if the number of degree constraints is fixed.

4. Tightness of Various Bounds

The following is *unsatisfactory*:

Proposition (ANS 2017)

For any $\epsilon > 0$, there exists a scale-factor k such that, if all DCs (X, Y, N) are scaled up into (X, Y, N^k) then

$$\textit{entropic-bound} = (1 - \epsilon) \log \textit{combinatorial-bound}$$

(Made use of group-characterizable entropic functions)

Outline

References (and references thereof)

- N. Alon, On the number of subgraphs of prescribed type of graphs with a given number of edges, Israel Journal of Mathematics 38 (1981), 116–130.

CGFS 1986 Fan R. K. Chung, Ronald L. Graham, Peter Frankl, James B. Shearer: Some intersection theorems for ordered sets and graphs. J. Comb. Theory, Ser. A 43(1): 23-37 (1986)

- Ehud Friedgut, Jeff Kahn, On the number of copies of one hypergraph in another. Israel J. Math. 105 (1998), 251–256.
- Ehud Friedgut: Hypergraphs, Entropy, and Inequalities. Am. Math. Mon. 111(9): 749-760 (2004)

AGM 08 Albert Atserias, Martin Grohe, Dániel Marx: Size Bounds and Query Plans for Relational Joins. FOCS 2008: 739-748

GLVV 12 Georg Gottlob, Stephanie Tien Lee, Gregory Valiant, Paul Valiant: Size and Treewidth Bounds for Conjunctive Queries. J. ACM 59(3): 16:1-16:35 (2012)

ANS 16 Mahmoud Abo Khamis, Hung Q. Ngo, Dan Suciu: Computing Join Queries with Functional Dependencies. PODS 2016: 327-342

ANS 17 Mahmoud Abo Khamis, Hung Q. Ngo, Dan Suciu: What Do Shannon-type Inequalities, Submodular Width, and Disjunctive Datalog Have to Do with One Another? PODS 2017: 429-444

AKNS1 20 Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, Dan Suciu: Decision Problems in Information Theory. ICALP 2020: 106:1-106:20

AKNS2 20 Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, Dan Suciu: Bag Query Containment and Information Theory. PODS 2020: 95-112

- Hung Q. Ngo: On an Information Theoretic Approach to Cardinality Estimation (Invited Talk). ICDT 2022: 1:1-1:21
- Dan Suciu: Applications of Information Inequalities to Database Theory Problems. LICS 2023: 1-30

IMNPS 2023 Sungjin Im, Ben Moseley, Hung Q. Ngo, Kirk Pruhs, and Alireza Samadian: New Efficiently Computable Size Bounds for Joins. 2023.

Thank You!

Outline

Appendix

Outline

Appendix

3. Parameterlized Complexity of Polymatroid Bound

- A set of DCs are simple if $|X| \leq 1$ for all DCs (X, Y, N)
- A set of DCs are acyclic if the constraint dependency graph is acyclic
  - *Constraint dependency graph*: for every degree constraint (X, Y, D), add edges $x \to y$ for all $x \in X$, $y \in Y$.

3. Parameterlized Complexity of Polymatroid Bound

Proposition

There is a polymatroid-bound-preserving reduction from an arbitrary instance to an instance where the degree constraints are a union of two sets: DC $=$ DC$_a \cup$ DC$_s$, where DC$_s$ is simple, and DC$_a$ is acyclic. Furthermore, for every non-simple DC $(X, Y, N) \in$ DC$_a$, we have $|X| = 2$ and $|Y| \leq 3$.

Proposition

There is a poly-time algorithm computing a variable ordering σ_0 so that:

- *If all input DCs are simple, then flow-bound$(Q, s, \sigma_0) =$ polymatroid-bound(Q,s)*
- *If all input DCs are acyclic, then flow-bound$(Q, s, \sigma_0) =$ polymatroid-bound(Q,s)*

4. Tightness of Various Bounds

Proposition

If all DCs are acyclic, then there is a σ_0 for which

$$\textit{chain-bound}(Q, s, \sigma_0) = \textit{modular-bound}(Q, s).$$

All bounds in between collapse, and are all $\Theta(\textit{combinatorial bound})$ in data-complexity.

Proposition

If all DCs are simple, then there is a σ_0 for which

$$\textit{flow-bound}(Q, s, \sigma_0) = \textit{normal-bound}(Q, s).$$

All bounds in between collapse, and are all $\Theta(\textit{combinatorial bound})$ in data-complexity.

Outline

Appendix

Frequency-Moment Constraints

Relation R(actor, movie, role), imagine a frequency vector $\boldsymbol{d}_{\text{actor}}$ (billions of entries)

actor	movie	role
alice		
bob		
bob		
bob		
bob		
carol		
carol		

$$d_{\text{actor}}(\text{alice}) = 1$$
$$d_{\text{actor}}(\text{bob}) = 4$$
$$d_{\text{actor}}(\text{carol}) = 2$$
$$d_{\text{actor}}(v) = 0 \qquad v \notin \{\text{alice}, \text{bob}, \text{carol}\}$$

The (very small) profile $s(\boldsymbol{D})$ contains

- $\|\boldsymbol{d}_{\text{actor}}\|_\infty = 4$ (heaviest frequency)
- $\|\boldsymbol{d}_{\text{actor}}\|_0 = 3$ (distinct counts)
- $\|\boldsymbol{d}_{\text{actor}}\|_1 = 7 = |R|$

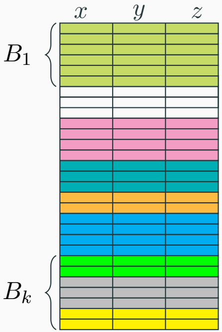
Similarly, we may have $\|\boldsymbol{d}_{\text{movie}}\|_p$, $\|\boldsymbol{d}_{\text{actor,movie}}\|_p$, $\|\boldsymbol{d}_{\text{role}}\|_p$, etc.

- Partition: $\mathrm{Dom}(x) = B_1 \cup B_2 \cup \cdots \cup B_k$ (x-histogram)
  - Typically about $k \approx 200$ buckets (e.g., MS SQL Server)
  - Top 10 heavy hitters are in their own singleton buckets
  - For the rest, equi-depth
- $\mathcal{B} := \{B_1, \ldots, B_k\}$
- Give rise to per-bucket frequency vectors:

$$\boldsymbol{d}_{Y|X\in B}(\boldsymbol{x}) := \begin{cases} |\pi_Y \sigma_{X=\boldsymbol{x}} R| & x \in B \\ 0 & \text{o.w.} \end{cases}$$

- HFM-Constraints $(\mathcal{B}, X, Y, \boldsymbol{c}, \ell, R)$

$$F_\ell(\boldsymbol{d}_{Y|X\in B}) \leq c_B \qquad B \in \mathcal{B}$$



Outline

Appendix

Some known results on the parametererized complexity of the polymatroid bound

Frequency Moment Constraints

Histograms

Histograms are More Intricate

Main Question

How to turn $s(\boldsymbol{D})$ into constraints on polymatroids h?

Answer sketch:

Theorem (KNN 2022, WIP)

Let Q be a conjunctive query and $\mathcal{C}$ be a given set of simple HFM-constraints, then for any database $\boldsymbol{D}$ satisfying $\mathcal{C}$, we have

$$\sup_{\boldsymbol{D} \models \mathcal{C}} \log |Q(\boldsymbol{D})| \leq \max\{h(\boldsymbol{V}) \mid h \in \overline{\Gamma}^*_{n+m}, (h, \boldsymbol{P}, \boldsymbol{C}) \in \text{HC}\} \quad \textit{(entropic bound)}$$

$$\leq \max\{h(\boldsymbol{V}) \mid h \in \Gamma_{n+m}, (h, \boldsymbol{P}, \boldsymbol{C}) \in \text{HC}\} \quad \textit{(polymatroid bound)}$$

- Let $\boldsymbol{f}_y$ and $\boldsymbol{g}_y$ be the y-frequency vectors in R and S, respectively
- Assume the same partition $\mathrm{dom}(y) = B_1 \cup \cdots \cup B_k$ on both sides
- Suppose $s(\boldsymbol{D})$ contains the following statistics:

$$r_B := \|\boldsymbol{f}_{y \in B}\|_\infty \qquad d_B := \|\boldsymbol{f}_{y \in B}\|_0 \qquad s_B := \|\boldsymbol{g}_{y \in B}\|_\infty \qquad B \in \mathcal{B}$$

  - r_B = maximum number of x per y in bucket B
  - d_B = number of distinct y's in B
  - s_B = maximum number of z per y in bucket B

Question: What are the constraints C on h?

Turning $s(D)$ into Constraints on h

- Consider the uniform distribution on (X, Y, Z) chosen from $R(x, y) \wedge S(y, z)$
- Let $J \in \mathcal{B}$ be a categorical random variable, where $J = B$ iff $Y \in B$

$$p_B := \text{Prob}[J = B]$$

- Then,

$$\begin{aligned}
h(J \mid Y) &= 0 && (1)\\
h(Y \mid J = B) &\leq \log d_B = \log \|\boldsymbol{f}_{y \in B}\|_0 && (2)\\
h(X \mid J = B) &\leq \log r_B = \log \|\boldsymbol{f}_{y \in B}\|_\infty && (3)\\
h(Z \mid J = B) &\leq \log s_B = \log \|\boldsymbol{g}_{y \in B}\|_\infty && (4)\\
h(J) &= -\langle \boldsymbol{p}, \log \boldsymbol{p} \rangle \qquad \|\boldsymbol{p}\|_1 = 1 \qquad \boldsymbol{p} \geq \boldsymbol{0} && (5)
\end{aligned}$$

Simplifying the constraints

$$h(J \mid Y) = 0 \tag{6}$$

$$h(Y \mid J) = \sum_B h(Y \mid J = B) \cdot p_B \leq \langle \log \boldsymbol{d}, \boldsymbol{p} \rangle \tag{7}$$

$$h(X \mid J) = \sum_B h(X \mid J = B) \cdot p_B \leq \langle \log \boldsymbol{r}, \boldsymbol{p} \rangle \tag{8}$$

$$h(Z \mid J) = \sum_B h(Z \mid J = B) \cdot p_B \leq \langle \log \boldsymbol{s}, \boldsymbol{p} \rangle \tag{9}$$

$$h(J) = -\langle \boldsymbol{p}, \log \boldsymbol{p} \rangle \tag{10}$$

$$\|\boldsymbol{p}\|_1 = 1 \tag{11}$$

$$\boldsymbol{p} \geq \boldsymbol{0} \tag{12}$$

The Estimator with Constraints on h — Concave Optimization Problem

$$
\begin{aligned}
\max \quad & h(XYZ) && (13)\\
\text{s.t.} \quad & h(Y \mid J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{d} \rangle && (14)\\
& h(X \mid J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{r} \rangle && (15)\\
& h(Z \mid J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{s} \rangle && (16)\\
& h \in \Gamma_4 \qquad \text{join distribution on } (X, Y, Z, J) && (17)\\
& \boldsymbol{p} \geq 0, && (18)\\
& h(J) = -\langle \boldsymbol{p}, \lg \boldsymbol{p} \rangle && (19)\\
& h(J \mid Y) = 0 && (20)\\
& \|\boldsymbol{p}\|_1 = 1. && (21)
\end{aligned}
$$

Sanity Check: Estimator Makes Sense Combinatorially!

$$\lg |Q| = h(XYZ) \tag{22}$$

$$(\text{since } H[JY] = H[Y]) = H[XYZJ] = H[XYZ|J] + H[J] \tag{23}$$

$$(\text{since } H \in \Gamma_4) \leq H[X|J] + H[Y|J] + H[Z|J] + H[J] \tag{24}$$

$$\leq \sum_{j \in \mathcal{B}} (\lg r_B + \lg d_B + \lg s_B - \lg p_B)) \cdot p_B \tag{25}$$

$$= \sum_{j \in \mathcal{B}} (\lg(r_B d_B s_B / p_B)) \cdot p_B \tag{26}$$

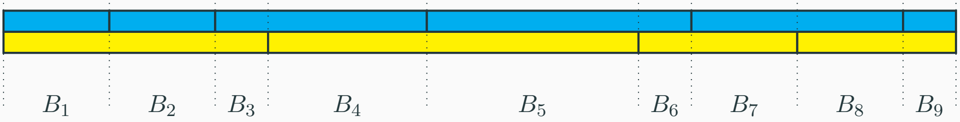
$$(\text{Jensen}) \leq \lg \left(\sum_B r_B d_B s_B \right). \tag{27}$$

$$|Q| \leq \sum_B r_B d_B s_B \tag{28}$$

Example – Removing the Simplifying Assumption

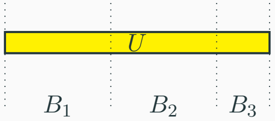
$$R(x, y) \wedge S(y, z)$$

- $\mathrm{dom}(y) = U_1 \cup \cdots \cup U_k$ on the R-side
- $\mathrm{dom}(y) = V_1 \cup \cdots \cup V_\ell$ on the S-side



Back to the "boundary aligned" model that $\mathrm{dom}(y) = B_1 \cup B_2 \cup \cdots \cup B_m$

Example – Removing the Simplifying Assumption



- For $q \in \{0, 1\}$ every constraint $\|\boldsymbol{f}_U^y\|_q$ is broken up into

$$\|\boldsymbol{f}_U^y\|_q = \|\boldsymbol{f}_{B_1}^y\|_q + \|\boldsymbol{f}_{B_2}^y\|_q + \|\boldsymbol{f}_{B_3}^y\|_q$$

  where $\|\boldsymbol{f}_{B_j}^y\|_q$ are *new variables* in the optimization problem
- For $q = \infty$, set $\|\boldsymbol{f}_{B_i}^y\|_\infty \leq \|\boldsymbol{f}_{B_i}^y\|_1$ (or Hölder-type)

$$\max\{\|\boldsymbol{f}_{B_1}^y\|_\infty, \|\boldsymbol{f}_{B_2}^y\|_\infty, \|\boldsymbol{f}_{B_3}^y\|_\infty\} \leq \|\boldsymbol{f}_U^y\|_\infty$$

Under histogrammed FM constraints

Repeat the 4 (classes of) questions

- Computability
- Complexity
- Parameterized Complexity
- Tightness