Finite Model Theory
Unit 5

Dan Suciu

Spring 2018

599c: Finite Model Theory

Unit 5: Algorithmic Aspects of FMT

The Problem

Given a query Q, and a structure (database) $\boldsymbol{D}$, what is the algorithmic complexity for computing $Q(\boldsymbol{D})$?

We are interested in data complexity only: Q is fixed, and the input is $\boldsymbol{D}$.

And we will consider only Conjunctive Queries: $\exists \boldsymbol{x}(R_1 \wedge R_2 \wedge \cdots)$.

The Problem

Suppose Q is in prenex normal form with k variables.
Suppose the domain size is $n = |D|$.
A naive algorithm computes $Q(\boldsymbol{D})$ in time $\tilde{O}(n^k)$. why the $\log n$ factor?

In general, we know the sizes of the input relations $|R_1| = N_1, |R_2| = N_2, \ldots$
Want an algorithm that is optimal in $N_1, N_2, \ldots$

Maximal Output Size

A *cardinality constraint* (or cardinality statistics) is an assertion

$$|R_i| \leq N_i$$

A set of cardinality constraints (statistics) is $\Sigma = \{|R_1| \leq N_1, |R_2| \leq N_2, \ldots\}$.

A database satisfies Σ, $\boldsymbol{D} \vDash \Sigma$, if $|R_1^D| \leq N_1, |R_2^D| \leq N_2, \ldots$

Q' maximal output size is $\max_{\boldsymbol{D} \vDash \Sigma} |Q(\boldsymbol{D})|$; written $\max_\Sigma |Q|$ or $\max |Q|$.

Observation Any algorithm takes time $\Omega(\max |Q|)$ on some inputs.

Examples

Assume $|R| \leq N_1, |S| \leq N_2, |T| \leq N_3$.
What is $\max_\Sigma |Q|$ in each case below? In class
Start with the simpler case: $N_1 = N_2 = N_3 = N$.

$$Q_1(x,y,z) = R(x,y) \wedge S(y,z) \quad \text{// One join}$$
$$Q_2(x,y) = R(x) \wedge S(x,y) \wedge T(y) \quad \text{// Bow-tie}$$
$$Q_3(x,y,z,u) = R(x,y) \wedge S(y,z) \wedge T(z,u) \quad \text{// Two joins}$$
$$Q_4(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x) \quad \text{// Triangles}$$
$$Q_5 = \exists x \exists y \exists z (R(x,y) \wedge S(y,z) \wedge T(z,x))$$

Full CQ and Boolean CQ

- Q is full if it all its variables are head variables.

  An algorithm is *worst case optimal* if it runs in time $\tilde{O}(\max_\Sigma |Q|)$.

  This week (two lectures): worst-case optimal algorithms for full CQ.

- Q is Boolean if all its variables are existentially quantified.

  A worst case optimal algorithm is impossible why?. Best techniques use *tree decomposition*.

  Next week, two guest lectures by Hung Ngo.

Full CQ

Fix statistics Σ and a full conjunctive query Q.

Problem: compute $\max_\Sigma |Q|$.

The Hypergraph of a Query

A hypergraph is $G = (V, E)$, where every hyperedge $e \in E$ is $e \subseteq V$.

An undirected graph is the special case when $|e| = 2$ forall $e \in E$.

An *edge cover* is a subset $E' \subseteq E$ s.t. every node $x \in V$ occurs in some edge $e \in E'$.

Every full query $Q(x_1, \ldots, x_k) = R_1(\boldsymbol{X}_1) \wedge \cdots \wedge R_m(\boldsymbol{X}_m)$ is associated to the hypergraph $(\{x_1, \ldots, x_k\}, \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\})$.

An edge cover for Q is a subset of atoms $R_{i_1}, R_{i_2}, \ldots$ that contain all variables.

Full CQ: Main Result

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

Fact

If $R_{i_1}, \ldots, R_{i_w}$ *is an* edge-cover, *then* $|Q| \leq |R_{i_1}| \cdot |R_{i_2}| \cdots |R_{i_w}|$

Example: $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$
Then $|Q| \leq |R| \cdot |S|$ and $|Q| \leq |R| \cdot |S|$ and $|Q| \leq |S| \cdot |T|$.

Theorem (Atserias,Grohe,Marx (AGM Bound))

If $w_1, \ldots, w_m \in [0, 1]$ *is a* fractional edge cover,[a] $|Q| \leq |R_1|^{w_1} \cdot |R_2|^{w_2} \cdots |R_m|^{w_m}$.

[a] Will define later; but what could it be?.

$Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$ then $|Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}$

Entropy

Definition

Fix a random variable X with N outcomes, with probabilities $p_1, \ldots, p_N$.
Its *entropy* is $H(X) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i$.

What everyone should know:

- $H(X) \geq 0$.
- $H(X) = 0$ iff X is deterministic: $\exists i, p_i = 1$ and $\forall j \neq i, p_j = 0$.
- $H(X) \leq \log N$, where N = number of possible outcomes. proof in class
- $H(X) = \log N$ iff X is uniform: $p_1 = \cdots = p_N = \frac{1}{N}$.

Entropy of Multiple Variables

Consider k random variables $X_1, \ldots, X_k$.

The tuple $(X_1, \ldots, X_k)$ is call the joint random variable.

Its entropy is $H(X_1 \cdots X_k)$.

Thus, we may talk about $H(XY)$, $H(X)$, $H(Z)$, $H(XYZ)$ etc.

In class: what is $H(\emptyset) =$?

We call the function $2^{\{X_1,\ldots,X_k\}} \rightarrow \mathbb{R}$, $\{X_{i_1}, \ldots, X_{i_m}\} \mapsto H(X_{i_1} \ldots X_{i_m})$ an *entropic function*.

The Entropic Bound

Fix a full CQ and constraints:

$$Q(X_1, \ldots, X_k) = R_1(\boldsymbol{X}_1) \wedge \cdots R_m(\boldsymbol{X}_m)$$
$$\Sigma = \{|R_i| \leq N_i \mid i = 1, m\}$$

We say that H *satisfies* the constraints if $H(\boldsymbol{X}_i) \leq \log N_i$ for $i = 1, m$.

Theorem (The Entropic Bound)

$$\log\left(\max_{\Sigma} |Q|\right) = \max_{\textit{entropic } H \models \Sigma} H(X_1 \cdots X_k)$$

Proof of $\log |Q(\boldsymbol{D})| \leq \max_{H \models \Sigma} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \wedge S(y, z) \wedge T(z, x)$
Consider the answer $Q(\boldsymbol{D})$ on some $\boldsymbol{D}$.
Define the uniform probability space on the joint random variables XYZ.
This induces marginal probabilities X, Y, and Z.

$Q(\boldsymbol{D})$:

x	y	z	
a	3	r	$\frac{1}{5}$
a	2	q	$\frac{1}{5}$
b	2	q	$\frac{1}{5}$
d	3	r	$\frac{1}{5}$
a	3	q	$\frac{1}{5}$

R^D:

x	y	
a	3	$\frac{2}{5}$
a	2	$\frac{1}{5}$
b	2	$\frac{1}{5}$
d	3	$\frac{1}{5}$

S^D:

y	z	
3	r	$\frac{2}{5}$
2	q	$\frac{2}{5}$
3	q	$\frac{1}{5}$
4	q	0

T^D:

x	z	
a	r	$\frac{1}{5}$
a	q	$\frac{2}{5}$
b	q	$\frac{1}{5}$
d	r	$\frac{1}{5}$

$H(XYZ) = \log 5$, and $H(XY) \leq \log |R^D| = \log 4$; $H(YZ), H(XZ) \leq \log 4$.
In general, for any input $\boldsymbol{D}$: $\log |Q(\boldsymbol{D})| = H(XYZ) \leq \max_{H \models \Sigma} H(XYZ)$

Discussion

- Our problem is to compute $\max_{\boldsymbol{D} \models \Sigma} |Q(\boldsymbol{D})|$.
- We observed that this is the same as computing $\max_{H \models \Sigma} H(X_1 \cdots X_k)$.
- Doesn't look like great progress.
- But will show next how to upper bound H.

Shannon's Inequalities

What everyone should know about the entropy:

Emptyset $H(\varnothing) = 0$

Monotonicity If $\boldsymbol{X} \subseteq \boldsymbol{Y}$ then $H(\boldsymbol{X}) \leq H(\boldsymbol{Y})$.

Submodularity $H(\boldsymbol{X} \cap \boldsymbol{Y}) + H(\boldsymbol{X} \cup \boldsymbol{Y}) \leq H(\boldsymbol{X}) + H(\boldsymbol{Y})$.

Definition

A function $H : 2^{\{X_1,\ldots,X_k\}} \to \mathbb{R}$ with these properties is called polymatroid.

Every entropic function is a polymatroid; converse fails when $k \geq 4$.

Example

$Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$
Claim: $|R|,|S|,|T| \leq N$ implies $|Q| \leq N^{3/2}$.

Proof:

$$\begin{aligned}
3\log N = \log|R| + \log|S| + \log|T| \geq & H(XY) + H(YZ) + H(XZ) \\
\geq & H(XYZ) + H(Y) + H(XZ) \quad \text{why?} \\
\geq & H(XYZ) + H(XYZ) + H(\varnothing) \quad \text{why?} \\
= & 2H(XYZ) = 2\log|Q|
\end{aligned}$$

This inequality is a special case of Shearer's inequality (next).

Covers in a Hypergraph

Let (V, E) be a hypergraph,
where $V = \{X_1, \ldots, X_k\}$, $E = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\}$.

Definition

A fractional edge cover is a vector $\boldsymbol{w} = (w_1, \ldots, w_m)$ s.t.
"every variable X_i is covered": $\sum_{j: X_i \in \boldsymbol{X}_j} w_j \geq 1$.

Definition

A fractional vertex packing is a vector $\boldsymbol{v} = (v_1, \ldots, v_k)$ s.t.
"every edge $\boldsymbol{X}_j$ is packed": $\sum_{i: X_i \in \boldsymbol{X}_j} v_i \leq 1$.

Theorem

$\min_{\boldsymbol{w}} \sum_j w_j = \max_{\boldsymbol{v}} \sum_i v_i \stackrel{def}{=} \rho^*$;
This is called the fractional edge covering number of the hypergraph.

Proof on the next slide.

Proof of $\min_{\mathbf{w}} \sum_j w_j = \max_{\mathbf{v}} \sum_i v_i$

We use the strong duality theorem for linear programs.
Will illustrate on the triangle query:
$G = (\{x_1, x_2, x_3\}, \{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_1\})$.

$$
\begin{array}{lllll}
 & \multicolumn{4}{l}{\text{minimize } w_1 + w_2 + w_3} \\
\text{Cover } x_1: & w_1 + & & w_3 & \geq 1 \\
\text{Cover } x_2: & w_1 + & w_2 & & \geq 1 \\
\text{Cover } x_3: & & w_2 + & w_3 & \geq 1
\end{array}
\qquad
\begin{array}{lllll}
 & \multicolumn{4}{l}{\text{maximize } v_1 + v_2 + v_3} \\
\text{Pack } \{x_1, x_2\}: & v_1 + & v_2 & & \leq 1 \\
\text{Pack } \{x_2, x_3\}: & & v_2 + & v_3 & \leq 1 \\
\text{Pack } \{x_3, x_1\}: & v_1 + & & v_3 & \geq 1
\end{array}
$$

These two linear programs are dual, hence
$\min(w_1 + w_2 + w_3) = \max(v_1 + v_2 + v_3)$.

Discussion

- Optimal fractional edge cover = optimal fractional vertex packing.
- Useful exercise: check this statement for these hypegraphs:

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

- For integral edge covers / vertex packings, we only have $\geq$.

Shearer's Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m\}$. H = entropic function.

Theorem (Shearer version 1)

If $w_1, \ldots, w_m$ is a fractional edge cover then
$w_1 H(\boldsymbol{X}_1) + \cdots + w_m H(\boldsymbol{X}_m) \geq H(X_1 \cdots X_k)$

Theorem (Shearer version 2)

If every variable X_i is k-covered (i.e. occurs in at least k hyperedges), then
$H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq kH(X_1 \cdots X_k)$

Example:

$$\frac{1}{2}H(XY) + \frac{1}{2}H(YZ) + \frac{1}{2}H(ZX) \geq H(XYZ)$$

$$H(XY) + H(YZ) + H(ZX) \geq 2H(XYZ)$$

The two formulations are equivalent why?

We will prove version 2, by generalizing the proof in the triangle query.

Proof of $H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq kH(X_1 \cdots X_k)$

A *sub-modularity step* consists of replacing $H(\boldsymbol{X}_i) + H(\boldsymbol{X}_j)$ with $H(\boldsymbol{X}_i \cap \boldsymbol{X}_j) + H(\boldsymbol{X}_i \cup \boldsymbol{X}_j)$

Claim 1: Invariant After an SM step, every variable remains k-covered

Proof: A variable X can occur in 0,1 or 2 times in $H(\boldsymbol{X}_i) + H(\boldsymbol{X}_j)$;
it occurs the same number of times in $H(\boldsymbol{X}_i \cap \boldsymbol{X}_j) + H(\boldsymbol{X}_i \cup \boldsymbol{X}_j)$. why?

Proof of $H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq kH(X_1 \cdots X_k)$

Claim 2: Progress If $\boldsymbol{X}_i \not\subseteq \boldsymbol{X}_j$ and $\boldsymbol{X}_j \not\subseteq \boldsymbol{X}_i$ then, after an SM step, the quantity $\sum_\ell |\boldsymbol{X}_\ell|^2$ strictly increases.

Proof: $|\boldsymbol{X}_i|^2 + |\boldsymbol{X}_j|^2 < |\boldsymbol{X}_i \cap \boldsymbol{X}_j|^2 + |\boldsymbol{X}_i \cup \boldsymbol{X}_j|^2$ why?

Proof of $H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq kH(X_1 \cdots X_k)$

Claim 3: Termination We have proven:

$$H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq H(\boldsymbol{Y}_1) + \cdots + H(\boldsymbol{Y}_m)$$

where every variable is k-covered by $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ (invariant!)
and $\boldsymbol{Y}_1 \supseteq \boldsymbol{Y}_2 \supseteq \boldsymbol{Y}_3 \supseteq \cdots$ (no more progress!)

That means that $\boldsymbol{Y}_1 = \boldsymbol{Y}_2 = \cdots = \boldsymbol{Y}_k = \{X_1, \ldots, X_k\}$ why?, thus:

$$H(\boldsymbol{X}_1) + \cdots + H(\boldsymbol{X}_m) \geq kH(X_1 \cdots X_k) + [\text{stuff}] \geq H(X_1 \cdots X_k)$$

Discussion

- We proved something stronger: Shearer's inequality holds for all polymatroids H.
- The converse also holds: if $\sum_j w_j H(\boldsymbol{X}_j) \geq H(X_1 \ldots X_k)$ for all entropic functions, then $w_1, \ldots, w_k$ is a fractional edge cover.
- Next: the AGM bound is Sheare's lemma restated in terms of a query PLUS a proof that the inequality is tight.

AGM Bound for $Q(X_1,\dots,X_k) = R_1(\boldsymbol{X}_1)\wedge\cdots\wedge R_m(\boldsymbol{X}_m)$

Assume uniform statistics $|R_1|, |R_2|, \dots, |R_m| \leq N$.

Lemma

(a) If $w_1,\dots,w_m$ is a fractional edge cover, then $\forall \boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq N^{w_1+\cdots+w_m}$.
(b) If $v_1,\dots,v_k$ is a fractional vertex packing, then $\exists \boldsymbol{D}$, $|Q(\boldsymbol{D})| = N^{v_1+\cdots+v_k}$

Proof. (a) $\log\max|Q(\boldsymbol{D})| \leq \max H(\boldsymbol{X}) \leq \sum_j w_j H(\boldsymbol{X}_j)$ (Shearer)

(b) "Product database": $R_j^D \stackrel{\text{def}}{=} \prod_{X_i\in\boldsymbol{X}_j}[N^{v_i}]$.
Then $|R_j^D| \leq N$, $\forall j$, and $Q(\boldsymbol{D}) = N^{v_1+\cdots+v_k}$

E.g. $Q(x,y,z) = R(x,y)\wedge S(y,z)\wedge T(z,x)$; $\quad v_x = v_y = v_z = \frac{1}{2}$.

$$R^D \stackrel{\text{def}}{=} [N^{1/2}]\times[N^{1/2}] \quad S^D \stackrel{\text{def}}{=} [N^{1/2}]\times[N^{1/2}] \quad T^D \stackrel{\text{def}}{=} [N^{1/2}]\times[N^{1/2}]$$

Then $|R^D|, |S^D|, |T^D| \leq N$, and $Q(\boldsymbol{D}) = [N^{1/2}]\times[N^{1/2}]\times[N^{1/2}]$

AGM Bound

Theorem (AGM Bound - Uniform cardinalities)

$\max |Q(\boldsymbol{D})| = \max 2^{H(\boldsymbol{X})} = N^{\rho^*}$

We denote this quantity by $AGM(Q)$.

Proof:

- $\log \max |Q(\boldsymbol{D})| \leq \max H(\boldsymbol{X})$ was the proof by example.
- $H(\boldsymbol{X}) \leq \sum w_j H(\boldsymbol{X}_j) = \rho^* \log N$ Shearer's inequality.
- $N^{\rho^*} \leq \max |Q(\boldsymbol{D})|$ worst-case (product) instance $\boldsymbol{D}$.

AGM Bound for $Q(X_1, \ldots, X_k) = R_1(\boldsymbol{X}_1) \wedge \cdots \wedge R_m(\boldsymbol{X}_m)$

Assume general statistics $|R_1| \leq N_1, \ldots, |R_m| \leq N_m$.
A *generalized fractional vertex packing* is $v_1, \ldots, v_k$ s.t. for every edge $R_j(\boldsymbol{X}_j)$: $\sum_{i:X_i \in \boldsymbol{X}_j} v_i \leq \log N_j$.

Lemma

(a) If $w_1, \ldots, w_m$ is a fractional edge cover, then $\forall \boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq N_1^{w_1} \cdots N_m^{w_m}$.
(b) If $v_1, \ldots, v_k$ is a generalized frac vertex packing, $\exists \boldsymbol{D}$, $|Q(\boldsymbol{D})| = 2^{v_1 + \cdots + v_k}$

Proof: straightforward generalization of the previous arguments. (Will skip in class, but it really helps if you review it at home.)

AGM Bound

Theorem (AGM Bound - general cardinalities)

$\max |Q(\boldsymbol{D})| = \max 2^{H(\boldsymbol{X})} = \min_{\boldsymbol{w}} \prod_j |R_j|^{w_j}$.

We denote this quantity by $AGM(Q)$.

Example

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

Find max $Q(\boldsymbol{D})$

For any fractional edge cover w_R, w_S, w_T: $|Q| \leq |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T}$.

w_R	w_S	w_T	$\lvert N_R\rvert^{w_R} \cdot \lvert N_S\rvert^{w_S} \cdot \lvert N_T\rvert^{w_T}$
1/2	1/2	1/2	$\sqrt{N_R N_S N_T}$
1	1	0	$N_R N_S$
0	1	1	$N_S N_T$
1	0	1	$N_R N_T$

The smallest of these values is the tight bound of $|Q(\boldsymbol{D})|$.

In class: what is the worst case instance $\boldsymbol{D}$?

Example

In class:

$$Q(x, y) = R(x) \wedge S(x, y) \wedge T(y)$$

Find $\max Q(\boldsymbol{D})$

Discussion

- The worst case database, where $Q(\boldsymbol{D}) = AGM(Q)$ is a *product* database.
- To compute $AGM(Q)$ we need to compute $\min_{\boldsymbol{w}} N_j^{w_j}$ where $\boldsymbol{w}$ ranges over all fractional edge covers.
- There are infinitely many $\boldsymbol{w}$'s!
- Good news: suffices to check *vertices of the edge covering polytope*, of which there are only finitely many.

Vertices of the Edge Covering Polytope

A polytope $P \subseteq \mathbb{R}^k$ is the intersection of semi-spaces:

$$P = \bigcap_i \{\boldsymbol{w} \mid \sum_j a_{ij} w_j \leq b_j\}$$

A polytope is convex: if $\boldsymbol{w}_1, \boldsymbol{w}_2 \in P$ then $(1-\lambda)\boldsymbol{w}_1 + \lambda \boldsymbol{w}_2 \in P$.

Call $\boldsymbol{w} \in P$ a vertex if it is no strict convex combination[1] of points in P.

For any linear function $f(\boldsymbol{w}) \stackrel{\text{def}}{=} \sum_j b_j w_j$ its minimum is at a vertex of the polytope why?

It follows, for the edge-covering polytope:

$$\min_{\boldsymbol{w} \in P} N_j^{w_j} = \min_{\boldsymbol{w} \in \text{vertices}(P)} N_j^{w_j}$$

In class find the vertices of $R(x,y) \wedge S(y,z) \wedge T(z,u) \wedge K(u,x)$.

[1] A strict convex combination is $\boldsymbol{w} = (1-\lambda)\boldsymbol{w}_1 + \lambda \boldsymbol{w}_2$ with $\lambda \neq 0, \lambda \neq 1$.

Discussion

- The AGM bound is Shearer's inequality PLUS tightness proof.
- The bound is reached by some "product" database instance.
- To be of practical value (in databases) the AGM bound needs to be extended to handle more complex statistics: this is not trivial. Next: a simple extension that *is* trivial.

Simple Functional Dependencies

Fix a relation $R(A_1, \ldots, A_\ell)$.
A simple functional dependency is of the form $A_i \to A_j$.
Meaning: every two tuples in R that agree on A_i must also agree on A_j.

Let Σ = set of statistics; Γ = set of simple FD's.

Problem: find $AGM_\Gamma(Q) \stackrel{\text{def}}{=} \max_{\boldsymbol{D} \models \Sigma, \Gamma} |Q(\boldsymbol{D})|$.

In general, $AGM_\Gamma(Q) \leq AGM(Q)$, but it is not tight.

Simple Functional Dependencies

Given Q, Γ, denote $\bar{Q}$ the query obtained as follows:

- If some relation R_j satisfies the simple FD $A \to B$ and R_i contains the attribute (variable) A, then add B to R_i (and increase its arity).
- Repeat until no more change.

Then $AGM_\Gamma(Q) = AGM(\bar{Q})$.

Examples

Assume $|R|, |S|, |T| \leq N$.

Example 1: $Q(x,y,z) = R(x,y) \wedge S(y,z)$
Compute $AGM_{S.y \to S.z}(Q)$.

- $AGM(Q) = N^2$
- $y \to z$ implies $\bar{Q}(x,y,z) = R(x,y,z) \wedge S(y,z)$
- $AGM_{S.y \to S.z}(Q) = N$

Example 2: $Q(x,y,z) = R(x,y) \wedge S(y,z) \wedge T(z,x)$
Compute $AGM_{S.y \to S.z}(Q)$

- $AGM(Q) = N^{3/2}$
- $y \to z$ implies $\bar{Q}(x,y,z) = R(x,y,z) \wedge S(y,z) \wedge T(z,x)$
- $AGM_{S.y \to S.z}(Q) = N$

Worst Case Optimal Algorithm

Problem: find an algorithm to compute $Q(\boldsymbol{D})$ in time $\tilde{O}(AGM(Q))$.

First such algorithm described by [Ngo, Porat, Re, Rudra]; it was a breakthrough but too complex. Later they simplified it significantly to an algorithm called *Generic Join*. Everyone should know GJ.

Generic Join

$Q(x_1, \dots, x_k) = R_1(\boldsymbol{X}_1) \wedge \cdots \wedge R_m(\boldsymbol{X}_m)$

Compute by calling Generic-join(Q,k,$()$):

Generic-join(Q, k, $\boldsymbol{a}$):
    if $k = 0$ **then** print $\boldsymbol{a}$
    choose any variable x
    let $J = \{j \mid x \in \boldsymbol{X}_j\}$ // atoms containing x
    let $D_j = \Pi_x(R_j)$, forall $j \in J$ // domains of x
    for v **in** $\bigcap_{j \in J} D_j$
        // must compute intersection in time $O(\min(|D_j|))$
        Generic-join($Q[v/x]$, $k-1$, $(\boldsymbol{a}, v)$)

$Q[v/x]$ is the *residual query*, where x is substituted with constant v.

Example

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

```
let D_R = Π_x(R), D_T = Π_x(T)
for u in D_R ∩ D_T do
    // compute query R(u, y) ∧ S(y, z) ∧ T(z, u)
    let D_R = Π_y(σ_{x=u}(R)), D_S = Π_y(S)
    for v in D_R ∩ D_S do
        // compute query R(u, v) ∧ S(v, z) ∧ T(z, u)
        let D_S = Π_z(σ_{y=v}(S)), D_T = Π_z(σ_{x=u}(T))
        for w in D_S ∩ D_T do
            print u, v, w
```

Next: we will prove its runtime.

Runtime of GJ

$$Q(x_1, \dots, x_k) = R_1(\boldsymbol{X}_1) \wedge \cdots \wedge R_m(\boldsymbol{X}_m)$$

Let $T_{GJ}(Q)$ be the runtime of GJ, assuming every relation $R_j^D(\boldsymbol{X}_j)$ is sorted lexicographically, by the attribute order in GJ.

Theorem

Let $w_1, \dots, w_m$ be any fractional edge cover. Then $T_{GJ}(Q) = \tilde{O}(\prod_j N_j^{w_j})$.

It follows that $T_{GJ}(Q) = \tilde{O}(AGM(Q))$.

We will prove the theorem by induction on the number of variables in Q.

Background: Intersection

Given 2 sorted lists (of numbers, or strings) D_1, D_2, compute $D_1 \cap D_2$.

In class:

- Describe an algorithm that runs in time $\tilde{O}(|D_1| + |D_2|)$.
  (this is $= \tilde{O}(\max(|D_1|, |D_2|))$).

- Describe a better algorithm that runs in time $\tilde{O}(\min(|D_1|, |D_2|))$.
  Example: if $|D_1| = 1$ then compute intersection in time $\tilde{O}(1) = O(\log n)$. who is n?

Runtime of GJ: Base Case: Q has a single variable x

$Q(x) = R_1(x) \wedge \cdots \wedge R_k(x)$

Let $w_1, \ldots, w_k$ be a fractional edge cover.

Then the runtime is $T_{GJ}(Q) = \tilde{O}(\min(N_1, \ldots, N_k))$

Claim: $\min(N_1, \ldots, N_k) \leq N_1^{w_1} \cdots N_k^{w_k}$ why?

This proves $T_{GJ}(Q) = \tilde{O}(N_1^{w_1} \cdots N_k^{w_k})$.

Background: Hölder's Generalized Inequality

Cauchy-Schwartz:

$$\sum_i a_i^{\frac{1}{2}} b_i^{\frac{1}{2}} \leq \left(\sum_i a_i\right)^{\frac{1}{2}} \left(\sum_i b_i\right)^{\frac{1}{2}}$$

Hölder: if $w_1 + w_2 \geq 1$, then

$$\sum_i a_i^{w_1} b_i^{w_2} \leq \left(\sum_i a_i\right)^{w_1} \left(\sum_i b_i\right)^{w_2}$$

Generalized Hölder: if $w_1 + w_2 + w_3 + \ldots \geq 1$, then

$$\sum_i a_i^{w_1} b_i^{w_2} c_i^{w_3} \cdots \leq \left(\sum_i a_i\right)^{w_1} \left(\sum_i b_i\right)^{w_2} \left(\sum_i c_i\right)^{w_3} \cdots$$

Runtime of GJ: Induction Step; GJ iterates over x_1

$$Q(x_1,\dots,x_k) = \underbrace{R_1(\boldsymbol{X}_1)\wedge\cdots\wedge R_{j_0}(\boldsymbol{X}_{j_0})}_{\text{Contain } x_1}\wedge\underbrace{R_{j_0+1}(\boldsymbol{X}_{j_0+1})\wedge\cdots\wedge R_m(\boldsymbol{X}_m)}_{\text{don't contain } x_1}$$

We prove $T_{GJ}(Q) = \tilde{O}(N_1^{w_1}\cdots N_m^{w_m})$.

- Time for $\Pi_x(R_1)\cap\cdots\cap\Pi_x(R_{j_0})$ is $\tilde{O}(N_1^{w_1}\cdots N_{j_0}^{w_{j_0}}) \leq \tilde{O}(N_1^{w_1}\cdots N_m^{w_m})$
- Time for residual query $Q[a/x]$. By induction:

$$T_{GJ}(Q[a/x_1]) = \underbrace{N_{1,a}^{w_1}}_{\stackrel{\text{def}}{=}|\sigma_{x_1=a}(R_1)|}\cdots\underbrace{N_{j_0,a}^{w_{j_0}}}_{\stackrel{\text{def}}{=}|\sigma_{x_1=a}(R_{j_0})|}\cdot N_{j_0+1}^{w_{j_0+1}}\cdots N_m^{w_m}$$

Total runtime is obtained by summing on a:

$$\sum_a N_{1,a}^{w_1}\cdots N_{j_0,a}^{w_{j_0}}\cdot N_{j_0+1}^{w_{j_0+1}}\cdots N_m^{w_m} \leq \underbrace{\left(\sum_a N_{1,a}\right)^{w_1}}_{=(N_1)^{w_1}}\cdots\underbrace{\left(\sum_a N_{j_0,a}\right)^{w_{j_0}}}_{=(N_{j_0})^{w_{j_0}}}\cdot N_{j_0+1}^{w_{j_0+1}}\cdots N_m^{w_m}$$

Discussion

- The AGM bound can be smaller than $\max_j N_j$. This means that GJ may not necessarily read all the data.
  E.g. computing $R_1 \cap R_2$ when $N_1 \ll N_2$: do a binary search in R_2.

- Hölder's generalized inequality only holds when $w_1 + w_2 + \cdots \geq 1$.
  Thus, it is necessary that x_1 be "covered" (and same for $x_2, x_3, \ldots$).

- Our proof of the runtime also implies $Q(\boldsymbol{D}) \leq \prod_j N_j^{w_j}$. But this means that we have proven Shearer's inequality again! What is the clean proof of Shearer's inequality that corresponds to GJ?

Conditional Polymatroid/Entropy

We will define the conditional polymatroid as $H(\boldsymbol{Z}|\boldsymbol{Y}) \stackrel{\text{def}}{=} H(\boldsymbol{Y}\boldsymbol{Z}) - H(\boldsymbol{Y})$. When H is entropic, then the conditional entropy has a meaning the entropy of a conditional probability space. We don't need this here.

Lemma

(1) $H(\boldsymbol{Z}|\boldsymbol{Y}) \geq H(\boldsymbol{Z}|\boldsymbol{X}\boldsymbol{Y})$ *(2)* $H'(\boldsymbol{Z}) \stackrel{\text{def}}{=} H(\boldsymbol{Z}|\boldsymbol{Y})$ *is a polymatroid.*

Proof: (1)

$$H(\boldsymbol{X}\boldsymbol{Y}) + H(\boldsymbol{Y}\boldsymbol{Z}) \geq H(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) + H(\underbrace{(\boldsymbol{X}\boldsymbol{Y}) \cap (\boldsymbol{Y}\boldsymbol{Z})}_{\text{not necessarily } \boldsymbol{Y} \text{ why?}})$$

$$\geq H(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) + H(\boldsymbol{Y})$$

$$H(\boldsymbol{Y}\boldsymbol{Z}) - H(\boldsymbol{Y}) \geq H(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) - H(\boldsymbol{X}\boldsymbol{Y})$$

(2) exercise.

Proof #2 of Shearer's Inequality

We prove: for any polymatroid H: $\sum_j w_j H(\boldsymbol{X}_j) \geq H(X_1 \ldots X_k)$.
when $w_1, \ldots, w_m$ is a fractional edge cover.

$$\underbrace{\Big(w_1 H(\boldsymbol{X}_1) + \ldots + w_{j_0} H(\boldsymbol{X}_{j_0})\Big)}_{\text{contain } X_1} + \underbrace{(\ldots + w_m H(\boldsymbol{X}_m))}_{\text{do not contain } X_1} =$$

$$=(w_1 + \ldots + w_{j_0}) H(X_1) + \Big(w_1 H(\boldsymbol{X}_1|X_1) + \ldots + w_{j_0} H(\boldsymbol{X}_{j_0}|X_1)\Big) + (\ldots + H(\boldsymbol{X}_m$$

$$\geq H(X_1) + \Big(w_1 H(\boldsymbol{X}_1|X_1) + \ldots + w_{j_0} H(\boldsymbol{X}_{j_0}|X_1)\Big) + (\ldots + H(\boldsymbol{X}_m))$$

$$\geq H(X_1) + \Big(w_1 H(\boldsymbol{X}_1|X_1) + \ldots + w_{j_0} H(\boldsymbol{X}_{j_0}|X_1)\Big) + (\ldots + H(\boldsymbol{X}_m|X_1))$$

$$\geq H(X_1) + H(X_1 X_2 \ldots X_k | X_1)$$

$$= H(X_1 X_2 \ldots X_k)$$

Discussion

- Main take away: GJ is very simple *and* worst case optimal!
- Query engines in database systems are *not* worst case optimal.
- GJ requires all relations to be pre-sorted. If not, then sort them dynamically; the additional cost $\sum_j N_j \log N_j$ may exceed the AGM bound.
- GJ does *only* intersection: great candidate for vectorization.
- GJ is designed for on Full CQ. In practice, most data analytics queries are aggregates; e.g. $\exists$-aggregate (a.k.a. Boolean query), `count`, `sum`, etc. Next week, Thursday at $9:30$ and Friday at 10, Hung Ngo will give two lectures on the FAQ algorithm for aggregate queries.