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) $D$, what is the algorithmic complexity for computing $Q(D)$?

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

And we will consider only Conjunctive Queries: $\exists x (R_1 \land R_2 \land \cdots)$. 

The Problem

Suppose $Q$ is in prenex normal form with $k$ variables. Suppose the domain size is $n = |D|$. A naive algorithm computes $Q(D)$ 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 $\Sigma$, $D \models \Sigma$, if $|R_1^D| \leq N_1, |R_2^D| \leq N_2, \ldots$

$Q'$ maximal output size is $\max_{D \models \Sigma} |Q(D)|$; 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: $N_1 = N_2 = N_3 = N$.

$Q_1(x, y, z) = R(x, y) \land S(y, z)$ // One join

$Q_2(x, y) = R(x) \land S(x, y) \land T(y)$ // Bow-tie

$Q_3(x, y, z, u) = R(x, y) \land S(y, z) \land T(z, u)$ // Two joins

$Q_4(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$ // Triangles

$Q_5 = \exists x \exists y \exists z (R(x, y) \land S(y, z) \land T(z, x))$
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: $N_1 = N_2 = N_3 = N$.

\[
Q_1(x, y, z) = R(x, y) \land S(y, z) \quad / \quad \text{One join}
\]
\[
Q_2(x, y) = R(x) \land S(x, y) \land T(y) \quad / \quad \text{Bow-tie}
\]
\[
Q_3(x, y, z, u) = R(x, y) \land S(y, z) \land T(z, u) \quad / \quad \text{Two joins}
\]
\[
Q_4(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \quad / \quad \text{Triangles}
\]
\[
Q_5 = \exists x \exists y \exists z (R(x, y) \land S(y, z) \land T(z, x))
\]
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: $N_1 = N_2 = N_3 = N$.

\[ Q_1(x, y, z) = R(x, y) \land S(y, z) \] // One join
\[ Q_2(x, y) = R(x) \land S(x, y) \land T(y) \] // Bow-tie
\[ Q_3(x, y, z, u) = R(x, y) \land S(y, z) \land T(z, u) \] // Two joins
\[ Q_4(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \] // Triangles
\[ Q_5 = \exists x \exists y \exists z (R(x, y) \land S(y, z) \land T(z, x)) \]
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: $N_1 = N_2 = N_3 = N$.

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

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

\[
Q_1(x, y, z) = R(x, y) \land S(y, z) \quad // \text{One join}
\]
\[
Q_2(x, y) = R(x) \land S(x, y) \land T(y) \quad // \text{Bow-tie}
\]
\[
Q_3(x, y, z, u) = R(x, y) \land S(y, z) \land T(z, u) \quad // \text{Two joins}
\]
\[
Q_4(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \quad // \text{Triangles}
\]
\[
Q_5 = \exists x \exists y \exists z (R(x, y) \land S(y, z) \land T(z, x))
\]
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 $\Sigma$ 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 $Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m)$ is associated to the hypergraph $\langle \{x_1, \ldots, x_k\}, \{X_1, \ldots, X_m\} \rangle$.

An edge cover for $Q$ is a subset of atoms $R_{i_1}, R_{i_2}, \ldots$ that contain all variables.
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 $|e| = 2$ for all $e \in E$.

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 $Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m)$ is associated to the hypergraph $(\{x_1, \ldots, x_k\}, \{X_1, \ldots, X_m\})$.

An edge cover for $Q$ is a subset of atoms $R_{i_1}, R_{i_2}, \ldots$ that contain all variables.
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 $|e| = 2$ for all $e \in E$.

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 $Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m)$ is associated to the hypergraph $(\{x_1, \ldots, x_k\}, \{X_1, \ldots, X_m\})$.

An edge cover for $Q$ is a subset of atoms $R_{i_1}, R_{i_2}, \ldots$ that contain all variables.
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 \( Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m) \) is associated to the hypergraph (\( \{x_1, \ldots, x_k\}, \{X_1, \ldots, X_m\} \)).

An edge cover for \( Q \) is a subset of atoms \( R_{i_1}, R_{i_2}, \ldots \) that contain all variables.
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 \( |e| = 2 \) for all \( e \in E \).

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 \( Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m) \) is associated to the hypergraph \( (\{x_1, \ldots, x_k\}, \{X_1, \ldots, X_m\}) \).

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

\[ Q(\mathbf{X}) = R_1(\mathbf{X}_1) \land \cdots \land R_m(\mathbf{X}_m) \]

**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:** \( Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \)

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, then* \(|Q| \leq |R_1|^{w_1} \cdot |R_2|^{w_2} \cdots |R_m|^{w_m}\).

\(^a\)Will define later; but what could it be?.

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \text{ then } |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2} \]
Full CQ: Main Result

\[ Q(\mathbf{X}) = R_1(\mathbf{X}_1) \land \cdots \land R_m(\mathbf{X}_m) \]

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: \( Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \)

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?

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \text{ then } |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2} \]
**Problem Definition**

**AGM Bound**

**Worst Case Algorithm**

---

## Full CQ: Main Result

\[ Q(x) = R_1(x_1) \land \cdots \land R_m(x_m) \]

### 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:

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

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?*

---

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \text{ then } |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2} \]
**Full CQ: Main Result**

\[ Q(\mathbf{X}) = R_1(\mathbf{X}_1) \land \cdots \land R_m(\mathbf{X}_m) \]

**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: \( Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \)

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, then \( |Q| \leq |R_1|^{w_1} \cdot |R_2|^{w_2} \cdots |R_m|^{w_m} \).

\(^\text{a}\)Will define later; but what could it be?

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \text{ then } |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2} \]
Entropy

Definition

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

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 $H(X_1 \ldots X_k)$.

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:

\[
Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m)
\]
\[
\Sigma = \{ |R_i| \leq N_i \mid i = 1, m \}
\]

We say that \( H \) satisfies the constraints if \( H(X_i) \leq \log N_i \) for \( i = 1, m \).

**Theorem (The Entropic Bound)**

\[
\log \left( \max_{\Sigma} |Q| \right) = \max_{\text{entropic } H} \ H(X_1 \cdots X_k)
\]
Proof of $\log |Q(D)| \leq \max_{H=\sum} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$
Consider the answer $Q(D)$ on some $D$. 
Proof of $\log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$

Consider the answer $Q(D)$ on some $D$.

<table>
<thead>
<tr>
<th>$Q(D)$:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
</tr>
<tr>
<td>$a$</td>
<td>2</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
</tr>
<tr>
<td>$d$</td>
<td>3</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
</tr>
</tbody>
</table>

Dan Suciu
Finite Model Theory – Unit 5
Spring 2018 14 / 49
Proof of $\log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$
Consider the answer $Q(D)$ on some $D$.
Define the uniform probability space on the joint random variables $XYZ$.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$z$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
<td>$r$</td>
</tr>
<tr>
<td>$a$</td>
<td>2</td>
<td>$q$</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
<td>$q$</td>
</tr>
<tr>
<td>$d$</td>
<td>3</td>
<td>$r$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
<td>$q$</td>
</tr>
</tbody>
</table>

$Q(D)$:

$\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$
Proof of \( \log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \cdots X_k) \)

By example: \( Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \)

Consider the answer \( Q(D) \) on some \( D \).

Define the uniform probability space on the joint random variables \( XYZ \).

This induces marginal probabilities \( X, Y, \) and \( Z \).

\[
Q(D) :
\begin{array}{ccc}
  x & y & z \\
  a & 3 & r & \frac{1}{5} \\
  a & 2 & q & \frac{1}{5} \\
  b & 2 & q & \frac{1}{5} \\
  d & 3 & r & \frac{1}{5} \\
  a & 3 & q & \frac{1}{5}
\end{array}
\]
Proof of $\log|Q(D)| \leq \max_{H=\Sigma} H(X_1\cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$

Consider the answer $Q(D)$ on some $D$.

Define the uniform probability space on the joint random variables $XYZ$. This induces marginal probabilities $X$, $Y$, and $Z$.

<table>
<thead>
<tr>
<th>$Q(D)$</th>
<th>$R^D$</th>
<th>$S^D$</th>
<th>$T^D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$z$</td>
<td>$x$</td>
</tr>
<tr>
<td>$a$</td>
<td>$3$</td>
<td>$r$</td>
<td>$a$</td>
</tr>
<tr>
<td>$a$</td>
<td>$2$</td>
<td>$q$</td>
<td>$a$</td>
</tr>
<tr>
<td>$b$</td>
<td>$2$</td>
<td>$q$</td>
<td>$b$</td>
</tr>
<tr>
<td>$d$</td>
<td>$3$</td>
<td>$r$</td>
<td>$d$</td>
</tr>
<tr>
<td>$a$</td>
<td>$3$</td>
<td>$q$</td>
<td></td>
</tr>
</tbody>
</table>
Proof of $\log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \ldots X_k)$

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

<table>
<thead>
<tr>
<th>$Q(D)$:</th>
<th>$R^D$:</th>
<th>$S^D$:</th>
<th>$T^D$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$z$</td>
<td>$x$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
<td>$r$</td>
<td>$a$</td>
</tr>
<tr>
<td>$a$</td>
<td>2</td>
<td>$q$</td>
<td>$a$</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
<td>$q$</td>
<td>$b$</td>
</tr>
<tr>
<td>$d$</td>
<td>3</td>
<td>$r$</td>
<td>$d$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
<td>$q$</td>
<td></td>
</tr>
</tbody>
</table>

$H(XYZ) = \log 5$,
Proof of $\log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$

Consider the answer $Q(D)$ on some $D$.

Define the uniform probability space on the joint random variables $XYZ$. This induces marginal probabilities $X$, $Y$, and $Z$.

$Q(D)$:

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td>r</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>d</td>
<td>3</td>
<td>r</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
</tbody>
</table>

$R^D$:

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td>$\frac{2}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>2</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>d</td>
<td>3</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>$\frac{1}{5}$</td>
</tr>
</tbody>
</table>

$S^D$:

<table>
<thead>
<tr>
<th></th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>r</td>
<td>$\frac{2}{5}$</td>
</tr>
<tr>
<td>2</td>
<td>q</td>
<td>$\frac{2}{5}$</td>
</tr>
<tr>
<td>3</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>4</td>
<td>q</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
</tbody>
</table>

$T^D$:

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>r</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>q</td>
<td>$\frac{2}{5}$</td>
</tr>
<tr>
<td>a</td>
<td>q</td>
<td>$\frac{1}{5}$</td>
</tr>
<tr>
<td>b</td>
<td>q</td>
<td>$\frac{2}{5}$</td>
</tr>
<tr>
<td>d</td>
<td>r</td>
<td>$\frac{1}{5}$</td>
</tr>
</tbody>
</table>

$H(XYZ) = \log 5$, and $H(XY) \leq \log |R^D| = \log 4$;
Proof of $\log |Q(D)| \leq \max_{H=\Sigma} H(X_1 \cdots X_k)$

By example: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$

Consider the answer $Q(D)$ on some $D$.

Define the uniform probability space on the joint random variables $XYZ$. This induces marginal probabilities $X$, $Y$, and $Z$.

For $Q(D)$:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>3</td>
<td>$r$</td>
</tr>
<tr>
<td>$a$</td>
<td>2</td>
<td>$q$</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
<td>$q$</td>
</tr>
<tr>
<td>$d$</td>
<td>3</td>
<td>$r$</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
<td>$q$</td>
</tr>
</tbody>
</table>

$\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$

For $R^D$:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>3</td>
</tr>
<tr>
<td>$a$</td>
<td>2</td>
</tr>
<tr>
<td>$b$</td>
<td>2</td>
</tr>
<tr>
<td>$d$</td>
<td>3</td>
</tr>
<tr>
<td>$a$</td>
<td>3</td>
</tr>
</tbody>
</table>

$\frac{2}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$

For $S^D$:

<table>
<thead>
<tr>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$r$</td>
</tr>
<tr>
<td>2</td>
<td>$q$</td>
</tr>
<tr>
<td>3</td>
<td>$q$</td>
</tr>
<tr>
<td>4</td>
<td>$q$</td>
</tr>
<tr>
<td>0</td>
<td>$d$</td>
</tr>
</tbody>
</table>

$\frac{2}{5}$ $\frac{2}{5}$ $\frac{1}{5}$ $\frac{1}{5}$ $\frac{1}{5}$

For $T^D$:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$r$</td>
</tr>
<tr>
<td>$a$</td>
<td>$q$</td>
</tr>
<tr>
<td>$b$</td>
<td>$q$</td>
</tr>
<tr>
<td>$d$</td>
<td>$r$</td>
</tr>
</tbody>
</table>

$\frac{1}{5}$ $\frac{2}{5}$ $\frac{2}{5}$ $\frac{1}{5}$ $\frac{1}{5}$

$H(XYZ) = \log 5$, and $H(XY) \leq \log |R^D| = \log 4$; $H(YZ), H(XZ) \leq \log 4$. 

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

By example: \( Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \)

Consider the answer \( Q(D) \) on some \( D \).

Define the uniform probability space on the joint random variables \( XYZ \).

This induces marginal probabilities \( X, Y, \) and \( Z \).

\[
\begin{align*}
Q(D) : & \quad R^D : & \quad S^D : & \quad T^D :
\begin{array}{|c|c|c|}
\hline
x & y & z \\
\hline
a & 3 & r & 1/5 \\
b & 2 & q & 1/5 \\
d & 3 & r & 1/5 \\
a & 3 & q & 1/5 \\
\hline
\end{array}
\begin{array}{|c|c|}
\hline
x & y \\
\hline
a & 3 & 2/5 \\
b & 2 & 1/5 \\
\hline
\end{array}
\begin{array}{|c|c|}
\hline
y & z \\
\hline
3 & r & 2/5 \\
3 & q & 1/5 \\
4 & q & 0 \\
\hline
\end{array}
\begin{array}{|c|c|}
\hline
x & z \\
\hline
a & r & 1/5 \\
a & q & 2/5 \\
b & q & 1/5 \\
d & r & 1/5 \\
\hline
\end{array}
\end{align*}
\]

\( H(XYZ) = \log 5 \), and \( H(XY) \leq \log |R^D| = \log 4 \); \( H(YZ), H(XZ) \leq \log 4 \).

In general, for any input \( D \): \( \log |Q(D)| = H(XYZ) \leq \max_{H \subseteq \Sigma} H(XYZ) \)
Discussion

- Our problem is to compute $\max_{D \vdash \Sigma} |Q(D)|$.

- We observed that this is the same as computing $\max_{H \vdash \Sigma} H(X_1 \cdots X_k)$.

- 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  \( H(\emptyset) = 0 \)

Monotonicity  If \( X \subseteq Y \) then \( H(X) \leq H(Y) \).

Submodularity  \( H(X \cap Y) + H(X \cup Y) \leq H(X) + H(Y) \).

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

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

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ) \\
\geq H(XYZ) + H(Y) + H(XZ) \quad \text{why?} \\
\geq H(XYZ) + H(XYZ) + H(\emptyset) \quad \text{why?} \\
= 2H(XYZ) = 2 \log |Q|
\]

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

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Claim: \(|R|, |S|, |T| \leq N\) implies \(|Q| \leq N^{3/2}\).

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ)
\]

\[
\geq H(XYZ) + H(Y) + H(XZ) \quad \text{why?}
\]

\[
\geq H(XYZ) + H(XYZ) + H(\emptyset) \quad \text{why?}
\]

\[
= 2H(XYZ) = 2 \log |Q|
\]

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

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Claim: \( |R|, |S|, |T| \leq N \) implies \( |Q| \leq N^{3/2} \).

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ) \\
\geq H(XYZ) + H(Y) + H(XZ) \quad \text{why?} \\
\geq H(XYZ) + H(XYZ) + H(\emptyset) \quad \text{why?} \\
= 2H(XYZ) = 2 \log |Q| 
\]

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

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Claim: \(|R|, |S|, |T| \leq N \) implies \(|Q| \leq N^{3/2} \).

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ)
\]

\[
\geq H(XYZ) + H(Y) + H(XZ) \quad \text{why?}
\]

\[
\geq H(XYZ) + H(XYZ) + H(\emptyset) \quad \text{why?}
\]

\[
= 2H(XYZ) = 2 \log |Q|
\]

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

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Claim: \(|R|, |S|, |T| \leq N\) implies \(|Q| \leq N^{3/2}\).

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ)
\]

\[
\geq H(XYZ) + H(Y) + H(XZ)
\]

\[
\geq H(XYZ) + H(XYZ) + H(\emptyset)
\]

\[
= 2H(XYZ) = 2 \log |Q|
\]

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

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

Proof:

\[
3 \log N = \log |R| + \log |S| + \log |T| \geq H(XY) + H(YZ) + H(XZ)
\]
\[
\geq H(XYZ) + H(Y) + H(XZ) \quad \text{why?}
\]
\[
\geq H(XYZ) + H(XYZ) + H(\emptyset) \quad \text{why?}
\]
\[
= 2H(XYZ) = 2 \log |Q|
\]

This inequality is a special case of Shearer’s inequality (next).
Covers in a Hypergraph

Let $(V, E)$ be a hypergraph, where $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$.

**Definition**

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

**Definition**

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

**Theorem**

$$\min_w \sum_j w_j = \max_v \sum_i v_i \overset{\text{def}}{=} \rho^*;$$

This is called the fractional edge covering number of the hypergraph.

Proof on the next slide.
Covers in a Hypergraph

Let $(V, E)$ be a hypergraph, where $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$.

**Definition**

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

**Definition**

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

**Theorem**

$$\min_w \sum_j w_j = \max_v \sum_i v_i \overset{\text{def}}{=} \rho^*;$$

This is called the fractional edge covering number of the hypergraph.

Proof on the next slide.
Covers in a Hypergraph

Let \((V, E)\) be a hypergraph, where \(V = \{X_1, \ldots, X_k\}\), \(E = \{X_1, \ldots, X_m\}\).

Definition

A fractional edge cover is a vector \(\mathbf{w} = (w_1, \ldots, w_m)\) s.t.
“every variable \(X_i\) is covered”: \(\sum_{j: X_i \in X_j} w_j \geq 1\).

Definition

A fractional vertex packing is a vector \(\mathbf{v} = (v_1, \ldots, v_k)\) s.t.
“every edge \(X_j\) is packed”: \(\sum_{i: X_i \in X_j} v_i \leq 1\).

Theorem

\[ \min_{\mathbf{w}} \sum w_j = \max_{\mathbf{v}} \sum v_i \overset{\text{def}}{=} \rho^*; \]

This is called the fractional edge covering number of the hypergraph.

Proof on the next slide.
Covers in a Hypergraph

Let \((V, E)\) be a hypergraph, where \(V = \{X_1, \ldots, X_k\}\), \(E = \{X_1, \ldots, X_m\}\).

Definition

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

Definition

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

Theorem

\[ \min_w \sum_j w_j = \max_v \sum_i v_i \overset{\text{def}}{=} \rho^*; \]

This is called the fractional edge covering number of the hypergraph.

Proof on the next slide.
Proof of $\min_w \sum_j w_j = \max_v \sum_i v_i$

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

minimize $w_1 + w_2 + w_3$

Cover $x_1$: $w_1 + w_3 \geq 1$

Cover $x_2$: $w_1 + w_2 \geq 1$

Cover $x_3$: $w_2 + w_3 \geq 1$

maximize $v_1 + v_2 + v_3$

Pack $\{x_1, x_2\}$: $v_1 + v_2 \leq 1$

Pack $\{x_2, x_3\}$: $v_2 + v_3 \leq 1$

Pack $\{x_3, x_1\}$: $v_1 + v_3 \geq 1$

These two linear programs are dual, hence
$\min(w_1 + w_2 + w_3) = \max(v_1 + v_2 + v_3)$. 
Proof of $\min_w \sum_j w_j = \max_v \sum_i v_i$

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

minimize $w_1 + w_2 + w_3$

Cover $x_1$: $w_1 + w_3 \geq 1$
Cover $x_2$: $w_1 + w_2 \geq 1$
Cover $x_3$: $w_2 + w_3 \geq 1$

maximize $v_1 + v_2 + v_3$

Pack $\{x_1, x_2\}$: $v_1 + v_2 \leq 1$
Pack $\{x_2, x_3\}$: $v_2 + v_3 \leq 1$
Pack $\{x_3, x_1\}$: $v_1 + v_3 \geq 1$

These two linear programs are dual, hence
$\min(w_1 + w_2 + w_3) = \max(v_1 + v_2 + v_3)$. 
Proof of $\min_w \sum_j w_j = \max_v \sum_i v_i$

We use the strong duality theorem for linear programs. Will illustrate on the triangle query:

$G = (\{x_1, x_2, x_3\}, \{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_1\})$.

\[
\begin{align*}
\text{minimize} & \quad w_1 + w_2 + w_3 \\
\text{Cover } x_1: & \quad w_1 + w_3 \geq 1 \\
\text{Cover } x_2: & \quad w_1 + w_2 \geq 1 \\
\text{Cover } x_3: & \quad w_2 + w_3 \geq 1 \\
\end{align*}
\begin{align*}
\text{maximize} & \quad v_1 + v_2 + v_3 \\
\text{Pack } \{x_1, x_2\}: & \quad v_1 + v_2 \leq 1 \\
\text{Pack } \{x_2, x_3\}: & \quad v_2 + v_3 \leq 1 \\
\text{Pack } \{x_3, x_1\}: & \quad v_1 + v_3 \geq 1 \\
\end{align*}
\]

These two linear programs are dual, hence

$\min(w_1 + w_2 + w_3) = \max(v_1 + v_2 + v_3)$. 
Proof of \( \min_w \sum_j w_j = \max_v \sum_i v_i \)

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

- **minimize** \( w_1 + w_2 + w_3 \)
  - Cover \( x_1 \): \( w_1 + w_3 \geq 1 \)
  - Cover \( x_2 \): \( w_1 + w_2 \geq 1 \)
  - Cover \( x_3 \): \( w_2 + w_3 \geq 1 \)

- **maximize** \( v_1 + v_2 + v_3 \)
  - Pack \( \{x_1, x_2\} \): \( v_1 + v_2 \leq 1 \)
  - Pack \( \{x_2, x_3\} \): \( v_2 + v_3 \leq 1 \)
  - Pack \( \{x_3, x_1\} \): \( v_1 + v_3 \geq 1 \)

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 hypergraphs:

\[
R(x, y) \land S(y, z) \land T(z, x) \\
R(x, y) \land S(y, z) \land T(z, u) \land K(u, v) \\
R(x, y, z) \land S(y, z, u) \land T(z, u, x) \land K(u, x, y)
\]

- For integral edge covers / vertex packings, we only have ≥.
Shearer’s Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$. $H = \text{entropic function.}$

**Theorem (Shearer version 1)**

If $w_1, \ldots, w_m$ is a fractional edge cover then

$$w_1 H(X_1) + \cdots + w_m H(X_m) \geq H(X_1 \cdots X_k)$$

**Theorem (Shearer version 2)**

If every variable $X_i$ is $k$-covered (i.e. occurs in at least $k$ hyperedges), then

$$H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$$

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.
Shearer’s Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$. $H = \text{entropic function.}$

**Theorem (Shearer version 1)**

If $w_1, \ldots, w_m$ is a fractional edge cover then

$$w_1 H(X_1) + \cdots + w_m H(X_m) \geq H(X_1 \cdots X_k)$$

**Theorem (Shearer version 2)**

If every variable $X_i$ is $k$-covered (i.e. occurs in at least $k$ hyperedges), then

$$H(X_1) + \cdots + H(X_m) \geq k H(X_1 \cdots X_k)$$

Example:

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

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

The two formulations are equivalent why?

We will prove version 2, by generalizing the proof in the triangle query.
Shearer’s Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$. $H$ = entropic function.

**Theorem (Shearer version 1)**

If $w_1, \ldots, w_m$ is a fractional edge cover then

$$w_1 H(X_1) + \cdots + w_m H(X_m) \geq H(X_1 \cdots X_k)$$

**Theorem (Shearer version 2)**

If every variable $X_i$ is $k$-covered (i.e. occurs in at least $k$ hyperedges), then

$$H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$$

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.
Shearer’s Inequality

Hypergraph \( V = \{X_1, \ldots, X_k\} \), \( E = \{X_1, \ldots, X_m\} \). \( H = \) entropic function.

**Theorem (Shearer version 1)**

If \( w_1, \ldots, w_m \) is a fractional edge cover then

\[
w_1 H(X_1) + \cdots + w_m H(X_m) \geq H(X_1 \cdots X_k)
\]

**Theorem (Shearer version 2)**

If every variable \( X_i \) is \( k \)-covered (i.e. occurs in at least \( k \) hyperedges), then

\[
H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)
\]

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.
Shearer’s Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$. $H = \text{entropic function}$.

**Theorem (Shearer version 1)**

If $w_1, \ldots, w_m$ is a fractional edge cover then

$$w_1 H(X_1) + \cdots + w_m H(X_m) \geq H(X_1 \cdots X_k)$$

**Theorem (Shearer version 2)**

If every variable $X_i$ is $k$-covered (i.e. occurs in at least $k$ hyperedges), then

$$H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$$

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.
Shearer’s Inequality

Hypergraph $V = \{X_1, \ldots, X_k\}$, $E = \{X_1, \ldots, X_m\}$. $H = \text{entropic function}$.

**Theorem (Shearer version 1)**

If $w_1, \ldots, w_m$ is a fractional edge cover then

$$w_1 H(X_1) + \ldots + w_m H(X_m) \geq H(X_1 \ldots X_k)$$

**Theorem (Shearer version 2)**

If every variable $X_i$ is $k$-covered (i.e. occurs in at least $k$ hyperedges), then

$$H(X_1) + \ldots + H(X_m) \geq k H(X_1 \ldots X_k)$$

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 $H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$

A *sub-modularity step* consists of replacing $H(X_i) + H(X_j)$ with $H(X_i \cap X_j) + H(X_i \cup X_j)$

**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 $H(X_i) + H(X_j)$; it occurs the same number of times in $H(X_i \cap X_j) + H(X_i \cup X_j)$. why?
Proof of $H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$

A sub-modularity step consists of replacing
$H(X_i) + H(X_j)$ with $H(X_i \cap X_j) + H(X_i \cup X_j)$

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 $H(X_i) + H(X_j)$; it occurs the same number of times in $H(X_i \cap X_j) + H(X_i \cup X_j)$. why?
Proof of \( H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k) \)

A *sub-modularity step* consists of replacing 
\[ H(X_i) + H(X_j) \text{ with } H(X_i \cap X_j) + H(X_i \cup X_j) \]

**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 \( H(X_i) + H(X_j) \); it occurs *the same* number of times in \( H(X_i \cap X_j) + H(X_i \cup X_j) \). why?
Proof of $H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$

Claim 2: Progress If $X_i \not\subseteq X_j$ and $X_j \not\subseteq X_i$ then, after an SM step, the quantity $\sum_{\ell} |X_\ell|^2$ strictly increases.

Proof: $|X_i|^2 + |X_j|^2 < |X_i \cap X_j|^2 + |X_i \cup X_j|^2$ why?
Proof of $H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k)$

Claim 2: Progress If $X_i \not\subseteq X_j$ and $X_j \not\subseteq X_i$ then, after an SM step, the quantity $\sum_{\ell} |X_{\ell}|^2$ strictly increases.

Proof: $|X_i|^2 + |X_j|^2 < |X_i \cap X_j|^2 + |X_i \cup X_j|^2$ why?
Proof of \( H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k) \)

**Claim 3: Termination** We have proven:

\[
H(X_1) + \cdots + H(X_m) \geq H(Y_1) + \cdots + H(Y_m)
\]

where every variable is \( k \)-covered by \( Y_1, \ldots, Y_m \) (invariant!) and \( Y_1 \supseteq Y_2 \supseteq Y_3 \supseteq \cdots \) (no more progress!)

That means that \( Y_1 = Y_2 = \cdots = Y_k = \{X_1, \ldots, X_k\} \) why?, thus:

\[
H(X_1) + \cdots + H(X_m) \geq kH(X_1 \cdots X_k) + [\text{stuff}] \geq H(X_1 \cdots X_k)
\]
Proof of $H(X_1) + \cdots + H(X_m) \geq kH(X_1\cdots X_k)$

Claim 3: Termination We have proven:

$$H(X_1) + \cdots + H(X_m) \geq H(Y_1) + \cdots + H(Y_m)$$

where every variable is $k$-covered by $Y_1, \ldots, Y_m$ (invariant!)
and $Y_1 \supseteq Y_2 \supseteq Y_3 \supseteq \cdots$ (no more progress!)

That means that $Y_1 = Y_2 = \cdots = Y_k = \{X_1, \ldots, X_k\}$ why?, thus:

$$H(X_1) + \cdots + H(X_m) \geq kH(X_1\cdots X_k) + [\text{stuff}] \geq H(X_1\cdots X_k)$$
Discussion

- We proved something stronger: Shearer’s inequality holds for all polymatroids $H$.

- The converse also holds: if $\sum_j w_j H(X_j) \geq H(X_1 \ldots X_k)$ for all entropic functions, then $w_1, \ldots, w_k$ is a fractional edge cover.

- Next: the AGM bound is Shear’s lemma restated in terms of a query PLUS a proof that the inequality is tight.
AGM Bound for \( Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m) \)

Assume uniform statistics \(|R_1|, |R_2|, \ldots, |R_m| \leq N\).

**Lemma**

(a) If \( w_1, \ldots, w_m \) is a fractional edge cover, then \( \forall D, |Q(D)| \leq N^{w_1 + \cdots + w_m} \).

(b) If \( v_1, \ldots, v_k \) is a fractional vertex packing, then \( \exists D, |Q(D)| = N^{v_1 + \cdots + v_k} \).

Proof. (a) \( \log \max |Q(D)| \leq \max H(X) \leq \sum_j w_j H(X_j) \) (Shearer)

(b) “Product database”: \( R_j^D \overset{\text{def}}{=} \prod_{X_i \in X_j} [N^{v_i}] \).

Then \( |R_j^D| \leq N, \forall j, \) and \( Q(D) = N^{v_1 + \cdots + v_k} \)

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

\[
R^D \overset{\text{def}}{=} \left[ \frac{N}{2} \right] \times \left[ \frac{N}{2} \right] \\
S^D \overset{\text{def}}{=} \left[ \frac{N}{2} \right] \times \left[ \frac{N}{2} \right] \\
T^D \overset{\text{def}}{=} \left[ \frac{N}{2} \right] \times \left[ \frac{N}{2} \right]
\]

Then \(|R^D|, |S^D|, |T^D| \leq N, \) and \( Q(D) = \left[ \frac{N}{2} \right] \times \left[ \frac{N}{2} \right] \times \left[ \frac{N}{2} \right] \)
AGM Bound for \( Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m) \)

Assume uniform statistics \( |R_1|, |R_2|, \ldots, |R_m| \leq N \).

**Lemma**

(a) If \( w_1, \ldots, w_m \) is a fractional edge cover, then \( \forall D, \ |Q(D)| \leq N^{w_1 + \cdots + w_m} \).

(b) If \( v_1, \ldots, v_k \) is a fractional vertex packing, then \( \exists D, \ |Q(D)| = N^{v_1 + \cdots + v_k} \).

Proof. (a) \( \log \max |Q(D)| \leq \max H(X) \leq \sum_j w_j H(X_j) \) (Shearer)

(b) “Product database”: \( R_j^D \overset{\text{def}}{=} \prod_{X_i \in X_j} \left[ N^{v_i} \right] \).

Then \( |R_j^D| \leq N, \forall j, \) and \( Q(D) = N^{v_1 + \cdots + v_k} \).

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

\[
R^D = \left[ N^{1/2} \right] \times \left[ N^{1/2} \right] \quad S^D = \left[ N^{1/2} \right] \times \left[ N^{1/2} \right] \quad T^D = \left[ N^{1/2} \right] \times \left[ N^{1/2} \right]
\]

Then \( |R^D|, |S^D|, |T^D| \leq N, \) and \( Q(D) = \left[ N^{1/2} \right] \times \left[ N^{1/2} \right] \times \left[ N^{1/2} \right] \).
AGM Bound for \( Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m) \)

Assume uniform statistics \(|R_1|, |R_2|, \ldots, |R_m| \leq N\).

**Lemma**

(a) If \( w_1, \ldots, w_m \) is a fractional edge cover, then \( \forall D, |Q(D)| \leq N^{w_1 + \cdots + w_m} \).

(b) If \( v_1, \ldots, v_k \) is a fractional vertex packing, then \( \exists D, |Q(D)| = N^{v_1 + \cdots + v_k} \).

Proof. (a) \( \log \max |Q(D)| \leq \max H(X) \leq \sum_j w_j H(X_j) \) (Shearer)

(b) “Product database”: \( R^D_j \overset{\text{def}}{=} \prod_{X_i \in X_j} [N^{v_i}] \).

Then \( |R^D_j| \leq N, \ \forall j, \) and \( Q(D) = N^{v_1 + \cdots + v_k} \)

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

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

Then \( |R^D|, |S^D|, |T^D| \leq N, \) and \( Q(D) = [N^{1/2}] \times [N^{1/2}] \times [N^{1/2}] \).
AGM Bound for $Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m)$

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

**Lemma**

(a) If $w_1, \ldots, w_m$ is a fractional edge cover, then $\forall D, |Q(D)| \leq N^{w_1 + \cdots + w_m}$.

(b) If $v_1, \ldots, v_k$ is a fractional vertex packing, then $\exists D, |Q(D)| = N^{v_1 + \cdots + v_k}$

**Proof.** (a) $\log \max |Q(D)| \leq \max H(X) \leq \sum_j w_j H(X_j)$ (Shearer)

(b) “Product database”: $R_j^D \overset{\text{def}}{=} \prod_{x_i \in X_j} [N^{v_i}]$. Then $|R_j^D| \leq N, \forall j$, and $Q(D) = N^{v_1 + \cdots + v_k}$

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

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

Then $|R^D|, |S^D|, |T^D| \leq N$, and $Q(D) = [N^{1/2}] \times [N^{1/2}] \times [N^{1/2}]$
AGM Bound for \( Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m) \)

Assume uniform statistics \(|R_1|, |R_2|, \ldots, |R_m| \leq N\).

**Lemma**

(a) If \( w_1, \ldots, w_m \) is a fractional edge cover, then \( \forall D, |Q(D)| \leq N^{w_1 + \cdots + w_m} \).

(b) If \( v_1, \ldots, v_k \) is a fractional vertex packing, then \( \exists D, |Q(D)| = N^{v_1 + \cdots + v_k} \).

**Proof.** (a) \( \log \max |Q(D)| \leq \max H(X) \leq \sum_j w_j H(X_j) \) (Shearer)

(b) “Product database”: \( R^D_j \defeq \prod_{X_i \in X_j} [N^{v_i}] \).

Then \( |R^D_j| \leq N, \forall j \), and \( Q(D) = N^{v_1 + \cdots + v_k} \)

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

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

Then \( |R^D|, |S^D|, |T^D| \leq N \), and \( Q(D) = [N^{1/2}] \times [N^{1/2}] \times [N^{1/2}] \)
Theorem (AGM Bound - Uniform cardinalities)

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

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

Proof:

- \( \log \max |Q(D)| \leq \max H(X) \) was the proof by example.

- \( H(X) \leq \sum w_j H(X_j) = \rho^* \log N \) Shearer’s inequality.

- \( N^{\rho^*} \leq \max |Q(D)| \) worst-case (product) instance \( D \).
AGM Bound

Theorem (AGM Bound - Uniform cardinalities)

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

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

Proof:

- \( \log \max |Q(D)| \leq \max H(X) \) was the proof by example.

- \( H(X) \leq \sum w_j H(X_j) = \rho^* \log N \) Shearer’s inequality.

- \( N^{\rho^*} \leq \max |Q(D)| \) worst-case (product) instance \( D \).
AGM Bound

Theorem (AGM Bound - Uniform cardinalities)
\[
\max |Q(D)| = \max 2^{H(X)} = N^{\rho^*}
\]

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

Proof:
- \( \log \max |Q(D)| \leq \max H(X) \) was the proof by example.
- \( H(X) \leq \sum w_j H(X_j) = \rho^* \log N \) Shearer’s inequality.
- \( N^{\rho^*} \leq \max |Q(D)| \) worst-case (product) instance \( D \).
AGM Bound

**Theorem (AGM Bound - Uniform cardinalities)**

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

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

**Proof:**

- \( \log \max |Q(D)| \leq \max H(X) \) was the proof by example.

- \( H(X) \leq \sum w_j H(X_j) = \rho^* \log N \) Shearer’s inequality.

- \( N^{\rho^*} \leq \max |Q(D)| \) worst-case (product) instance \( D \).
AGM Bound for $Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m)$

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 $R_j(X_j)$: $\sum_{i: X_i \in X_j} v_i \leq \log N_j$.

Lemma

(a) If $w_1, \ldots, w_m$ is a fractional edge cover, then $\forall D$, $|Q(D)| \leq N_1^{w_1} \cdots N_m^{w_m}$.

(b) If $v_1, \ldots, v_k$ is a generalized frac vertex packing, $\exists D$, $|Q(D)| = 2^{v_1 + \cdots + v_k}$

Proof: straightforward generalization of the previous arguments. (Will skip in class, but it really helps if you review it at home.)
AGM Bound for \( Q(X_1, \ldots, X_k) = R_1(X_1) \land \cdots \land R_m(X_m) \)

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 \( R_j(X_j) \): \( \sum_{i : X_i \in X_j} v_i \leq \log N_j \).

**Lemma**

(a) If \( w_1, \ldots, w_m \) is a fractional edge cover, then \( \forall D, |Q(D)| \leq N_1^{w_1} \cdots N_m^{w_m} \).

(b) If \( v_1, \ldots, v_k \) is a generalized frac vertex packing, \( \exists D, |Q(D)| = 2^{v_1+\cdots+v_k} \)

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)

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

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

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find max \( Q(D) \)

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

\[
\begin{array}{ccc|ccc}
 w_R & w_S & w_T & N_R^{w_R} \cdot N_S^{w_S} \cdot N_T^{w_T} \\
 1/2 & 1/2 & 1/2 & \sqrt{N_R N_S N_T} \\
 1 & 1 & 0 & N_R N_S \\
 0 & 1 & 1 & N_S N_T \\
 1 & 0 & 1 & N_R N_T \\
\end{array}
\]

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

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

**Find max** \( Q(D) \)

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

| \( w_R \) | \( w_S \) | \( w_T \) | \( |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T} \) |
|---|---|---|---|
| 1/2 | 1/2 | 1/2 | \( \sqrt{N_RN_SN_T} \) |
| 1 | 1 | 0 | \( N_RN_S \) |
| 0 | 1 | 1 | \( NSN_T \) |
| 1 | 0 | 1 | \( N_RN_T \) |

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

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find \( \max Q(D) \)

For any fractional edge cover \( w_R, w_S, w_T \):

\[ |Q| \leq |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T}. \]

| \( w_R \) | \( w_S \) | \( w_T \) | \( |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T} \) |
|---|---|---|---|
| 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 \( |Q(D)| \).

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find max \( Q(D) \)

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

| \( w_R \) | \( w_S \) | \( w_T \) | \( |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T} \) |
|---|---|---|---|
| 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 \( |Q(D)| \).

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find \( \max Q(D) \)

For any fractional edge cover \( w_R, w_S, w_T \):

\[ |Q| \leq |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T}. \]

| \( w_R \) | \( w_S \) | \( w_T \) | \( |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T} \) |
|---|---|---|---|
| 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 \( |Q(D)| \).

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find \( \max Q(D) \)

For any fractional edge cover \( w_R, w_S, w_T \):

\[ |Q| \leq |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T}. \]

<table>
<thead>
<tr>
<th>( w_R )</th>
<th>( w_S )</th>
<th>( w_T )</th>
<th>( \sqrt{N_R N_S N_T} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>( \sqrt{N_R N_S N_T} )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>( N_R N_S )</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>( N_S N_T )</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>( N_R N_T )</td>
</tr>
</tbody>
</table>

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

In class: what is the worst case instance \( D \)?
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

Find \( \max Q(D) \)

For any fractional edge cover \( w_R, w_S, w_T \):

\[ |Q| \leq |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T}. \]

| \( w_R \) | \( w_S \) | \( w_T \) | \( |N_R|^{w_R} \cdot |N_S|^{w_S} \cdot |N_T|^{w_T} \) |
|---|---|---|---|
| 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 \(|Q(D)|\).

**In class:** what is the worst case instance \( D \)?
Example

In class:

\[ Q(x, y) = R(x) \land S(x, y) \land T(y) \]

Find max \( Q(D) \)
The worst case database, where $Q(D) = AGM(Q)$ is a product database.

To compute $AGM(Q)$ we need to compute $\min_{w} \mathcal{N}_j^{w_j}$ where $w$ ranges over all fractional edge covers.

There are infinitely many $w$'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:

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

A polytope is convex: if $w_1, w_2 \in P$ then $(1 - \lambda) w_1 + \lambda w_2 \in P$.

Call $w \in P$ a vertex if it is no strict convex combination\(^1\) of points in $P$.

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

It follows, for the edge-covering polytope:

$$\min_{w \in P} N_{jw} = \min_{w \in \text{vertices}(P)} N_{jw}$$

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

---

\(^1\)A strict convex combination is $w = (1 - \lambda) w_1 + \lambda w_2$ with $\lambda \neq 0, \lambda \neq 1$. 
Vertices of the Edge Covering Polytope

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

\[
P = \cap_i \{ \mathbf{w} \mid \sum_j a_{ij} w_j \leq b_j \}
\]

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

Call \( \mathbf{w} \in P \) a vertex if it is no strict convex combination\(^1\) of points in \( P \).

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

It follows, for the edge-covering polytope:

\[
\min_{\mathbf{w} \in P} N^w_j = \min_{\mathbf{w} \in \text{vertices}(P)} N^w_j
\]

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

\(^1\)A strict convex combination is \( \mathbf{w} = (1 - \lambda) \mathbf{w}_1 + \lambda \mathbf{w}_2 \) with \( \lambda \neq 0, \lambda \neq 1 \).
Vertices of the Edge Covering Polytope

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

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

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

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

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

It follows, for the edge-covering polytope:

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

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

\[\text{1A strict convex combination is } \mathbf{w} = (1 - \lambda) \mathbf{w}_1 + \lambda \mathbf{w}_2 \text{ with } \lambda \neq 0, \lambda \neq 1.\]
Vertices of the Edge Covering Polytope

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

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

A polytope is convex: if $w_1, w_2 \in P$ then $(1 - \lambda)w_1 + \lambda w_2 \in P$.

Call $w \in P$ a vertex if it is no strict convex combination of points in $P$.

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

It follows, for the edge-covering polytope:

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

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

---

1 A strict convex combination is $w = (1 - \lambda)w_1 + \lambda w_2$ with $\lambda \neq 0, \lambda \neq 1$. 
Vertices of the Edge Covering Polytope

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

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

A polytope is convex: if $w_1, w_2 \in P$ then $(1 - \lambda) w_1 + \lambda w_2 \in P$.

Call $w \in P$ a vertex if it is no strict convex combination\(^1\) of points in $P$.

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

It follows, for the edge-covering polytope:

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

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

---

\(^1\)A strict convex combination is $w = (1 - \lambda) w_1 + \lambda w_2$ 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 $A_i \rightarrow A_j$.
Meaning: every two tuples in $R$ that agree on $A_i$ must also agree on $A_j$.

Let $\Sigma =$ set of statistics; $\Gamma =$ set of simple FD’s.

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

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

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

- If some relation $R_j$ satisfies the simple FD $A \rightarrow B$ 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: $Q(x, y, z) = R(x, y) \land S(y, z)$

Compute $AGM_{S. y \rightarrow S. z}(Q)$. 
Examples

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

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

- $AGM(Q) = N^2$
- $y \rightarrow z$ implies $\overline{Q}(x, y, z) = R(x, y, z) \land S(y, z)$
Examples

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

Example 1: $Q(x, y, z) = R(x, y) \land S(y, z)$
Compute $AGM_{S_y \rightarrow S_z}(Q)$.

- $AGM(Q) = N^2$
- $y \rightarrow z$ implies $\bar{Q}(x, y, z) = R(x, y, z) \land S(y, z)$
- $AGM_{S_y \rightarrow S_z}(Q) = N$
Examples

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

Example 1: $Q(x, y, z) = R(x, y) \land S(y, z)$
Compute $AGM_{S.y \rightarrow S.z}(Q)$.
- $AGM(Q) = N^2$
- $y \rightarrow z$ implies $\bar{Q}(x, y, z) = R(x, y, z) \land S(y, z)$
- $AGM_{S.y \rightarrow S.z}(Q) = N$

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

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

Example 1: $Q(x, y, z) = R(x, y) \land S(y, z)$

Compute $AGM_{S.y \rightarrow S.z}(Q)$.

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

Example 2: $Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x)$

Compute $AGM_{S.y \rightarrow S.z}(Q)$

- $AGM(Q) = N^{3/2}$
- $y \rightarrow z$ implies $\bar{Q}(x, y, z) = R(x, y, z) \land S(y, z) \land T(z, x)$
Examples

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

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

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

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

- $AGM(Q) = N^{3/2}$
- $y \rightarrow z$ implies $\bar{Q}(x, y, z) = R(x, y, z) \land S(y, z) \land T(z, x)$
- $AGM_{S.y\rightarrow S.z}(Q) = N$
Worst Case Optimal Algorithm

Problem: find an algorithm to compute $Q(D)$ 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

\[ Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m) \]

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

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

\(Q[v/x]\) is the residual query, where \(x\) is substituted with constant \(v\).
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

\[
\text{let } D_R = \Pi_x(R), \ D_T = \Pi_x(T) \\
\text{for } u \text{ in } D_R \cap D_T \text{ do} \\
    // compute query \ R(u, y) \land S(y, z) \land T(z, u) \\
\text{let } D_R = \Pi_y(\sigma_{x=u}(R)), \ D_S = \Pi_y(S) \\
\text{for } v \text{ in } D_R \cap D_S \text{ do} \\
    // compute query \ R(u, v) \land S(v, z) \land T(z, u) \\
\text{let } D_S = \Pi_z(\sigma_{y=v}(S)), \ D_T = \Pi_z(\sigma_{x=u}(T)) \\
\text{for } w \text{ in } D_S \cap D_T \text{ do} \\
    \text{print } u, v, w
\]

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

**let** \( D_R = \Pi_x(R), D_T = \Pi_x(T) \)

**for** \( u \) in \( D_R \cap D_T \) **do**

// compute query \( R(u, y) \land S(y, z) \land T(z, u) \)

**let** \( D_R = \Pi_y(\sigma_{x=u}(R)), D_S = \Pi_y(S) \)

**for** \( v \) in \( D_R \cap D_S \) **do**

// compute query \( R(u, v) \land S(v, z) \land T(z, u) \)

**let** \( D_S = \Pi_z(\sigma_{y=v}(S)), D_T = \Pi_z(\sigma_{x=u}(T)) \)

**for** \( w \) in \( D_S \cap D_T \) **do**

print \( u, v, w \)

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

let \( D_R = \Pi_x(R) \), \( D_T = \Pi_x(T) \)
for \( u \) in \( D_R \cap D_T \) do

// compute query \( R(u, y) \land S(y, z) \land T(z, u) \)
let \( D_R = \Pi_y(\sigma_{x=u}(R)) \), \( D_S = \Pi_y(S) \)
for \( v \) in \( D_R \cap D_S \) do

// compute query \( R(u, v) \land S(v, z) \land T(z, u) \)
let \( D_S = \Pi_z(\sigma_{y=v}(S)) \), \( D_T = \Pi_z(\sigma_{x=u}(T)) \)
for \( w \) in \( D_S \cap D_T \) do

print \( u, v, w \)

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

\[
\text{let } D_R = \Pi_x(R), \quad D_T = \Pi_x(T) \\
\text{for } u \text{ in } D_R \cap D_T \text{ do} \\
\quad \text{// compute query } R(u, y) \land S(y, z) \land T(z, u) \\
\text{let } D_R = \Pi_y(\sigma_{x=u}(R)), \quad D_S = \Pi_y(S) \\
\text{for } v \text{ in } D_R \cap D_S \text{ do} \\
\quad \text{// compute query } R(u, v) \land S(v, z) \land T(z, u) \\
\text{let } D_S = \Pi_z(\sigma_{y=v}(S)), \quad D_T = \Pi_z(\sigma_{x=u}(T)) \\
\text{for } w \text{ in } D_S \cap D_T \text{ do} \\
\quad \text{print } u, v, w
\]

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

let \( D_R = \Pi_x(R) \), \( D_T = \Pi_x(T) \)

for \( u \) in \( D_R \cap D_T \) do

// compute query \( R(u, y) \land S(y, z) \land T(z, u) \)

let \( D_R = \Pi_y(\sigma_{x=u}(R)) \), \( D_S = \Pi_y(S) \)

for \( v \) in \( D_R \cap D_S \) do

// compute query \( R(u, v) \land S(v, z) \land T(z, u) \)

let \( D_S = \Pi_z(\sigma_{y=v}(S)) \), \( D_T = \Pi_z(\sigma_{x=u}(T)) \)

for \( w \) in \( D_S \cap D_T \) do

print \( u, v, w \)

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

let \( D_R = \Pi_x(R), \quad D_T = \Pi_x(T) \)

for \( u \) in \( D_R \cap D_T \) do

// compute query \( R(u, y) \land S(y, z) \land T(z, u) \)
let \( D_R = \Pi_y(\sigma_{x=u}(R)), \quad D_S = \Pi_y(S) \)

for \( v \) in \( D_R \cap D_S \) do

// compute query \( R(u, v) \land S(v, z) \land T(z, u) \)
let \( D_S = \Pi_z(\sigma_{y=v}(S)), \quad D_T = \Pi_z(\sigma_{x=u}(T)) \)

for \( w \) in \( D_S \cap D_T \) do

print \( u, v, w \)

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

let \( D_R = \Pi_x(R) \), \( D_T = \Pi_x(T) \)
for \( u \) in \( D_R \cap D_T \) do

// compute query \( R(u, y) \land S(y, z) \land T(z, u) \)
let \( D_R = \Pi_y(\sigma_{x=u}(R)) \), \( D_S = \Pi_y(S) \)
for \( v \) in \( D_R \cap D_S \) do

// compute query \( R(u, v) \land S(v, z) \land T(z, u) \)
let \( D_S = \Pi_z(\sigma_{y=v}(S)) \), \( D_T = \Pi_z(\sigma_{x=u}(T)) \)
for \( w \) in \( D_S \cap D_T \) do

print \( u, v, w \)

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

```plaintext
let \( D_R = \Pi_x(R) \), \( D_T = \Pi_x(T) \)
for \( u \) in \( D_R \cap D_T \) do
    // compute query \( R(u, y) \land S(y, z) \land T(z, u) \)
    let \( D_R = \Pi_y(\sigma_{x=u}(R)) \), \( D_S = \Pi_y(S) \)
    for \( v \) in \( D_R \cap D_S \) do
        // compute query \( R(u, v) \land S(v, z) \land T(z, u) \)
        let \( D_S = \Pi_z(\sigma_{y=v}(S)) \), \( D_T = \Pi_z(\sigma_{x=u}(T)) \)
            for \( w \) in \( D_S \cap D_T \) do
                print \( u, v, w \)
```

Next: we will prove its runtime.
Example

\[ Q(x, y, z) = R(x, y) \land S(y, z) \land T(z, x) \]

\[
\text{let } D_R = \Pi_x(R), \quad D_T = \Pi_x(T) \\
\text{for } u \text{ in } D_R \cap D_T \text{ do} \\
\quad // \text{ compute query } R(u, y) \land S(y, z) \land T(z, u) \\
\quad \text{let } D_R = \Pi_y(\sigma_{x=u}(R)), \quad D_S = \Pi_y(S) \\
\quad \text{for } v \text{ in } D_R \cap D_S \text{ do} \\
\quad \quad // \text{ compute query } R(u, v) \land S(v, z) \land T(z, u) \\
\quad \quad \text{let } D_S = \Pi_z(\sigma_{y=v}(S)), \quad D_T = \Pi_z(\sigma_{x=u}(T)) \\
\quad \quad \text{for } w \text{ in } D_S \cap D_T \text{ do} \\
\quad \quad \quad \text{print } u, v, w
\]

Next: we will prove its runtime.
Runtime of GJ

\[ Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_m(X_m) \]

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

**Theorem**

Let \( w_1, \ldots, w_m \) 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 $x$

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

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

\[
\sum_{i} a_i^{w_1} b_i^{w_2} c_i^{w_3} \ldots \leq \left( \sum_{i} a_i \right)^{w_1} \left( \sum_{i} b_i \right)^{w_2} \left( \sum_{i} c_i \right)^{w_3} \ldots
\]
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

$$\sum_i a_i^{w_1} b_i^{w_2} c_i^{w_3} \ldots \leq \left( \sum_i a_i \right)^{w_1} \left( \sum_i b_i \right)^{w_2} \left( \sum_i c_i \right)^{w_3} \ldots$$
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
\[ \sum_i a_i^{w_1} b_i^{w_2} c_i^{w_3} \ldots \leq \left( \sum_i a_i \right)^{w_1} \left( \sum_i b_i \right)^{w_2} \left( \sum_i c_i \right)^{w_3} \ldots \]
Runtime of GJ: Induction Step; GJ iterates over $x_1$

$$Q(x_1, \ldots, x_k) = R_1(x_1) \land \cdots \land R_{j_0}(x_{j_0}) \land R_{j_0+1}(x_{j_0+1}) \land \cdots \land R_m(x_m)$$

- **Contain $x_1$**
- **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:

$$T_{GJ}(Q[a/x_1]) = N_1^{w_1}_{1,a} \cdot N_{j_0}^{w_{j_0}}_{j_0,a} \cdot N_{j_0+1}^{w_{j_0+1}} \cdots N_m^{w_m}$$

$$= |\sigma_{x_1=a}(R_1)| \quad \text{def} = |\sigma_{x_1=a}(R_{j_0})|$$

Total runtime is obtained by summing on $a$:

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

$$= (N_1)^{w_1} \cdot (N_{j_0})^{w_{j_0}}$$
Runtime of GJ: Induction Step; GJ iterates over $x_1$

\[
Q(x_1, \ldots, x_k) = R_1(x_1) \land \cdots \land R_{j_0}(x_{j_0}) \land R_{j_0+1}(x_{j_0+1}) \land \cdots \land R_m(x_m)
\]

- Contains $x_1$
- Doesn’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:

\[
T_{GJ}(Q[a/x_1]) = N_1^{w_1}_{1,a} \cdots N_{j_0}^{w_{j_0}}_{j_0,a} \cdot N_{j_0+1}^{w_{j_0+1}} \cdots N_m^{w_m}
\]

\[
= |\sigma_{x_1=a}(R_1)| \quad \text{def} \quad |\sigma_{x_1=a}(R_{j_0})| \quad \text{def}
\]

Total runtime is obtained by summing on $a$:

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

\[
= (N_1)^{w_1} \cdots (N_{j_0})^{w_{j_0}}
\]
Runtime of GJ: Induction Step; GJ iterates over $x_1$

$$Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_{j_0}(X_{j_0}) \land R_{j_0+1}(X_{j_0+1}) \land \cdots \land R_m(X_m)$$

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

- Time for $\prod_x(R_1) \cap \cdots \cap \prod_x(R_{j_0})$ 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:

$$T_{GJ}(Q[a/x_1]) = \underbrace{N_{1,a}^{w_1}}_a \cdots \underbrace{N_{j_0,a}^{w_{j_0}}}_a \cdot \underbrace{N_{j_0+1,a}^{w_{j_0+1}}}_a \cdots \underbrace{N_{m,a}^{w_m}}_a$$

$$\overset{\text{def}}{=} |\sigma_{x_1=a}(R_1)| \quad \overset{\text{def}}{=} |\sigma_{x_1=a}(R_{j_0})|$$

Total runtime is obtained by summing on $a$:

$$\sum_a N_{1,a}^{w_1} \cdots N_{j_0,a}^{w_{j_0}} \cdot N_{j_0+1,a}^{w_{j_0+1}} \cdots N_{m,a}^{w_m} \leq \left( \sum_a N_{1,a} \right)^{w_1} \cdots \left( \sum_a N_{j_0,a} \right)^{w_{j_0}} \cdot N_{j_0+1}^{w_{j_0+1}} \cdots N_{m}^{w_m}$$

$$=(N_1)^{w_1} \cdots (N_{j_0})^{w_{j_0}}$$
Runtime of GJ: Induction Step; GJ iterates over $x_1$

$$Q(x_1, \ldots, x_k) = R_1(X_1) \land \cdots \land R_{j_0}(X_{j_0}) \land R_{j_0+1}(X_{j_0+1}) \land \cdots \land R_m(X_m)$$

- **Contain $x_1$**
- **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:

$$T_{GJ}(Q[a/x_1]) = \underbrace{N_1^{w_1}}_{\sigma_{x_1=a}(R_1)} \cdot \underbrace{\cdots \cdot N_{j_0}^{w_{j_0}}}_{\sigma_{x_1=a}(R_{j_0})} \cdot \underbrace{N_{j_0+1}^{w_{j_0+1}} \cdots N_m^{w_m}}_{\sigma_{x_1=a}(R_m)}$$

Total runtime is obtained by summing on $a$:

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

\[= (N_1)^{w_1} \cdots (N_{j_0})^{w_{j_0}} \cdots (N_m)^{w_m} \]
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 $Q(D) \leq \prod_j N_j^{w_j}$. 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 \( H(Z|Y) \overset{\text{def}}{=} H(YZ) - H(Y) \).

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) \( H(Z|Y) \geq H(Z|XY) \) (2) \( H'(Z) \overset{\text{def}}{=} H(Z|Y) \) is a polymatroid.

Proof: (1)

\[
H(XY) + H(YZ) \geq H(XYZ) + H\left(\frac{XY \cap YZ}{Y}\right)
\]

not necessarily \( Y \) why?

\[
\geq H(XYZ) + H(Y)
\]

\[
H(YZ) - H(Y) \geq H(XYZ) - H(XY)
\]

(2) exercise.
Conditional Polymatroid/Entropy

We will define the conditional polymatroid as $H(Z|Y) \overset{\text{def}}{=} H(YZ) - H(Y)$. 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) $H(Z|Y) \geq H(Z|XY)$  (2) $H'(Z) \overset{\text{def}}{=} H(Z|Y)$ is a polymatroid.

**Proof:** (1)

\[
H(XY) + H(YZ) \geq H(XYZ) + H((XY) \cap (YZ))
\]

\[
\geq H(XYZ) + H(Y)
\]

\[
H(YZ) - H(Y) \geq H(XYZ) - H(XY)
\]

(2) exercise.
Conditional Polymatroid/Entropy

We will define the conditional polymatroid as

$$H(Z|Y) \overset{\text{def}}{=} H(YZ) - H(Y).$$

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) $H(Z|Y) \geq H(Z|XY)$

(2) $H'(Z) \overset{\text{def}}{=} H(Z|Y)$ is a polymatroid.

Proof: (1)

$$H(XY) + H(YZ) \geq H(XYZ) + H((XY) \cap (YZ))$$

not necessarily $Y$ why?

$$\geq H(XYZ) + H(Y)$$

$$H(YZ) - H(Y) \geq H(XYZ) - H(XY)$$

(2) exercise.
Conditional Polymatroid/Entropy

We will define the conditional polymatroid as $H(Z|Y) \overset{\text{def}}{=} H(YZ) - H(Y)$. 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

$(1)$ $H(Z|Y) \geq H(Z|XY)$  
$(2)$ $H'(Z) \overset{\text{def}}{=} H(Z|Y)$ is a polymatroid.

Proof: (1)

\[
H(XY) + H(YZ) \geq H(XYZ) + H((XY) \cap (YZ)) \\
\geq H(XYZ) + H(Y) \\
H(YZ) - H(Y) \geq H(XYZ) - H(XY)
\]

(2) exercise.
Proof #2 of Shearer’s Inequality

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

\[
\begin{align*}
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) &+ \left( \ldots + w_m H(X_m) \right) = \\
\text{contain } X_1 &\quad \text{do not contain } X_1 \\
= (w_1 + \ldots + w_{j_0}) H(X_1) &+ \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + (\ldots + H(X_m)) \\
\geq &+ (w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1)) + (\ldots + H(X_m|X_1)) \\
\geq &+ H(X_1) + H(X_1 X_2 \ldots X_k|X_1) \\
= &+ H(X_1 X_2 \ldots X_k)
\end{align*}
\]
Proof #2 of Shearer’s Inequality

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

\[
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) + \left( \ldots + w_m H(X_m) \right) = \\
\underbrace{(w_1 + \ldots + w_{j_0}) H(X_1)}_{\text{contain } X_1} + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m) \right) \\
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m) \right) \\
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m|X_1) \right) \\
\geq H(X_1) + H(X_1X_2 \ldots X_k|X_1) \\
= H(X_1X_2 \ldots X_k)
\]
Proof #2 of Shearer’s Inequality

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

$$
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) + \left( \ldots + w_m H(X_m) \right) = \\
\begin{array}{l}
\text{contain } X_1 \\
\text{do not contain } X_1
\end{array}
$$

$$
= (w_1 + \ldots + w_{j_0}) H(X_1) + \left( w_1 H(X_1 | X_1) + \ldots + w_{j_0} H(X_{j_0} | X_1) \right) + (\ldots + H(X_m)) \\
\geq H(X_1) + \left( w_1 H(X_1 | X_1) + \ldots + w_{j_0} H(X_{j_0} | X_1) \right) + (\ldots + H(X_m)) \\
\geq H(X_1) + H(X_1 X_2 \ldots X_k | X_1) \\
= H(X_1 X_2 \ldots X_k)
$$
Proof #2 of Shearer’s Inequality

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

\[
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) + \left( \ldots + w_m H(X_m) \right) = \\
\underbrace{\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right)}_{\text{contain } X_1} + \underbrace{\left( \ldots + w_m H(X_m) \right)}_{\text{do not contain } X_1}
\]

\[
= (w_1 + \ldots + w_{j_0}) H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m) \right)
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m) \right)
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + \left( \ldots + H(X_m|X_1) \right)
= H(X_1 X_2 \ldots X_k | X_1)
\]
Proof #2 of Shearer’s Inequality

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

\[
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) + \left( \ldots + w_m H(X_m) \right) = \\
\text{contain } X_1 \quad \text{do not contain } X_1
\]

\[
=(w_1 + \ldots + w_{j_0}) H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + (\ldots + H(X_m))
\]

\[
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + (\ldots + H(X_m))
\]

\[
\geq H(X_1) + \left( w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1) \right) + (\ldots + H(X_m|X_1))
\]

\[
\geq H(X_1) + H(X_1 X_2 \ldots X_k|X_1)
\]

\[
=H(X_1 X_2 \ldots X_k)
\]
Proof #2 of Shearer’s Inequality

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

\[
\left( w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}) \right) + \left( \ldots + w_m H(X_m) \right) = \\
\begin{array}{c}
\text{contain } X_1 \\
\text{do not contain } X_1
\end{array}
\]

\[
= (w_1 + \ldots + w_{j_0}) H(X_1) + \left( w_1 H(X_1 | X_1) + \ldots + w_{j_0} H(X_{j_0} | X_1) \right) + (\ldots + H(X_m)) \\
\geq H(X_1) + \left( w_1 H(X_1 | X_1) + \ldots + w_{j_0} H(X_{j_0} | X_1) \right) + (\ldots + H(X_m)) \\
\geq H(X_1) + H(X_1 X_2 \ldots X_k | X_1) \\
= H(X_1 X_2 \ldots X_k)
\]
Proof #2 of Shearer’s Inequality

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

\[
(w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0})) + (\ldots + w_m H(X_m)) =
\]
\[
\underbrace{(w_1 H(X_1) + \ldots + w_{j_0} H(X_{j_0}))}_{\text{contain } X_1} + \underbrace{(\ldots + w_m H(X_m))}_{\text{do not contain } X_1}
\]
\[
=(w_1 + \ldots + w_{j_0}) H(X_1) + (w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1)) + (\ldots + H(X_m))
\]
\[
\geq H(X_1) + (w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1)) + (\ldots + H(X_m))
\]
\[
\geq H(X_1) + (w_1 H(X_1|X_1) + \ldots + w_{j_0} H(X_{j_0}|X_1)) + (\ldots + H(X_m|X_1))
\]
\[
\geq H(X_1) + H(X_1 X_2 \ldots X_k|X_1)
\]
\[
= H(X_1 X_2 \ldots X_k)
\]
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.