Size of the Query’s Output

• Fully conjunctive query Q

• Known cardinalities of input relations $|R|$, $|S|$, ...

• How large is the size of the output?
Example

• \( Q(x,y,z) :- R(x,y), S(y,z) \)

• \(|R| = N_1, |S| = N_2\)

• How large is \(|Q|?\)
  – Min =
  – Max =
Example

• $Q(x, y, z) ::= R(x, y), S(y, z)$

• $|R| = N_1, \quad |S| = N_2$

• How large is $|Q|$?
  – Min = 0
  – Max = $N_1 N_2$
Example

- Q(x,y,z) :- R(x,y), S(y,z)

- |R| = N_1, |S| = N_2

- How large is |Q|?
  - Min = 0
  - Max = N_1 N_2

- Thus 0 ≤ |Q| ≤ N_1 N_2
Example

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

- \(|R| = N_1, |S| = N_2, |T| = N_3\)

- How large is \(|Q|\)?
Example

• $Q(x,y,z) = R(x,y), S(y,z), T(z,x)$

• $|R| = N_1, \quad |S| = N_2, \quad |T| = N_3$

• How large is $|Q|$?

• $|Q| \leq N_1 N_2 N_3$
Example

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

• \(|R| = N_1, \quad |S| = N_2, \quad |T| = N_3\)

• How large is \(|Q|\)?

• \(|Q| \leq N_1N_2N_3\)
• \(|Q| \leq N_1N_2 \quad \text{and} \quad |Q| \leq N_1N_3 \quad \text{and} \quad |Q| \leq N_2N_3\)
Example

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

- \(|R| = N_1, \quad |S| = N_2, \quad |T| = N_3\)

- How large is \(|Q|\)?

- \(|Q| \leq N_1 N_2 N_3\)

- \(|Q| \leq N_1 N_2 \quad \text{and} \quad |Q| \leq N_1 N_3 \quad \text{and} \quad |Q| \leq N_2 N_3\)

- But also \(|Q| \leq (N_1 N_2 N_3)^{1/2}\)
Definition

• Let Q be a full conjunctive query without self-joins. The *hypergraph* associated to Q has:
  – Nodes = variables of Q
  – Hyperedges = atoms of Q
Examples

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

\[ Q(x,y,z) = R(x,y,z), S(x), T(y), K(z), M(x,u) \]
Edge Cover

- $G = (V, E)$ a hypergraph
  
  $V = \{x_1, \ldots, x_n\}$, \hspace{1em} $E = \{e_1, \ldots, e_m\}$

- An edge cover = set of edges $e_{i_1}, \ldots, e_{i_k}$
  \hspace{1em} s.t. forall $x \in V$, $\exists i \hspace{0.5em} x \in e_i$
Edge Cover $\rightarrow$ Query Bound

- Fact. If $R_{i1}, R_{i2}, \ldots, R_{ik}$ is an edge cover, then $|Q| \leq |R_{i1}| |R_{i2}| \ldots |R_{ik}|$

- Proof in class
Fractional Edge Cover

• $G = (V,E)$ a hypergraph
  $V = \{x_1, ..., x_n\}$,   $E = \{e_1, ..., e_m\}$

• A fractional edge cover = real numbers $w_1, ..., w_m \geq 0$
  s.t. for any $x \in V$: $\sum \{ w_i | x \in e_i \} \geq 1$

• Every edge cover is also a fractional edge cover. (Why?)
The AGM Bound

• **Theorem** [Atserias, Grohe, Marx]

  (1) If $w_1, ..., w_m \geq 0$ is a fractional edge cover, then $|Q| \leq |R_1|^{w_1} |R_2|^{w_2} ... |R_m|^{w_m}$

  (2) For any numbers $N_1, ..., N_m$, there exists a database s.t. $|R_1| \leq N_1, ..., |R_m| \leq N_m$ and a fractional edge cover $w_1, ..., w_m \geq 0$ such that $|Q| = |R_1|^{w_1} |R_2|^{w_2} ... |R_m|^{w_m}$

• We denote $\text{AGM}(Q) = \min_w |R_1|^{w_1} |R_2|^{w_2} ... |R_m|^{w_m}$
Proof of Part (2)

• $G = (V,E)$ a hypergraph
  $V = \{x_1,...,x_n\}$, $E = \{e_1, ..., e_m\}$

• $n_1, ..., n_m \geq 0$ are given numbers

• A generalized fractional vertex packing =
  = real numbers $v_1, ..., v_n \geq 0$
  s.t. for any $e_j \in E$: $\sum \{ v_i | x_i \in e_j \} \leq n_j = \log N_j$

• **Theorem** (strong duality of LP programs)
  $\min_{\text{w=frac. edge cover}} w_1 n_1 + ... + w_m n_m =$
  $= \max_{v=\text{gen. frac. vertex packing}} v_1 + ... + v_n$
Proof of the Theorem on Special Case

Q(x,y,z) = R(x,y), S(y,z), T(z,x)

Hypergraph = variables + relations

(Generalized) fractional vertex packing:

\[
\text{max}(v_R + v_S + v_T)
\]

R: \(v_x + v_y \leq \log |R|\)
S: \(v_y + v_S \leq \log |S|\)
T: \(v_x + v_z \leq \log |T|\)

Th. For any feasible \(v_R, v_S, v_T\):
\[
\log |Q| \geq \text{objective}
\]
\[
|Q| \geq n^{vx} \times n^{vy} \times n^{vz}
\]

Fractional edge cover:

\[
\text{min}(w_R \log |R| + w_S \log |S| + w_T \log |T|)
\]

x: \(w_R + w_T \geq 1\)
y: \(w_R + w_S \geq 1\)
z: \(w_S + w_T \geq 1\)

Th. For any feasible \(w_R, w_S, w_T\):
\[
\log |Q| \leq \text{objective}
\]
\[
|Q| \leq |R|^{wR} \times |S|^{wS} \times |T|^{wT}
\]

Proof “Free” instance

R(x,y) = \([n^{vx}] \times [n^{vy}]\)
S(y,z) = \([n^{vy}] \times [n^{vz}]\)
T(z,x) = \([n^{vx}] \times [n^{vz}]\)

CSE 544 - Fall 2016
Examples (in Class)

- Assume $|R|=|S|=|T|=... = N$
- Find max $|Q|$
- Describe database on which $Q$ is max

- $Q = R(x,y), S(y,z)$
- $Q = R(x,y), S(y,z), T(z,x)$
- $Q = R(x,y), S(y,z), T(z,u)$
- $Q = R(x,y), S(y,z), T(z,u), K(u,v)$
- $Q = R(x,y,z), S(x,y,u), T(x,z,u), K(y,z,u)$
- $Q = R(x,y,z,u), S(x,y,z,w), T(x,y,u,w), K(x,z,u,w), L(y,z,u,w)$
Shannon Entropy

• X = random variable (usually X = x₁, x₂, ..., xₖ)
  Has N outcomes
  with probabilities p₁, ..., pₙ
• The entropy of X is: H(X) = - [p₁ log p₁ + ... + pₙ log pₙ]

• Facts about the entropy
  • H(X) ≤ log N and it is “=” iff p₁=p₂=...=pₙ
  • H(∅) = 0
  • H(X) ≤ H(XY) monotonicity
  • H(X ∩ Y) + H(X ∪ Y) ≤ H(X) + H(Y) submodularity
Entropy for Query Bounds

\[ H = - (p_1 \log p_1 + p_2 \log p_2 + \ldots + p_N \log p_N) \]

Q(x,y,z) = R(x,y), S(y,z), T(z,x)

Probability space:

\[ H(xyz) = \log n \]

R, S, T are marginal probabilities:

\[ H(xy) \leq \log|R| \]
\[ H(yz) \leq \log|S| \]
\[ H(xz) \leq \log|T| \]
Shearer’s Inequality

- Let $X_1, ..., X_m \subseteq X$ be sets of random variables.
- Let $w_1, ..., w_m$ = fractional edge cover of this hypergraph
  nodes = $X$
  hyperedges = $X_1, ..., X_m$

- Then: $w_1 \text{H}(X_1) + ... + w_m \text{H}(X_m) \geq \text{H}(X)$ (Shearer)

- Example: $\frac{1}{2} \text{H}(xy) + \frac{1}{2} \text{H}(yz) + \frac{1}{2} \text{H}(xz) \geq \text{H}(xyz)$

- Proof: $\text{H}(xy) + \text{H}(yz) + \text{H}(xz) \geq \text{H}(xyz) + \text{H}(y) + \text{H}(xz)$
  $\geq \text{H}(xyz) + \text{H}(xyz) + \text{H}(\emptyset) = 2\text{H}(xyz)$
Proof of Shearer’s Lemma

• Restated: let $X_1, \ldots, X_m \subseteq X$ be sets of random variables such that every $x \in X$ is covered at least $k$ times

• Then: $H(X_1) + H(X_2) + \ldots + H(X_m) \geq k \cdot H(X)$

• Proof. replace $X_i, X_j$ s.t. $X_i \nsubseteq X_j$ and $X_j \nsubseteq X_i$ with $X_i \cup X_j, X_i \cap X_j$
  – Every variable $x$ continues to be $k$-covered (why?)
  – $|X_i|^2 + |X_j|^2 < |X_i \cup X_j|^2 + |X_i \cap X_j|^2$ (why?) we use $X_i \nsubseteq X_j$ and $X_j \nsubseteq X_i$

• When we stop: $X_1 \supseteq X_2 \supseteq X_3 \supseteq \ldots$

• Since each variable is $k$-covered: $X_1 = X_2 = \ldots = X_k = X$ (why?)

• Hence $H(X_1) + H(X_2) + \ldots + H(X_m) \geq k \cdot H(X) + \ldots$[rest]
Proof of AGM Part (1)

• If $w_1, \ldots, w_m \geq 0$ is a fractional edge cover, then $|Q| \leq |R_1|^{w_1} |R_2|^{w_2} \ldots |R_m|^{w_m}$

• Fix any input database, let $H$ be the entropy of the probability space defined by the output of $Q$
  – $\log |Q| = H(X)$
  – $\log |R_i| \geq H(X_i)$
  – Shearer’s inequality: $w_1 H(X_1) + \ldots + w_m H(X_m) \geq H(X)$

• It follows: $w_1 \log |R_1| + \ldots + w_m \log |R_m| \geq \log |Q|$
Summary of AGM Bound

• For any fractional vertex cover
  \[ |Q| \leq |R_1|^{w_1} |R_2|^{w_2} \ldots |R_m|^{w_m} \] and this is tight

• No query should take time more than AGM(Q)!

• However, for certain queries, any query plan has a data complexity >> AGM(Q)

• Next time: novel worst-case optimal algorithms, which run in time O(AGM(Q))