CSE 544
Principles of Database Management Systems

Fall 2016

Lecture 10 -- AGM Bound

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)
- $|R| = N_1$, $|S| = N_2$
- How large is |Q|?
  - Min = 0
  - Max = $N_1 N_2$

Example

- $Q(x,y,z) \text{ :- } R(x,y), S(y,z)$
- $|R| = N_1$, $|S| = N_2$
- How large is $|Q|$?
  - Min = 0
  - Max = $N_1 N_2$
- Thus $0 \leq |Q| \leq N_1 N_2$

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

- $Q(x,y,z) = R(x,y),S(y,z),T(z,x)$
- $|R| = N_1$, $|S| = N_2$, $|T| = N_3$
- 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 $|Q|$?
- $|Q| \leq N_1N_2N_3$
- $|Q| \leq N_1N_2$ and $|Q| \leq N_1N_3$ and $|Q| \leq N_2N_3$

Example

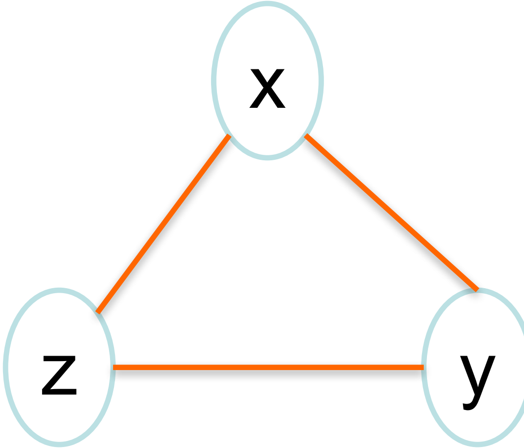
- $Q(x,y,z) = R(x,y),S(y,z),T(z,x)$
- $|R| = N_1$, $|S| = N_2$, $|T| = N_3$
- How large is $|Q|$?
- $|Q| \leq N_1N_2N_3$
- $|Q| \leq N_1N_2$ and $|Q| \leq N_1N_3$ and $|Q| \leq N_2N_3$
- But also $|Q| \leq (N_1N_2N_3)^{½}$

Definition

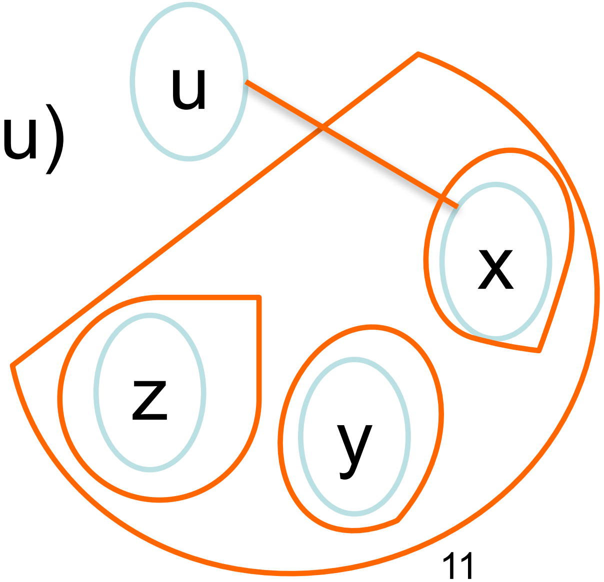
- Let Q be a full conjunctive query without self-joins. The *hypergraph* associated to Q has:
  – Nodes = variables of Q
  – Hyperedges = atoms of Q

Examples

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

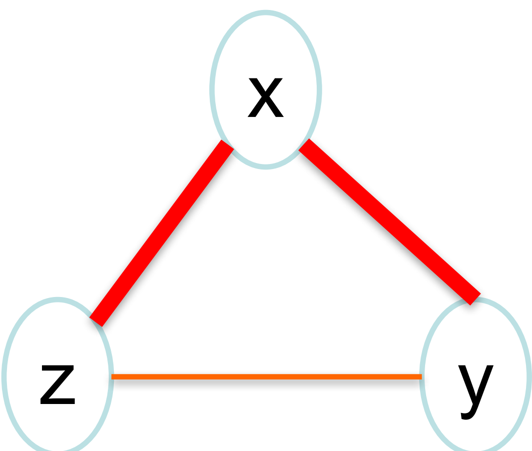


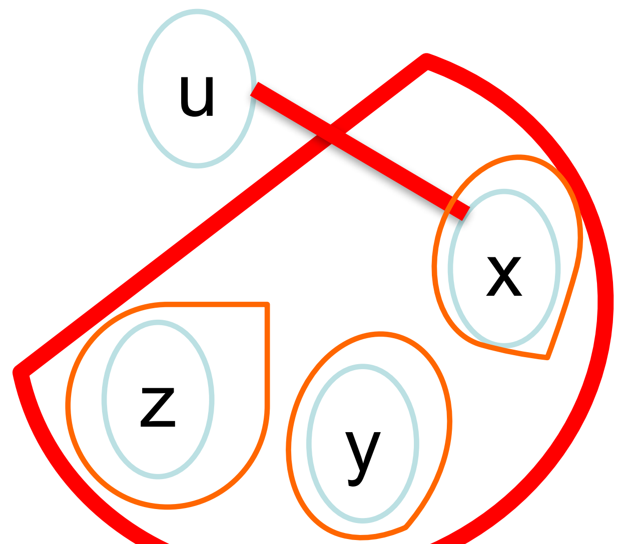
$Q(x,y,z) = R(x,y,z), S(x), T(y), K(z), M(x,u)$



Edge Cover

- G = (V,E) a hypergraph
  $V = \{x_1,\ldots, x_n\}$, $E = \{e_1, \ldots, e_m\}$
- An edge cover = set of edges $e_{i1}, \ldots, e_{ik}$
  s.t. forall $x \in V$, $\exists$ i $x \in e_i$





Edge Cover → Query Bound

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

- 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 $x \in V$: $\sum \{ w_i \mid x \in e_i \} \geq 1$

- 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 $|Q| \leq |R_1|^{w1} |R_2|^{w2} ... |R_m|^{wm}$

  (2) For any numbers N1, ..., Nm, there exists
  a database s.t. |R1| ≤ N1, ..., |Rm| ≤ Nm
  and a fractional edge cover $w_1, ..., w_m \geq 0$
  such that $|Q| = |R_1|^{w1} |R_2|^{w2} ... |R_m|^{wm}$

- We denote $AGM(Q) = \min_w |R_1|^{w1} |R_2|^{w2} ... |R_m|^{wm}$

Proof of Part (2)

- G = (V,E) a hypergraph
  $V = \{x_1, \ldots, x_n\}$, $E = \{e_1, \ldots, e_m\}$
- $n_1, \ldots, n_m \geq 0$ are given numbers
- A generalized fractional vertex packing =
  = real numbers $v_1, \ldots, v_n \geq 0$
  s.t. for any $e_j \in E$: $\sum \{ v_i \mid x_i \in e_j \} \leq n_j = \log N_j$

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

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:

$\max(v_R + v_S + v_T)$

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

Fractional edge cover:

$\min(w_R \log|R| + w_S \log|S| + w_T \log|T|)$

x:	$w_R + w_T$	≥ 1
y:	$w_R + w_S$	≥ 1
z:	$w_S + w_T$	≥ 1

Th. For any feasible v_R,v_S,v_T
$\log|Q| \geq$ objective
$|Q| \geq n^{vx} \times n^{vy} \times n^{vz}$

$\leq$

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

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

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 $X = x_1,x_2,...,x_k$)
  Has N outcomes
  with probabilities $p_1, ..., p_N$
- The entropy of X is: $H(X) = - [p_1 \log p_1 + ... + p_N \log p_N]$

- Facts about the entropy
- $H(X) \leq \log N$ and it is "=" iff $p_1=p_2=...=p_N$

- $H(\varnothing) = 0$
- $H(X) \leq H(XY)$ monotonicity
- $H(X \cap Y) + H(X \cup Y) \leq H(X) + H(Y)$ submodularity

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

[GLVV'2012]

Entropy for Query Bounds

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

Probability space:

$H(xyz) = \log n$

x	y	z	
a	3	m	1/5
a	2	q	1/5
b	2	q	1/5
d	3	m	1/5
a	3	q	1/5

R, S, T are marginal probabilities:

$H(xy) \leq \log|R|$
$H(yz) \leq \log|S|$
$H(xz) \leq \log|T|$

x	y	
a	3	2/5
a	2	1/5
b	2	1/5
d	3	1/5

y	z	
3	m	2/5
2	q	2/5
3	q	1/5

x	z	
a	m	1/5
a	q	2/5
b	q	1/5
d	m	1/5

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: $w_1 H(X_1) + ... + w_m H(X_m) \geq H(X)$ (Shearer)

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

- Proof: $H(xy) + H(yz) + H(xz) \geq H(xyz) + H(y) + H(xz)$

  $\geq H(xyz) + H(xyz) + H(\emptyset) = 2H(xyz)$

Proof of Shearer's Lemma

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

- Then: $H(X_1) + H(X_2) + ... + H(X_m) \geq k\, H(X)$

- Proof. replace X_i, X_j s.t. $X_i \not\subseteq X_j$ and $X_j \not\subseteq X_i$ 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 $X_i \not\subseteq X_j$ and $X_j \not\subseteq X_i$
- When we stop: $X_1 \supseteq X_2 \supseteq X_3 \supseteq ...$
- Since each variable is k-covered: $X_1 = X_2 = ... = X_k = X$ (why?)
- Hence $H(X_1) + H(X_2) + ... + H(X_m) \geq k\, H(X) + ...$[rest]

Proof of AGM Part (1)

- If $w_1, ..., w_m \geq 0$ is a fractional edge cover, then $|Q| \leq |R_1|^{w1} |R_2|^{w2} ... |R_m|^{wm}$
- 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: $w_1 H(X_1) + ... + w_m H(X_m) \geq H(X)$
- It follows: $w_1 \log |R_1| + ... + w_m \log |R_m| \geq \log |Q|$

Summary of AGM Bound

- For *any* fractional vertex cover
  $|Q| \leq |R_1|^{w1} |R_2|^{w2} \ldots |R_m|^{wm}$ 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))