Multi-join Query Evaluation on Big Data
Section 1

Dan Suciu

March, 2015

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

When $|R| = |S| = |T| = m$ then the optimal cover is $(1/2, 1/2, 1/2)$ and $AGM(Q) = m^{3/2}$.

Problem 1

Prove that this bound is tight. Construct 3 relations R, S, T each of size m s.t. there are $m^{3/2}$ triangles.

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

When $|R| = |S| = |T| = m$ then the optimal cover is $(1/2, 1/2, 1/2)$ and $AGM(Q) = m^{3/2}$.

Problem 1

Prove that this bound is tight. Construct 3 relations R, S, T each of size m s.t. there are $m^{3/2}$ triangles.

Solution: $R = S = T = [m^{1/2}] \times [m^{1/2}] \times [m^{1/2}]$

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

Problem 2

Prove that this AGM bound is tight for arbitrary cardinalities m_R, m_S, m_T.
Construct relations R, S, T that have $\min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$ triangles.

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

Solution: write the primal and the dual LP:

$$\text{minimize}(u_R \log m_R + u_S \log m_S + u_T \log m_T)$$
$$u_R + u_S \geq 1$$
$$u_R + u_T \geq 1$$
$$u_S + u_T \geq 1$$

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

Solution: write the primal and the dual LP:

minimize($u_R \log m_R + u_S \log m_S + u_T \log m_T$)	maximize($v_x + v_y + v_z$)
$u_R + u_S \geq 1$	$v_x + v_y \leq \log m_R$
$u_R + u_T \geq 1$	$v_y + v_z \leq \log m_S$
$u_S + u_T \geq 1$	$v_x + v_z \leq \log m_T$

Prove that the AGM Bound is Tight

$Q(x, y, z) = R(x, y), S(y, z), T(z, x)$
$AGM(Q) = \min_{\mathbf{u}} m_R^{u_R} m_S^{u_S} m_T^{u_T}$
where u_R, u_S, u_T range over fractional edge covers.

Solution: write the primal and the dual LP:

minimize$(u_R \log m_R + u_S \log m_S + u_T \log m_T)$	maximize$(v_x + v_y + v_z)$
$u_R + u_S \geq 1$	$v_x + v_y \leq \log m_R$
$u_R + u_T \geq 1$	$v_y + v_z \leq \log m_S$
$u_S + u_T \geq 1$	$v_x + v_z \leq \log m_T$

Define: $R = [2^{v_x^*}] \times [2^{v_y^*}]$, $S = [2^{v_y^*}] \times [2^{v_z^*}]$, $T = [2^{v_z^*}] \times [2^{v_x^*}]$

Claim 1: $|R| \leq m_R$ (why?) Note: if $\neq$ the add arbitrary tuples.

Claim 2: Number of triangles is $AGM(Q)$ (why?).

To discuss in class: u^* is a vertex of the polytope, but v^* is not.

Adding Key Constraints

Assume all cardinalities = m.

$$Q_1(x, y, z) = R(x, y), S(y, z) \qquad |Q| \leq m^2$$

$$Q_2(x, y, z) = R(x, y), S(y, z), T(z, x) \qquad |Q| \leq m^{3/2}$$

Problem 3

Suppose y is a key in S. Give a formula for a tight bound for Q_1 and Q_2.

$$Q_1(x, y, z) = R(x, y), S(\underline{y}, z) \qquad |Q| \leq ?$$

$$Q_2(x, y, z) = R(x, y), S(\underline{y}, z), T(z, x) \qquad |Q| \leq ?$$

Adding Key Constraints

Assume all cardinalities = m.

$$Q_1(x,y,z) = R(x,y), S(y,z) \qquad |Q| \leq m^2$$
$$Q_2(x,y,z) = R(x,y), S(y,z), T(z,x) \qquad |Q| \leq m^{3/2}$$

Problem 3

Suppose y is a key in S. Give a formula for a tight bound for Q_1 and Q_2.

$$Q_1(x,y,z) = R(x,y), S(\underline{y},z) \qquad |Q| \leq ?$$
$$Q_2(x,y,z) = R(x,y), S(\underline{y},z), T(z,x) \qquad |Q| \leq ?$$

Claim: the answers of Q_1, Q_2 have the same sizes as those of Q_1', Q_2':

$$Q_1'(x,y,z) = R'(x,y,z), S(y,z)$$
$$Q_2'(x,y,z) = R'(x,y,z), S(y,z), T(z,x)$$

Their AGM bounds are $AGM(Q_1') = AGM(Q_2') = m$. Let's prove this.

AGM Bound for Relations with Keys

Consider only
$Q(x, y, z) = R(x, y), S(\underline{y}, z), T(z, x)$

Claim 1

Denote: $Q'(x, y, z) = R'(x, y, z), S'(y, z), T(z, x)$
where both R' and S' satisfy the functional dependency $y \rightarrow z$.
Any instance R, S, T can be transfomred into a canoncial instance R', S', T with the same cardinalities. The claim is that $|Q| = |Q'|$ on these instances.

AGM Bound for Relations with Keys

Consider only
$Q(x, y, z) = R(x, y), S(\underline{y}, z), T(z, x)$

Claim 1

Denote: $Q'(x, y, z) = R'(x, y, z), S'(y, z), T(z, x)$
where both R' and S' satisfy the functional dependency $y \rightarrow z$.
Any instance R, S, T can be transfomred into a canoncial instance R', S', T with the same cardinalities. The claim is that $|Q| = |Q'|$ on these instances.

Solution: simply expand each tuple $R(x, y)$ to $R'(x, y, z)$ with the unique value z from $S(y, z)$.

AGM Bound for Relations with Keys

Consider only
$Q(x, y, z) = R(x, y), S(\underline{y}, z), T(z, x)$

Claim 2

Denote $Q''(x, y, z) = R''(x, y, z), S''(y, z), T(z, x)$
where R'', S'' have no constraints.
Claim: Then $\max |Q'| = \max |Q''|$

AGM Bound for Relations with Keys

Consider only
$Q(x, y, z) = R(x, y), S(\underline{y}, z), T(z, x)$

Claim 2

Denote $Q''(x, y, z) = R''(x, y, z), S''(y, z), T(z, x)$
where R'', S'' have no constraints.
Claim: Then $\max |Q'| = \max |Q''|$

Solution: clearly $\max |Q'| \leq \max |Q''|$ because we can simply forget the functional dependencies.
Conversely, consider an instance $R''(x, y, z), S''(y, z), T(z, x)$. Modify the instance as follows: replace everywhere a value y with a pair (y, z). E.g. replace $R''(a, b, c)$ with $R'(a, (b, c), c)$, and replace $S''(b, c)$ with $S'((b, c), c)$. (Possible because every atom that contains y also contains z.) Clearly $Q' = Q''$.

AGM Bound for Relations with Keys: General case

Problem 4

Given a query Q with simple keys, find a tight upper bound formula.

Expand the query Q by repeating the following procedure: if x is a key in the atom $R_j(\mathbf{x}_j)$, then add all the variables $\mathbf{x}_j$ to all other atoms that contain x. Call Q' the modified query (it has no keys and no constraints).

Then $|Q| \leq AGM(Q')$ and this bound is tight.

Notice: upper bounds for non-simple keys, or general FD's are open.

The LeapFrog Trie-Join Algorithm

(time permitting, will discuss in class)