Multi-join Query Evaluation on Big Data Problem Set

Dan Suciu

March, 2015

Name: _____

This problem set contains 7 problems, totalling 100 points. You need 40 points to pass the course. Questions marked `[*]` are considered more challenging. Please email your respose to `mailto://suciu@cs.washington.edu`, no later than April 30, 2015.

Solutions to all problems are available by request, after April 30, 2015.

Question	Points	Score
1	5	
2	15	
3	10	
4	20	
5	20	
6	10	
7	20	
Total:	100	

1. (5 points) Prove that the minimum of the AGM bound:

$$AGM(Q) = \min_{\mathbf{u}} m_1^{u_1} \cdots m_\ell^{u_\ell}$$

is obtained when the fractional edge cover $\mathbf{u}$ is a vertex of the edge-covering polytope.

Hint: for all problems in this set you only need two simple properties of polytopes. (1) if $f(\mathbf{u})$ is a linear function defined on a polytope, then both its minimum and maximum values are obtained when $\mathbf{u}$ is a vertex of the polytope. (2) $\mathbf{u}$ is a vertex of the polytope iff it does not belong to any open segment of the form $\{(1-t)\mathbf{u_0} + t\mathbf{u_1} \mid t \in (0,1)\}$ where $\mathbf{u_0}, \mathbf{u_1}$ belong to the polytope. (In fact, (2) implies (1).)

2. (15points)

Consider the query:

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

Suppose the four relations have cardinalities m_1, m_2, m_3, m_4.

(a) (5 points) Give a formula that represents a tight upper bound on $|Q|$. Your formula should use the cardinalities m_1, m_2, m_3, m_4 and operations like $+, \times, /, \hat{}, \max$, for example $\max(m_1/m_2, m_3^{3/2} + m_4)$ (not a real answer).

(b) (10 points) Repeat your answer for the case when y is a key in S:

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

3. (10 points) Denote $\rho^*(Q)$ and $\tau^*(Q)$ the values of the optimal fractional edge cover, and the optimal fractional edge packing of Q. Give examples of queries Q in each of the cases below:

- $\rho^*(Q) > \tau^*(Q)$
- $\rho^*(Q) = \tau^*(Q)$
- $\rho^*(Q) < \tau^*(Q)$

4. (20points)

Consider the following query:

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

where $|R| = |T| = |L| = 10^{10}$ and $|S| = |K| = 4 \cdot 10^{10}$. We compute this query on a distributed system with p servers, using the HyperCube algorithm. Assume the database is without skew.

(a) (8 points) Compute the load/server when $p = 400$ servers, and when $p = 8000$ servers.

(b) (10 points) Find the speedup of the algorithm as a function of p. Your answer should be of the form *the speedup is* $1/p^{2/3}$ *when* $p < 300$ *and* $1/p$ *(linear) when* $p \geq 300$ (not a real answer).

(c) (2 points) Explain what changes if y is a key in S:

$$Q(x, y, z, u, v, w) = R(x, y), S(\underline{y}, z), T(z, u), K(u, v), L(v, w)$$

5. (20points)[$*$].

Consider the expression:

$$L(\mathbf{u}) = \left(\frac{m_1^{u_1} \cdots m_\ell^{u_\ell}}{p} \right)^{\frac{1}{u_1 + \ldots + u_\ell}}$$

where $m_j \geq p$ for all $j = 1, \ldots, p$.

(a) (10 points) Prove that $L(\mathbf{u})$ is, in general, not a convex function on its entire domain. (Hint: give a counterexample, by choosing concrete values for $\ell, m_1, \ldots, m_\ell$).

(b) (10 points) Prove that $L(\mathbf{u})$ is maximized when $\mathbf{u}$ is a vertex of the edge packing polytope.

6. (10 points) [$*$].

Let R be a bag (multiset) with m elements. Let d_i denote the number of occurrences of the distinct value i in R; thus, $\sum_i d_i = m$. We use a hash function $h : \text{Domain} \to [p]$ from a strongly universal family of hash functions to partition the elements of R into p bins: each element x is sent to bin $h(x)$. Denote L_u the number of elements hashed into bin $u \in [p]$. Prove that, if $\forall i$, $d_i \leq \frac{m}{\alpha \cdot p}$, where $\alpha > 0$ is some fixed constant, then, for all $\delta \geq 0$:

$$\forall u \in [p] : \mathbf{P}(L_u > (1 + \delta)\frac{m}{p}) < \frac{1}{2^{\alpha \cdot \delta}}$$

Derive from here:

$$\mathbf{P}(\max_u L_u > (1 + \delta)\frac{m}{p}) < \frac{p}{2^{\alpha \cdot \delta}}$$

Hint: use Bennett's theorem.

7. (20points)

Let $R(x, y), S(y, z)$ be two relations, and p the number of servers. Let m_R, m_S be their cardinalities, and $m_R[w], m_S[w], w \in \texttt{Domain}_y$ be their y-degree sequences, in other words $m_R[w]$ is the number of tuples in R with $y = w$, and $m_S[w]$ is the number of tuples in S with $y = w$.

(a) (10 points) [$*$] Prove that, any algorithm that computes the join of the two relations R, S on p servers in one round has a load $\Omega(L_{\text{lower}})$, where:

$$L_{\text{lower}} = \left(\frac{\sum_w m_R[w]m_S[w]}{p}\right)^{1/2}$$

Use a simple communication model where messages consists of tuples, as opposed to arbitrary bits; see the proof of the lower bound for the cartesian product in Lecture 3. There is no need to use entropy here.

(b) (10 points) Let $HH = HH_R \cap HH_S$ denote the set of heavy hitters in both R and S, where $HH_R = \{w \mid m_R[w] > m_R/p\}$, $HH_R = \{w \mid m_S[w] > m_S/p\}$. Denote

$$L'_{\text{lower}} = \max\left(m_R/p, m_S/p, \left(\frac{\sum_{w\in HH} m_R[w]m_S[w]}{p}\right)^{1/2}\right)$$

Prove that (a) $L_{\text{lower}} = O(L'_{\text{lower}})$, (b) the converse is false in general (hint: for m arbitrarily large, give a counterexample where $L_{\text{lower}} = 0$, yet $L'_{\text{lower}} = m/p$)