Multi-join Query Evaluation on Big Data
Section 2

Dan Suciu

March, 2015

Lower Bound

$Q(x, y, v, w) = R(x, y), S(y, z), T(z, v), L(v, w)$;
$|R| = |S| = |T| = |L| = m$ tuples.

Let $\mathbf{u} = (u_1, u_2, u_3, u_4)$ be any fractional edge packing.

Problem 1

Prove a lower bound for the load of computing Q on p servers.

Lower Bound

$Q(x, y, v, w) = R(x, y), S(y, z), T(z, v), L(v, w)$;
$|R| = |S| = |T| = |L| = m$ tuples.

We already know $\mathbf{E}[|K_1(\mathrm{msg}_1)|] \leq f_1 m$ tuples, and similarly for K_2, K_3, K_4.

Next step is to apply Friedgut's inequality. Problem: we need an edge cover, but $\mathbf{u}$ is an edge packing.
(In class)

Lower Bound for General Query

$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell)$
$|R_1| = \ldots = |R_\ell| = m.$

Problem 2

Prove a lower bound for arbitrary full conjunctive queries.

Assume that $R_1, \ldots, R_\ell$ are random matchings over a domain of size n: every $R_j \subseteq [n]^{a_j}$, where a_j is the arity of a_j, every attribute is a key, and $|R_j| = n$.

What is $\mathbf{P}(\mathbf{x}_j \in R_j) =$?

What is the entropy of R_j, $H(R_j) =$?

Lower Bound for General Query

$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell)$
$|R_1| = \ldots = |R_\ell| = m.$

Problem 2

Prove a lower bound for arbitrary full conjunctive queries.

Assume that $R_1, \ldots, R_\ell$ are random matchings over a domain of size n: every $R_j \subseteq [n]^{a_j}$, where a_j is the arity of a_j, every attribute is a key, and $|R_j| = n$.

What is $\mathbf{P}(\mathbf{x}_j \in R_j) =?$

What is the entropy of R_j, $H(R_j) =?$

$\mathbf{P}(\mathbf{x}_j \in R_j) = \frac{1}{n^{a_j-1}}$, $H(R_j) = (a_j - 1) \cdot \log n!$.

Lower Bound for General Query

$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell)$

Each server u receives a message $\mathtt{msg}_j(R_j)$ about R_j, of L_j bits. If $L_j \leq f_j H(R_j)$, then $|K_j(\mathtt{m}_j)| \leq f_j n$.

The proof is identical to that for permutations, and we won't prove it.

Lower Bound for General Query

$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell)$

Denote $a_j = \text{arity}(R_j)$ and $a = \sum_j a_j$.

What is $\mathbf{E}[|Q|] =$?

Lower Bound for General Query

$Q(x_1, \dots, x_k) = R_1(\mathbf{x}_1), \dots, R_\ell(\mathbf{x}_\ell)$

Denote $a_j = \text{arity}(R_j)$ and $a = \sum_j a_j$.

What is $\mathbf{E}[|Q|] =$?

$\mathbf{E}[|Q|] = \sum_{\mathbf{x}} \prod_j \frac{1}{n}^{a_j - 1} = n^{k+\ell-a}$

Lower Bound for General Query

$Q(x_1, \dots, x_k) = R_1(\mathbf{x}_1), \dots, R_\ell(\mathbf{x}_\ell)$

Define: $w_{j,\mathbf{x}_j} = \mathbf{P}(\mathbf{x}_j \in K_j(R_j))$. We want an upper bound on:

$$\mathbf{E}[|A_u|] = \sum_{\mathbf{x}} \prod_j w_{j,\mathbf{x}_j}$$

But $\mathbf{u}$ is a fractional edge *packing*; to apply Friedgut's inequality we need a *cover*. What do we do?

Lower Bound for General Query

$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell)$

Define: $w_{j,\mathbf{x}_j} = \mathbf{P}(\mathbf{x}_j \in K_j(R_j))$. We want an upper bound on:

$$\mathbf{E}[|A_u|] = \sum_{\mathbf{x}} \prod_j w_{j,\mathbf{x}_j}$$

But $\mathbf{u}$ is a fractional edge *packing*; to apply Friedgut's inequality we need a *cover*. What do we do?

Add a unary symbol $R'_i(x_i)$ for every variable x_i:
$Q(x_1, \ldots, x_k) = R_1(\mathbf{x}_1), \ldots, R_\ell(\mathbf{x}_\ell), R'_1(x_1), \ldots, R'_k(x_k)$
Transform a *packing* $\mathbf{u}$ into a *cover* $\mathbf{u}'$ by defining: $u'_i = 1 - \sum_{j:i \in R_j} u_j$.
Question: why is $\mathbf{u}'$ an edge cover?

Set $w'_{i,x_i} = 1$

Proof

Use Friedgut; $\sum_i u'_i = \sum_i (1 - \sum_{j:i \in R_j} w_j) = k - \sum_j a_j u_j$; $\mathbf{P}(w_{j,\mathbf{x}_j}) \le 1/n^{a_j - 1}$:

$$\mathbf{E}[|A_u|] = \sum_{\mathbf{x}} \prod_j w_{j,\mathbf{x}_j} \prod_i w'_{i,x_i} = \prod_j (\sum_{\mathbf{x_j}} w_{j,\mathbf{x}_j}^{1/u_j})^{u_j} \prod_i (\sum_{x_i} 1^{1/u'_i}))^{u'_i}$$

$$\le \prod_j (\sum_{\mathbf{x_j}} w_{j,\mathbf{x}_j}^{1/u_j})^{u_j} \prod_i (n)^{u'_i} = n^{k - \sum_j a_j u_j} \cdot \prod_j (\sum_{\mathbf{x_j}} w_{j,\mathbf{x}_j} \cdot w_{j,\mathbf{x}_j}^{1/u_j - 1})^{u_j}$$

$$\le n^{k - \sum_j a_j u_j} \cdot \prod_j \frac{1}{n^{(a_j - 1)(1 - u_j)}} \cdot \prod_j (\sum_{\mathbf{x_j}} w_{j,\mathbf{x}_j})^{u_j}$$

$$= n^{k - \sum_j (a_j u_j + a_j - 1 - a_j u_j + u_j)} \cdot \prod_j (f_j n)^{u_j} = n^{k - a + \ell - u_0} n^{u_0} \prod_j f_j^{u_j}$$

$$= n^{k - a + \ell} \prod_j \left(\frac{f_j M_j}{u_j}\right)^{u_j} \prod_j \left(\frac{u_j}{M_j}\right)^{u_j} \le n^{k - a + \ell} \left(\frac{\sum_j f_j M_j}{u_0}\right)^{u_0} \prod_j \left(\frac{u_j}{M_j}\right)^{u_j}$$

$$\le n^{k - a + \ell} \left(\frac{L}{u_0}\right)^{u_0} \prod_j \left(\frac{u_j}{M_j}\right)^{u_j} = \frac{\prod_j u_j^{u_j}}{u_0^{u_0}} \frac{L^{u_0}}{\prod_j M_j^{u_j}} \mathbf{E}[|Q|] \quad \text{note:} \sum_j f_j M_j = L$$

$$\mathbf{E}[|A|] \le p \mathbf{E}[|A_u|] \le \frac{\prod_j u_j^{u_j}}{u_0^{u_0}} \left(\frac{L}{\left(\frac{\prod_j M_j^{u_j}}{p}\right)^{1/u_0}} \right)^{u_0} \mathbf{E}[|Q|] = O(1) \left(\frac{L}{\frac{M}{p^{1/u_0}}}\right)^{u_0} \mathbf{E}[|Q|] \quad M_1 = \ldots = M_\ell = M$$