CS 784: Foundations of Data Management *Instructor: Paris Koutris*

Lecture 6: Size Bounds for Joins

As we discussed in previous lectures, the output size of a join query often dominates the running time of any algorithm, since any algorithm has to enumerate all the output tuples. Thus, being able to compute the output size, or even provide a good upper bound on the output size becomes an important task. In this lecture, we discuss the following question: given a Conjunctive Query q, where each relation R_j has size N_j, what is the largest possible output?

We start with two examples.

Example 1

Consider the join query $q(x, y, z)$:- $R_1(x, y), R_2(y, z)$ where the sizes of the relations are N_1 and N_2 respectively. The largest possible output is $N_1 \cdot N_2$, which occurs when the join behaves like a cartesian product (i.e. there is a single value of the y variable). One can also observe that we can construct a trivial algorithm that runs in time $N_1 \cdot N_2$ by considering all possible pairs of tuples and checking whether they join or not.

Example 2

Consider the triangle query $\Delta(x, y, z)$:- $R(x, y), S(y, z), T(z, x)$, where the three relations have sizes N_R, N_S, N_T. A first straightforward bound is $N_R \cdot N_S \cdot N_T$. We can get a better bound by noticing that the join of any two relations is an upper bound on the total size, so we get an improved bound of $\min\{N_R \cdot N_S, N_R \cdot N_T, N_T \cdot N_S\}$.

Can we do any better? We will see that another upper bound on the size of the query is $\sqrt{N_R \cdot N_S \cdot N_T}$. Notice that, depending on the relation between N_R, N_S, N_T, this can be a better or worse bound than the above three quantities.

6.1 The AGM Bound

We start by introducing some notation. Let $H(q)$ the hypergraph of a CQ q.

Definition 1: Fractional Edge Cover

The *fractional edge cover* of a hypergraph $H = (V, E)$ is a vector $\mathbf{u}$, which assigns a weight u_j to each hyperedge $e_j \in E$, such that for every vertex $x \in V$, we have that $\sum_{j:x \in e_j} u_j \geq 1$.

We say *fractional* edge cover to distinguish from the (integral) edge cover, which assigns to each hyperedge a weight of 0 or 1. The value of the *minimum fractional edge cover* of the hypergraph $H(q)$ is denoted by $\rho^*(q)$.

Example 3

Consider again the triangle query Δ. A possible fractional edge cover is $u_R = u_S = 1$, and $u_T = 0$. In this case, the sum of the weights is 2. Another fractional edge cover is $u_R = u_S = u_T = 1/2$, which has a smaller sum equal to $3/2$.

The AGM inequality, first proved in [AGM08], bounds the output size of a join query without projections using any fractional edge cover of the query.

Theorem 1: AGM Bound

Let q be a full Conjunctive Query that takes as an input relations S_j with size at most N_j. For every fractional edge cover $\mathbf{u}$ of $H(q)$, the output size is bounded as follows:

$$|q(I)| \leq \prod_{j=1}^{\ell} N_j^{u_j}$$

Notice that in the case we have the same upper bound N on the sizes of the relations, i.e. $N_j = N$, we have that $|q(I)| \leq \min_{\mathbf{u}} N^{\sum_j u_j} = N^{\rho^*(q)}$. In other words, the best bound is achieved by the minimum fractional edge cover $\rho^*(q)$.

Example 4

For the triangle query, the fractional edge cover is $u_R = u_S = u_S = 1/2$ gives the $\sqrt{N_R \cdot N_S \cdot N_T}$ bound. The fractional edge covers $(u_R, u_S, u_T) = (1,1,0), (1,0,1), (0,1,1)$ give the $N_R \cdot N_S$, $N_R \cdot N_T$ and $N_S \cdot N_T$ upper bounds respectively.

Example 5

Consider the Loomis Whitney join LW_k, where each relation has size at most N:

$$LW_k \text{ :- } R_1(x_2, \ldots, x_k), R_2(x_1, x_3, \ldots, x_k), \ldots, R_k(x_1, \ldots, x_{k-1})$$

The smallest fractional edge cover assigns an equal weight of $1/(k-1)$ to each R_j (observe that each variable belongs to exactly $k-1$ atoms). The bound we get then is

$$|LW_k(I)| \leq N^{\sum_{j=1}^{k} u_k} = N^{k/(k-1)}.$$

6.2 Proof of the AGM Bound

We will prove the AGM bound using a tool from information theory called Shannon entropy. Recall that the Shannon entropy of a random variable $\mathbb{X}$ that has N outcomes with probabilities $p_1, \ldots, p_N$ is defined as:

$$H(\mathbb{X}) = -\sum_{i=1}^{N} p_i \log p_i.$$

Let $x_1, \ldots, x_n$ be the variables in the query q. For each variable x_i, we define a *random variable* $\mathbb{X}_i$, such that the random variable $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$ is uniformly distributed over the output tuples in $q(I)$. In other words,

$$Pr[\mathbb{X} = t] = \begin{cases} 1/|q(I)|, & t \in q(I) \\ 0, & otherwise \end{cases}$$

Since $\mathbb{X}$ is defined as a uniform distribution over the output, we have $H(\mathbb{X}) = \log |q(I)|$. Moreover, for each hyperedge $e_j \in E$, the marginal distribution of $\mathbb{X}$ on e_j (denoted $\mathbb{Y}_j$) has support at most N_j. Hence, we also have that $H(\mathbb{Y}_j) \leq \log N_j$.

We now apply a powerful tool called *Shearer's lemma*. This tells us that for every fractional edge cover $\mathbf{u}$ of a hypergraph, we have:

$$H(\mathbb{X}) \leq \sum_j u_j H(\mathbb{Y}_j)$$

Applying Shearer's lemma, we have:

$$\log |q(I)| = H(\mathbb{X}) \leq \sum_j u_j \cdot H(\mathbb{Y}_j) \leq \sum_j u_j \log N_j.$$

6.3 Tightness of the AGM Bound

The AGM bound gives us an infinite number of upper bounds on the output size. Given the cardinalities of each relation, how can we find the best (minimum) possible bound? In the case of equal cardinalities N it suffices to find $\rho^*(q)$, but in the general case we can achieve this by minimizing the quantity $\prod_{j=1}^{\ell} N_j^{u_j}$ by solving the following linear program (LP):

$$\begin{aligned} \min \quad & \sum_j \log_2(N_j) \cdot u_j \\ \text{s.t.} \quad & \forall x_i \in V : \sum_{j: x_i \in e_j} u_j \geq 1 \\ & \forall e_j \in E : u_j \geq 0 \end{aligned}$$

The fractional edge cover obtained by the above LP will give the best possible bound. It turns out that this bound is *tight*; in other words, we can always find a database instance I, such that $|q(I)|$ is equal to the worst-case upper bound. The idea is simple: we first take the dual of the LP.

$$\begin{aligned} \max \quad & \sum_i w_i \\ \text{s.t.} \forall e_j \in E : & \sum_{i : x_i \in e_j} w_i \leq \log_2(N_j) \\ & \forall x_i \in V : w_i \geq 0 \end{aligned}$$

Let $\mathbf{w} = (w_1, w_2, \dots)$ be the optimal solution of the above dual LP. We create an instance I as follows. For every variable x_i, we assign a domain of size 2^{w_i}. Each relation is then created by taking the cartesian product of the domains of its variables. One can verify that each relation R_j in the instance has size at most $2^{\sum_{i:x_i \in e_j} w_i} \leq N_j$. Moreover, the output will be of size $2^{\sum_i w_i}$, which is exactly equal to the AGM bound by LP duality.

References

[Alice] S. ABITEBOUL, R. HULL and V. VIANU, "Foundations of Databases."

[AGM08] A. ATSERIAS, M. GROHE and D. MARX, "Size bounds and query plans for relational joins," *FOCS 2008*.