Homework 3

Due on December 13

A: Parallel Query Processing **[25%]**

The task is to compute the join query $Q(x_1, x_2, x_3) : -R_1(x_1), R_2(x_2), R_3(x_3)$ in parallel using p machines. The query is a cartesian product of three unary relations. Let the sizes of the relations be N_1, N_2, N_3 respectively.

1. **[10%]** Suppose that we compute Q in a single round by distributing the data once (using the HyperCube algorithm). What is the smallest possible load (maximum amount of data each machine receives)? What is the total communication?

2. **[10%]** Suppose that we compute Q in two rounds. In the first round we compute the cartesian product of R_1, R_2, and the second round the cartesian product of the intermediate result with R_3. What is the load in each round? What is the total communication across both rounds?

3. **[5%]** Which of the above two strategies achieves the best load? Which one the best total communication?

B: Data Streaming **[30%]**

1. **[15%]** In *reservoir sampling*, we want to produce a uniform sample of size k from a stream of unknown size. The algorithm works by placing the i-th item in the reservoir (of size k) with probability k/i (all the first k items go in the reservoir initially). Show that at any point where we have seen n total items, the probability of each item being in the reservoir is k/n.

2. **[15%]** Recall the application of the Misra-Gries algorithm for the case where we want to find the top-k most frequent elements. For any element j, let f_j be its actual frequency in the stream, and $\hat{f}_j$ its estimated frequency.

   Show that $f_j - \frac{m-\hat{m}}{k} \leq \hat{f}_j \leq f_j$, where $\hat{m}$ is the sum of the estimated frequencies $\hat{f}_j$, and m is the total length of the stream.

C: UNCERTAIN DATA **[30%]**

1. **[15%]** Consider a tuple-independent probabilistic database with the following relations: $R(A, B), S(A, C), T(A, D)$. Suppose we want to answer the boolean query

$$Q() : -R(x, y), S(x, 'a'), T(x, w)$$

   Describe a *safe plan* for computing the probability of Q if one exists; otherwise, explain why it is not possible to obtain one.

2. **[15%]** Consider an inconsistent database, where the integrity constraints are primary keys. The database consists of two relations $R(\underline{A}, B)$ and $S(\underline{B}, C, D)$. We want to obtain the consistent answers for the query $Q(x) : -R(x, y), S(y, z, 'a')$. Write a query in SQL that computes the consistent answers for Q.

D: PROVENANCE **[15%]**

Consider the following instance with two relations $R(A), S(A, B)$, where each tuple is tagged with a unique identifier.

$$\{R(a_1) : x_1, R(a_2) : x_2, S(a_1, b_1) : y_1, S(a_1, b_2) : y_2, S(a_2, b_2) : y_3\}$$

For each polynomial below, write a Boolean CQ (without inequalities or constants) having that provenance polynomial.

1. $x_1y_1 + x_1y_2 + x_2y_3$
2. $x_1y_1^2 + x_1y_2^2 + x_1y_2y_3 + x_2y_2y_3 + x_2y_3^2$
3. $(x_1 + x_2)(y_1 + y_2 + y_3)$

DELIVERABLES

Submit a single PDF document using `Canvas` (Homework 3). It is strongly suggested to use LaTeX to write your solution.