Homework 3
Due on May 5

A: Views and Rewriting [15%]

Find an equivalent rewriting of the query

\[ Q() : -R(X_1, X_2), R(X_2, X_3), R(X_3, X_4), R(X_4, X_5), R(X_5, X_1), X_1 < X_2 \]

using the following two views:

- \( V_1(X_1, X_3) : -R(X_1, X_2), R(X_2, X_3) \)
- \( V_2(X_1, X_3) : -R(X_3, X_4), R(X_4, X_5), R(X_5, X_1) \)

B: Parallel Query Processing [25%]

The task is to compute the 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 technique). What is the minimum amount of data that each machine can receive? 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 minimum amount of data that each machine can receive in each round? What is the total communication across both rounds?

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

C: 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. We showed in class that \( f_j - \frac{m}{k} \leq \hat{f}_j \leq f_j \), where \( m \) is the total length of the stream. Show that this bound can be improved to \( f_j - \frac{m-\hat{m}}{k} \leq \hat{f}_j \leq f_j \), where \( \hat{m} \) is the sum of the estimated frequencies \( \hat{f}_j \).

**D: 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

\[
Q() : \neg R(x, y), S(x, 'a'), T(x, w)
\]

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 \( R(A, B) \) and \( S(B, C, D) \). We want to obtain the consistent answers for the query \( Q(x) : \neg R(x, y), S(y, z, 'a') \). Write a query in SQL that computes the consistent answers for \( Q \).

**Deliverables**

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