

Assignment 2: Query Optimization

Due Thu March 9
at the beginning of class

No late assignments will be accepted.

1. (16 points) *Algebraic Query Optimization.*

Here is a schema for a banking database:

```
account[number, person, balance, bank, city, branch]
bank[bank, city, branch, street, manager]
lives[person, city, street, number]
```

Here is a view definition:

```
create view summary
as select lives.person, lives.city, lives.street, account.number,
        balance, account.bank, bank.city, bank.street
from account, bank, lives
where account.person = lives.person
and account.bank = bank.bank
and account.city = bank.city
and account.branch = bank.branch
```

- (a) Using the view, translate the following query into SQL: “Retrieve the person, bank city, and account number of all accounts at any branch of Royal Bank with a balance of over \$60,000 belonging to people who live in Toronto.” (1 point)
- (b) Translate the view and the query into relational algebra. (2 points)
- (c) Draw parse trees for the query, the view, and the query merged with the view. (3 points)
- (d) Produce an optimal program for evaluating the query, as shown in class. Show the parse tree after pushing selects, after pushing projections, after eliminating redundant operators, and after grouping the operators. Finally, show the assignment statements. (10 points)

2. (20 points) *Distributed Query Optimization.*

A distributed database has four relations $P[AB]$, $Q[BC]$, $R[CD]$, and $S[DE]$ distributed over three sites. P is stored at site 1, Q is stored at site 2, and R and S are stored at site 3. The relations have the following statistics:

$$\begin{array}{lll}
 |P| = 7,000 & |\pi_A P| = 4,200 & |\pi_B P| = 5,000 \\
 |Q| = 8,000 & |\pi_B Q| = 6,000 & |\pi_C Q| = 5,000 \\
 |R| = 5,000 & |\pi_C R| = 4,000 & |\pi_D R| = 3,500 \\
 |S| = 6,000 & |\pi_D S| = 4,000 & |\pi_E S| = 4,500
 \end{array}$$

In addition, the cost of opening a communication channel is equal to the cost of transmitting 1,200 tuples (i.e., $c_0 = 1,200$). We would like to join the four relations by splitting the 3-way join into pairwise joins as follows:

$$(P \bowtie (Q \bowtie R)) \bowtie S$$

- (a) (1 point) Draw the unordered binary tree for this partitioning of the join.
 (b) (8 points) As shown in class, determine the minimum cost of evaluating the root at site 3. That is, create a table with the following headers:

node	site of node	site of left child	site of right child	method	cost

Fill in the cost of each method, as shown in class, and then produce a new table retaining only the cheapest method for each node and site. For simplicity, you may ignore semi-join methods.

- (c) (8 points) Show your estimates of all image sizes and join sizes used to construct the table.
 (d) (3 points) From the table, write an optimal program for evaluating the root of the tree at site 3, as shown in class. The optimal program will be a sequence of statements like the following:

Ship Q from site 2 to site 3.
 Compute $Q \bowtie R$ at site 3.
 Ship $Q \bowtie R$ from site 3 to site 1.