CSE 544 Homework 3
April 2006

Due date: Wednesday, May 3, 2006.

1. (12 points) This is the famous drinkers-beers-bars problem, used by Ullman in his early textbook on databases. Consider the following schema:

   `Likes(drinker, beer), Frequents(drinker, bar), Serves(bar, beer)`

   We will abbreviate the table names with L, F, S. For example the following query finds all drinkers that like only `Bud-Light`:

   $$q(d) \quad :- \quad (\exists b.L(d,b)) \wedge (\forall b.L(d,b) \Rightarrow b = \texttt{Bud-Light})$$

   Note that the first condition ensures that the query is safe.

   Write FO formulas to compute the following:

   (a) Find all drinkers that frequent only bars that serve only beer they like. (Optimists)

   (b) Find all drinkers that frequent only bars that serve some beer they like. (Realists)

   (c) Find all drinkers that frequent some bar that serves only beers they like. (Prudents)

   (d) Find all drinkers that frequent only bars that serve none of the beers they like. (Flagellators)

2. (12 points) A formula $\varphi(x_1, \ldots, x_m)$ is *range-restricted* if (a) it is of the form:

$$\varphi(x_1, \ldots, x_m) \equiv adom(x_1) \wedge adom(x_2) \wedge \ldots \wedge adom(x_m) \wedge \psi(x_1, \ldots, x_m)$$

(b) every quantifier in ψ has one of these two forms: $\forall z.(adom(z) \Rightarrow \omega)$ or $\exists z.(adom(z) \wedge \omega)$, where ω is some other formula. Here $adom(u)$ is a formula that checks if u is in the active domain. Examples of range restricted formulas over the vocabulary $R(x, y)$ are:

$$\begin{aligned} \varphi_1(x) &\equiv adom(x) \wedge (\forall y.adom(y) \rightarrow R(x, y)) \\ \varphi_2(x, y) &\equiv adom(x) \wedge adom(y) \wedge (R(x, x) \vee \exists z.(adom(z) \wedge \neg R(y, z))) \end{aligned}$$

where $adom(u) \equiv (\exists v.R(u, v)) \vee (\exists v.(R(v, u)))$. Indicate for each of the statements below if it is true of false. You don't have to justify your answer:

(a) Every range restricted formula is safe (i.e. domain independent: that is, its answer depends only on the extent of the relations and not on the domain).

(b) Every range restricted formula is finite (i.e. on any structure that has finite relations but possible infinite domain it returns a finite set of answers).

(c) Every safe formula is range restricted.

(d) For every safe formula φ there is some range restricted formula φ' s.t. φ and φ' are equivalent, $\varphi \equiv \varphi'$.

(e) The set of range restricted formulas is decidable.

(f) The set of safe formulas is decidable.

3. (21 points) Consider three finite relations: $R(x, y), S(x), U(x, y)$.

(a) Write a formula $adom(x)$ that computes the active domain of a database with the schema R, S, U.

(b) For each of the FO queries below do the following: (1) indicate whether they are finite or not, (2) indicate whether they are safe or not, (3) give a range restricted formula that is equivalent, or indicate that no such formula exists.

i. $\{x \mid S(x) \wedge \forall y.(\neg R(x,y))\}$

ii. $\{x \mid S(x) \wedge (\forall y.(R(x,y) \Rightarrow \exists z.(S(z) \vee U(y,z))))\}$

iii. $\{x \mid \exists y.(S(y) \Rightarrow \forall z.(R(x,y) \wedge U(y,z)))\}$

iv. $\{x \mid S(x) \wedge \forall y.(S(y) \wedge R(x,y))\}$

v. $\{x \mid S(x) \wedge \forall y.(U(x,y) \vee \forall z.(\neg R(y,z)))\}$

vi. $\{(x,y) \mid \exists z.(R(x,z) \vee U(z,y))\}$

4. (18 points) Let $T(x,y,z)$ and $L(x)$ be two tables representing a binary tree: a triple (x,y,z) in T says that x is the parent of y and z, while a node x in L indicates that x is a leaf.

   (a) Two nodes u, v are on the same level in the tree if either u and v have the same parent, or their parents are on the same level. Write a datalog query that returns all nodes that are on the same level as given node a (here a is a constant).

   (b) Alice and Bob play the following pebble game on the tree T. Alice places the pebble on some node x. Next, Bob moves the pebble to one of the children of x, call it x_1. Next, Alice moves the pebble to one of the children of x_1 call it x_2. The game continues until the pebble reaches a leaf, x. If $A(x)$ is true then Alice wins, otherwise Bob wins. Here we assume that $A(x)$ is a predicate that is true at a leaf x if Alice wins at x. Write a datalog query that computes the set of all nodes x where Alice can start the game and have a winning strategy.

5. (22 points) Query containment.

   (a) Indicate for each pair of queries q, q' below, whether $q \subseteq q'$. If the answer is yes, provide a proof; if the answer is no, give a database instance I on which $q(I) \not\subseteq q'(I)$.

   i.

$$\begin{aligned} q(x) &:- \quad R(x,y), R(y,z), R(z,x) \\ q'(x) &:- \quad R(x,y), R(y,z), R(z,u), R(u,v), R(v,z) \end{aligned}$$

   ii.

$$\begin{aligned} q(x,y) &:- \quad R(x,u,u), R(u,v,w), R(w,w,y) \\ q'(x,y) &:- \quad R(x,u,v), R(v,v,v), R(v,w,y) \end{aligned}$$

iii.

$$\begin{aligned} q() &:- \quad R(u,u,x,y), R(x,y,v,w), v \neq w \\ q'() &:- \quad R(u,u,x,y), x \neq y \end{aligned}$$

iv.

$$\begin{aligned} q(x) &:- \quad R(x,y), R(y,z), R(z,v) \\ q'(x) &:- \quad R(x,y), R(y,z), y \neq z \end{aligned}$$

(b) Let:

$$\begin{aligned} q_1(x) &:- \quad R(x,y), R(y,z), R(z,u) \\ q_2(x) &:- \quad R(x,y), R(y,z) \end{aligned}$$

Notice that $q_1 \subseteq q_2$. Give an example of a conjunctive query q such that $q_1 \subset q$ and $q \subset q_2$. Here $q_1 \subset q$ means $q_1 \subseteq q$ and not $q \subseteq q_1$.

(c) Consider the following two queries:

$$\begin{aligned} q_1(x) &:- \quad R(x,y), R(y,z), R(a,z) \\ q_2(x) &:- \quad R(x,y), R(y,z), R(z,u), R(y,b) \end{aligned}$$

Here a and b are constants, while x, y, z, u are variables. Find two queries q and q' such that the following four conditions hold simultaneously: $q \subseteq q_1$, $q \subseteq q_2$, $q_1 \subseteq q'$, $q_2 \subseteq q'$. You should choose q and q' as "tight" as possible.

6. (15 points) For each statement below indicate whether it is true or false. You do not have to provide any proof. (Note: some answers below are trivial, but one statements has a difficult proof. You don't have to prove it, or find the proof in the literature: instead rely on your intuition to provide a true/false answer).

   (a) Every query in FO has a data complexity which is in PTIME
   (b) All queries in FO are monotone.
   (c) The query complexity of conjunctive queries is NP complete.
   (d) There exists a query in FO that is not expressible in datalog.
   (e) If a query can be expressed in FO and also in datalog, then it can be expressed in UCQ (= unions of conjunctive queries).