CSE-6490B: Assignment #2

1. **Query Containment.** *I can't contain myself.* (5 points)

   Consider the following conjunctive queries, Ullman-style. Note that 'c' in $\mathcal{Q}_4$ is a constant, not as variable.

$$\begin{aligned}
\mathcal{Q}_1&: p\ (X,\ Y) \leftarrow q\ (X,\ A),\ q\ (A,\ B),\ q\ (B,\ Y).\\
\mathcal{Q}_2&: p\ (X,\ Y) \leftarrow q\ (X,\ A),\ q\ (A,\ B),\ q\ (B,\ C),\ q\ (C,\ Y).\\
\mathcal{Q}_3&: p\ (X,\ Y) \leftarrow q\ (X,\ A),\ q\ (B,\ C),\ q\ (D,\ Y),\\
&\qquad\qquad\quad q\ (X,\ B),\ q\ (A,\ C),\ q\ (C,\ Y).\\
\mathcal{Q}_4&: p\ (X,\ Y) \leftarrow q\ (X,\ A),\ q\ (A,\ c),\ q\ (c,\ B),\ q\ (B,\ Y).
\end{aligned}$$

---

   a. (2 points) Find all containments and equivalences between $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, and $\mathcal{Q}_4$.

---

   b. (2 points) For each of $\mathcal{Q}_1$, $\mathcal{Q}_2$, $\mathcal{Q}_3$, and $\mathcal{Q}_4$, simplify it. This means find the minimal clause that is equivalent to $\mathcal{Q}'_i$, in each case.

   Simplify $\mathcal{Q}'_1 \cup \mathcal{Q}'_2 \cup \mathcal{Q}'_3 \cup \mathcal{Q}'_4$ (where $\mathcal{Q}'_i$ is your simplified $\mathcal{Q}_i$). This means eliminate any of the rules contained in any other, because these do not contribute anything additionally to $p\ (X,\ Y)$.

---

   c. (1 point) A containment mapping is sufficient and necessary to show containment for Datalog conjunctive queries without inequalities. It is still necessary but not sufficient to show containment for Datalog conjunctive queries with inequalities.

   What do you need to additionally show in these cases to prove containment?

2. **Integrity Constraints.** *You can't say that!* (5 points)

Most schema also have integrity constraints. We can extend our Datalog databases to include integrity constraints (ICs), and our notion of containment to account for ICs.

An integrity constraint can be written as a query, with the mandate that the IC "query" must evaluate to have *no* answers. A common convention is to use '$\Leftarrow$' instead of '$\leftarrow$' when writing an IC instead of a query, to distinguish ICs and queries.

Consider the following schema.

student(s#, sname, dob, d#)
  FK (d#) refs **dept** // Student's major
prof(p#, pname, d#)
  FK (d#) refs **dept** // Professor's home deparmtent
dept(d#, dname, building, p#)
  FK (p#) refs **prof** // Department's chair
course(d#, no, title)
  FK (d#) refs **dept** // Course offered by this deparmtent
class(d#, no, term, year, section, room, time, p#)
  FK (d#, no) refs **course** // Class is an offering of this course
  FK (p#) refs **prof** // Instructor of class
enroll(s#, d#, no, term, year, section, grade)
  FK (s#) refs **student** // This student is enrolled in
  FK (d#, no, term, year, section) refs **class** // this class

---

a. (2 points) Write an IC to represent the constraint that s# is the primary key of **student**; that is, a s# value can only appear once.

Write an IC to represent the constraint that d# is a foreign key of **student** referencing **dept**; that is, any d# value in **student** must be also a value in **dept**.

---

b. (2 points) Consider

$$\Leftarrow e\ (A,\ C),\ e\ (B,\ C),\ A \neq B.$$
$$d\ (A,\ C) \leftarrow e\ (A,\ B),\ e\ (B,\ C).$$
$$t\ (A,\ D) \leftarrow e\ (A,\ B),\ e\ (B,\ C),\ e\ (C,\ D).$$

Given d and t as resources, and assuming the IC for both, find the maximal contained foldings for

$$\leftarrow e\ (X,\ Y).$$

---

c. (1 point) Consider

$$\Leftarrow e\ (A,\ B),\ e\ (B,\ A).$$
$$tri\ (A,\ B,\ C) \leftarrow e\ (A,\ B),\ e\ (B,\ C),\ e\ (C,\ A).$$
$$nontrans\ (A,\ B,\ C) \leftarrow e\ (A,\ B),\ e\ (B,\ C),\ \mathsf{not}\ e\ (A,\ C).$$

Is *tri* contained by *nontrans*, given the IC?
Is *nontrans* contained by *tri*, given the IC?

3. **Information integration.**[1] *Schema scheming.* [exercise] (5 points)

   We are trying to integrate a set of booksellers. The schema we want to export to the user involves just a single relation:

   Book(author, title, subject, price, #pages)

   Assume we have the following two online bookstore sources.

   A) *AliensRComing*: Only sells sci-fi books. For each book it sells, it can give the information about author, title, price, and the number of pages.

   E) *EndIsNear*: Primarily sells religious books. However, it also stocks a few books on other subjects. For each book it sells, it has information about author, title, subject, and price.

   a. Consider integrating these sources using the global-as-view approach. Write down the view corresponding to the **Book** relation.

   b. Consider the following SQL query:

   ```
   select author
       from Book
       where subject = "sci-fi"
   ```

   Show how it gets rewritten in terms of the sources in the global-as-view approach.

   c. Now consider instead integrating these sources using the local-as-view approach. Show how the two sources will be modeled.

   d. Consider your local-as-view approach from Question 3c and the following candidate plan for the query from Question 3b:

   > Compute the results by calling the source **A** and selecting the authors from the returned tuples.

   Show, using the idea of containment mappings, that this plan is in fact a sound plan for the query. Is it also a complete plan? If not, what additional information about the sources would make it a complete plan?

   e. Consider your local-as-view approach from Question 3c. Show how the inverse-rule algorithm rewrites the query from Question 3b into a maximally contained plan.

[1] Thanks to Subbarao Kambhampati.

4. **XPath.** *Oops, I think we're on the Y path by mistake.* (5 points)

   Compose the following XPath queries with respect to *Bibliography* XML document (on the course webpage with this assignment). Look over the Bibliography document to understand its "schema".

   Your queries ought to be logically correct in that they would still do the job as expected if the Bibliography document were to have many new papers added.

   a. Extract the titles of the paper.
   b. Return the titles of the papers by Halevy.
   c. Extract the authors of papers that appeared in a session named "Invited Talk".
   d. Extract the titles of papers that have four or more authors.
   e. Return the names of the conferences *and* journals.
   f. Extract the titles of papers such that "XML" is mentioned *somewhere* in the paper's ancestors's attributes or in the descendents's attributes or text.
   g. Return the titles of papers co-authored by Hass and Kossmann.
   h. Return the titles of papers that appeared after the year 2000.
   i. Return the titles of papers that are longer than 20 pages.
   j. Return the names of conferences that contain a paper with "XML" in its title and 'Chaberlin' as one of its authors.