ECS289F — Homework 3 Solutions

March 3, 2010

Problem 1. *Prove that checking equivalence of CQs is NP-complete, by modifying the proof from class of the same for containment.*

Solution. Since checking containment is in NP and $Q \equiv Q'$ iff $Q \sqsubseteq Q'$ and $Q' \sqsubseteq Q$, it is clear that equivalence is also in NP. To show NP-hardness, here are two ways to do it:

1. Reduce containment to equivalence, by observing that $Q \sqsubseteq Q'$ iff $Q \cap Q' \equiv Q$, and noting that CQs are closed under intersection. Specifically, given CQs $Q = \langle \bar{u}, T \rangle$ and $Q' = \langle \bar{u}', T' \rangle$ with $\mathsf{vars}(T)$ and $\mathsf{vars}(T')$ disjoint and $|\bar{u}| = |\bar{u}'| = k$, their intersection is the CQ
$$Q \cap Q' \stackrel{\text{def}}{=} \langle \bar{u}, T \cup T' \cup \{u_i = u'_i \mid 1 \leq i \leq k\} \rangle$$
2. Modify the reduction from 3-coloring used to show NP-hardness of the recognition (and containment) problem for CQs. Recall that a graph $G(V, E)$ is 3-colorable iff there exists a graph homomorphism $h : G \to C_3$, where C_3 is the complete 3-vertex graph. Let G' be the disjoint union of G and C_3 (i.e., G' has a copy of C_3 and a copy of G, with disjoint vertex sets). We observe, using similar reasoning as in the earlier reduction, that the CQs corresponding to G' and C_3 are equivalent iff there exist graph homomorphisms $f : G' \to C_3$ and $g : C_3 \to G'$. But it is easy to see that G is 3-colorable iff G' is 3-colorable, hence we have a graph homomorphism $f : G' \to C_3$ iff we have one from G to C_3, and in the other direction, the identity mapping is a graph homomorpism $g : C_3 \to G'$.

(Note that the two proofs are essentially the same since the disjoint union of G and C_3 used in the second proof corresponds exactly to the graph of the intersection of their corresponding CQs.) $\square$

Problem 2. *A* union of conjunctive queries *(UCQ)* $\mathbf{Q}$ *is a set* $\mathbf{Q} = \{Q_1, \ldots, Q_n\}$ *of conjunctive queries, with the semantics defined*
$$[\![Q]\!]^I \stackrel{def}{=} \bigcup_{i=1}^{n} [\![Q_i]\!]^I.$$

(a) Show that the problems of containment and equivalence for UCQs are easily interreducible.

(b) Show that the Chandra-Merlin Theorem for containment of conjunctive queries extends to UCQs (you will need to restate it appropriately), and that the complexity of the problem remains the same, i.e., NP-complete.

(c) Give a procedure for minimizing UCQs and show that, as for CQs, the result of this procedure is unique up to isomorphism, where UCQs $\mathbf{Q}$ *and* $\mathbf{Q}'$ *are* isomorphic *if there is a bijection* $h : \mathbf{Q} \to \mathbf{Q}'$ *such that for every CQ* $Q \in \mathbf{Q}$*,* Q *is isomorphic to* $h(Q)$*.*

Solution.

(a) We have $\mathbf{Q} \sqsubseteq \mathbf{Q}'$ iff $\mathbf{Q} \cup \mathbf{Q}' \equiv \mathbf{Q}'$ (and of course $\mathbf{Q} \equiv \mathbf{Q}'$ iff $\mathbf{Q} \sqsubseteq \mathbf{Q}'$ and $\mathbf{Q}' \sqsubseteq \mathbf{Q}$).

(b) The extended version of the theorem (due to Sagiv and Yannakakis, 1980) can be stated as follows.

Theorem 3. *For UCQs $\mathbf{Q}, \mathbf{Q}'$ the following are equivalent:*

1. $\mathbf{Q} \sqsubseteq \mathbf{Q}'$

2. For every $Q_i \in \mathbf{Q}$ there exists a containment mapping $h : Q'_j \to Q_i$ for some $Q'_j \in \mathbf{Q}'$

3. For every $Q_i = \langle \bar{u}_i, T \rangle \in \mathbf{Q}$ we have $\bar{u}_i \in [\![\mathbf{Q}']\!]^{can(Q_i)}$

Proof. To prove this, it suffices to show that $\mathbf{Q} \sqsubseteq \mathbf{Q}'$ iff for every $Q_i \in \mathbf{Q}$ there exists $Q'_j \in \mathbf{Q}'$ such that $Q_i \sqsubseteq Q'_j$. (The equivalence of 1–3 then follows by Chandra-Merlin.) "$\Leftarrow$" is obvious. For "$\Rightarrow$," suppose $\mathbf{Q} \sqsubseteq \mathbf{Q}'$, and choose arbitrarily some $Q_i = \langle \bar{u}_i, T_i \rangle \in \mathbf{Q}$. Clearly $\bar{u}_i \in [\![\mathbf{Q}]\!]^{\mathsf{can}(Q_i)}$, hence $\bar{u}_i \in [\![\mathbf{Q}']\!]^{\mathsf{can}(Q_i)}$ and in particular $\bar{u}_i \in [\![Q'_j]\!]^{\mathsf{can}(Q_i)}$ for some $Q'_j \in \mathbf{Q}'$. Using Chandra-Merlin, we conclude $Q_i \sqsubseteq Q'_j$, as required. □

Given the theorem above, it is obvious that containment of UCQs is in NP (since we can just guess the containment mappings). Moreover, for CQs Q, Q' we have $Q \sqsubseteq Q'$ iff $\{Q\} \sqsubseteq \{Q'\}$, hence it is NP-hard as well.

(c) The procedure to minimize a UCQ $\mathbf{Q}$ is as follows: let $Q_1, \ldots, Q_k$ be an arbitrary ordering of the CQs in $\mathbf{Q}$; throw out any Q_i such that $Q_i \sqsubseteq Q_j$ for some $j > i$; then minimize each CQ in the result. It is clear that the result of this process will be unique up to isomorphism. Moreover, one can check that it will be minimal according to the following equivalent criteria:

- (global minimality) A UCQ $\mathbf{Q}$ is said to be *globally minimal* if for any equivalent UCQ $\mathbf{Q}'$, the total number of atoms in all CQs in $\mathbf{Q}$ is less than or equal to the same for $\mathbf{Q}'$.
- (local minimality) Define a *subquery* of a UCQ $\mathbf{Q}$ to be a UCQ obtained from $\mathbf{Q}$ as follows: choose some subset $\mathbf{Q}' \subseteq \mathbf{Q}$ (not necessarily a strict subset), then replace each CQ Q in $\mathbf{Q}'$ with a subquery of Q. Then $\mathbf{Q}$ is said to be *locally minimal* if any subquery $\mathbf{Q}'$ of $\mathbf{Q}$ that is equivalent to $\mathbf{Q}$ is identical to $\mathbf{Q}$.

Finally, using similar reasoning as for CQs, one can check that for minimal UCQs, equivalence implies isomorphism.

□

Problem 4. *Denote by SPCU ("select-project-cross product-union") the fragment of the relational algebra where set difference is disallowed.*

(a) Show that any SPCU query can be rewritten equivalently as a union of SPC queries and that this implies that SPCUs are expressively equivalent to UCQs.

(b) Give an example of an SPCU expression such that the corresponding union of SPC expressions (and hence the corresponding UCQ) is exponentially larger.

(c) *Recall that Π_2^p is the complement of Σ_2^p, the class of languages decidable in nondeterministic polynomial time with access to an oracle for NP. The canonical example of a Π_2^p-complete problem is Q3-SAT ("quantified 3-SAT"): given a Boolean formula $F = B_1 \wedge \cdots \wedge B_n$ in conjunctive normal form with three literals in each clause (3-CNF), and a subset X of the variables in F, decide if F is satisfiable for every assignment of the variables in X.*

Show that containment of SPCU queries is in Π_2^p.

(d) **Extra credit.** *Show that containment of SPCU queries is Π_2^p-hard.*

Solution.

(a) We can accomplish this rewriting by repeatedly applying the following transformations:

$$\begin{aligned}
E_1 \times (E_2 \cup E_3) &= (E_1 \times E_2) \cup (E_1 \times E_3) \\
(E_1 \cup E_2) \times E_3 &= (E_1 \times E_3) \cup (E_2 \times E_3) \\
\sigma_{i=j}(E_1 \cup E_2) &= \sigma_{i=j}(E_1) \cup \sigma_{i=j}(E_2) \\
\sigma_{i=c}(E_1 \cup E_2) &= \sigma_{i=c}(E_1) \cup \sigma_{i=c}(E_2) \\
\pi_{i_1,\ldots,i_k}(E_1 \cup E_2) &= \pi_{i_1,\ldots,i_k}(E_1) \cup \pi_{i_1,\ldots,i_k}(E_2)
\end{aligned}$$

Since we know that any SPC expression can be translated into an equivalent CQ, it follows that any SPCU expression can be translated into an equivalent UCQ. Moreover, any UCQ can be translated into an equivalent SPCU expression using the SPC-to-CQ translation as a subroutine.

(b) The SPCU expression

$$(R \cup S) \times \cdots \times (R \cup S)$$

with n union operators, $n-1$ cross product operators, and $2n$ predicate symbols contains $2^n - 1$ union operators, $(n-1)2^n$ cross product operators, and $n \cdot 2^n$ predicate symbols in normal form:

$$(R \times \cdots \times R) \cup (S \times R \times \cdots \times R) \cup (R \times S \times \cdots \times R) \cup \cdots \cup (S \times \cdots \times S)$$

(c) Let E_1, E_2 be SPCU queries. Note that if $E_1 \not\sqsubseteq E_2$, then we have some CQ $Q = \langle \bar{u}, T \rangle$ in the UCQ $\mathbf{Q}$ corresponding to E_1 such that $\bar{u} \notin [\![E_2]\!]^T$. Now, note that we can view each CQ Q in $\mathbf{Q}$ as being produced from E_1 by the following procedure:

- for each (binary) union operation in E_1, pick "left" or "right" and erase the subexpression on that side (along with the union operator itself);
- the result is an SPC query, which can be translated (in polynomial time) into an equivalent CQ.

Thus, the witness we use is a set of labels "left" or "right," one for each union operator in E_1. From such a labeling, the corresponding CQ can be produced in polynomial time, and non-containment in E_2 can be verified in NP. Thus the problem is in Π_2^p.

(d) See the proof of Theorem 19 in the paper by Sagiv and Yannakakis (*J. ACM*, 1980).

$\square$