4. Containment and Minimization of Conjunctive Queries

These are sample solutions to some of the exercises that were given as coursework. They are not intended as models but show each one way to approach the problem set in the exercise.

2. Containment of Selfjoin-free Conjunctive Queries with Built-Ins

We start with a characterization of containment of selfjoin-free queries with built-ins over rational numbers. In contrast with containment of general conjunctive queries with built-ins, such a characterization is possible when selfjoin-free queries are considered.

Theorem 1. *Let $Q(\vec{x}){:-}L, M$ and $Q'(\vec{x}){:-}L', M'$ be selfjoin-free conjunctive queries. Then the following are equivalent:*

- $Q \sqsubseteq Q'$
- *Q is unsatisfiable or there exists a homomorphism δ from Q' to Q.*

Recall that for queries with comparisons, a mapping $\delta\colon \mathit{Terms}(Q') \to \mathit{Terms}(Q)$ is a query homomorphism if

- $\delta(c) = c$ for every constant c;
- $\delta(x) = x$ for every distinguished variable x of Q';
- $\delta(L') \subseteq L$;
- $M \models \delta(M')$.

Proof. ($\Leftarrow$) Follows from the Homomorphism Theorem for queries with comparisons.
($\Rightarrow$) Suppose $Q \sqsubseteq Q'$, and Q is satisfiable. We construct a mapping $\delta\colon \mathit{Terms}(Q') \to \mathit{Terms}(Q)$ and show that it is a query homomorphism from Q' to Q.
Note that, without loss of generality, we may assume that M does not entail equality between different terms. Hence over the domain of rational numbers, we may assume that there always exists an injective assignment satisfying M.
First, Q is satisfiable, so there exists an injective assignment α for $\mathit{Terms}(Q)$ such that $\alpha \models M$ and $\alpha(c) = c$, for each constant $c \in \mathit{Terms}(Q)$.
Define database instance $\mathbf{I} = \alpha(L)$. Then we have that $\alpha(\vec{x}) \in Q(\mathbf{I})$, and from $Q \sqsubseteq Q'$ it follows that $\alpha(\vec{x}) \in Q'(\mathbf{I})$.

Consequently, we obtain a mapping γ such that $\gamma(\vec{x}) = \alpha(\vec{x})$, $\gamma(L') \subseteq \mathbf{I}$, and $\gamma \models M'$. Moreover, γ maps constants to themselves.
Now, let $\delta = \alpha^{-1} \circ \gamma$. Then δ is a mapping from *Terms*(Q') to *Terms*(Q). It remains to prove that δ is a query homomorphism from Q' to Q.

- Let $c \in$ *Terms*(Q') be a constant. Then $\delta(c) = (\alpha^{-1} \circ \gamma)(c) = \alpha^{-1}(\gamma(c)) = \alpha^{-1}(c) = c$.
- Let x be a distinguished variable of Q'. Then $\delta(x) = (\alpha^{-1} \circ \gamma)(x) = \alpha^{-1}(\gamma(x)) = \alpha^{-1}(\alpha(x)) = x$.
- $\delta(L') = (\alpha^{-1} \circ \gamma)(L') = \alpha^{-1}(\gamma(L')) \subseteq \alpha^{-1}(\mathbf{I}) = L$. Hence, δ is a relational homomorphism.
- Let $\alpha_2 \models M$ be an assignment (not necessarily injective). We show that $\alpha_2 \models \delta(M')$.

  We can repeat the argument above with α, $\mathbf{I}$ and γ, to obtain a database instance $\mathbf{I}_2$ and a mapping γ_2 from L' to $\mathbf{I}_2$ such that $\gamma_2 \models M'$.

  Then, we have that $\alpha_2 \circ \delta$ is a mapping from L' to $\mathbf{I}_2 = \alpha_2(L)$.
  Since there are no selfjoins in L and L', there exists a unique mapping from L' to $\mathbf{I}_2$. Therefore $\alpha_2 \circ \delta = \gamma_2$.
  It follows that $\alpha_2 \circ \delta \models M'$. We are working under the Standard Name Assumption and over the domain of rational numbers, so it holds that $\models (\alpha_2 \circ \delta)(M')$, which is equivalent to $\models \alpha_2(\delta(M'))$. Finally, we obtain that $\alpha_2 \models \delta(M')$.

□

Having proved the characterization of containment of selfjoin-free conjunctive queries, we are ready to devise a polynomial time algorithm that given two queries $Q{:-}L, M$ and $Q'{:-}L', M'$ decides whether $Q \sqsubseteq Q'$:

1. If Q is unsatisfiable, return true.
2. Find a relational homomorphism δ from L' to L. If it does not exist, return false.
3. Check whether $M \models \delta(M')$, that is,
   for each $C \in \delta(M')$, check whether $M \cup \{\neg C\}$ is unsatisfiable. If not, return false.
4. Return true.

The correctness of the algorithm follows from Theorem 1.
From the previous coursework and Exercise 2, Satisfiability of Comparisons, it follows that step 2 and steps 1 and 3 can be done in polynomial time. Hence, this is a polynomial time algorithm.

4. Minimization of Conjunctive Queries

Recall that *relational conjunctive queries* (RCQs) are conjunctive queries without equalities and inequalities. Recall as well that a conjunctive query Q_0 is a *subquery* of another conjunctive query Q if Q_0 can be obtained from Q by dropping some of the atoms in the body of Q.
Prove the following two propositions that provide the underpinnings for the algorithm of conjunctive query minimization.

Proposition 1. *Let Q be a RCQ with n atoms and Q' be an equivalent RCQ with m atoms where $m < n$. Then there exists a subquery Q_0 of Q such that Q_0 has at most m atoms in the body and Q_0 is equivalent to Q.*

Sample solution by Evgeny Kharlamov.

Proof. Given that Q is equivalent to Q', the containments $Q' \sqsubseteq Q$ and $Q \sqsubseteq Q'$ hold. Suppose Q and Q' have the form $Q(\bar{x}) \coloneq L$ and $Q'(\bar{x}) \coloneq L'$, respectively. By the Homomorphism Theorem, there exist substitutions $\delta\colon \mathit{Terms}(Q) \to \mathit{Terms}(Q')$ and $\delta'\colon \mathit{Terms}(Q') \to \mathit{Terms}(Q)$, such that δ' is a homomorphism from Q' to Q and δ is a homomorphism from Q to Q'.
Let us consider the composition $\gamma = \delta' \circ \delta$ of these homomorphisms. Let $L_0 := \tilde{\gamma}L$ be the set of atoms in the range of γ. For each substitution δ, δ', and γ we define corresponding mappings $\tilde{\delta}$, $\tilde{\delta}'$, and $\tilde{\gamma}$ that map atoms to atoms. The function $\tilde{\gamma}$ looks as follows:

$$\tilde{\gamma}\colon \quad L \xrightarrow{\tilde{\delta}} L' \xrightarrow{\tilde{\delta}'} L_0 \ .$$

Obviously, L_0 consists of at most $|L'| = m$ atoms. Observe that by construction $\gamma\bar{x} = \bar{x}$ and $\gamma c = c$ for any constant c in $\mathit{Terms}(Q)$. Hence, every variable in $\bar{x}$ occurs in L_0 and consequently $Q_0(\bar{x}) \coloneq L_0$ defines a query. Note that for the body of the query Q_0 the inclusion $\tilde{\gamma}L \subseteq L_0$ hold. We constructed the query Q_0 in such a way that it is a subquery of Q with at most m atoms in the body. Now we will show that Q_0 is equivalent to Q. Observe that we constructed Q_0 in such a way that γ maps $\mathit{Terms}(Q)$ to $\mathit{Terms}(Q_0)$ and satisfies the conditions from the definition of a homomorphism. Hence, γ is a homomorphism from Q to Q_0 and $Q_0 \sqsubseteq Q$. The containment $Q \sqsubseteq Q_0$ holds, since there exists a homomorphism from Q_0 to Q, which is simply the identity function on $\mathit{Terms}(Q_0)$. Hence, $Q_0 \equiv Q$ and Q_0 has at most m atoms in the body. □

Proposition 2. *Let Q and Q' be two equivalent minimal RCQs. Then Q and Q' are identical up to renaming of variables.*

Sample solution by Evgeny Kharlamov.

Proof. Given the equivalence of Q and Q', there exist two homomorphisms δ from Q to Q' and δ' from Q' to Q, which are mappings

$$\delta\colon \ \mathit{Terms}(Q) \to \mathit{Terms}(Q') \ \text{ and } \ \delta'\colon \mathit{Terms}(Q') \to \mathit{Terms}(Q).$$

Let us consider the composition $\gamma = \delta' \circ \delta$ of these homomorphisms. Using the same reasoning as in Proposition 1, one can show that γ is a homomorphism from Q to its subquery Q_0 and Q is equivalent to Q_0. Using the minimality of Q we obtain that Q_0 coincides with Q. Hence, the composition $\tilde{\gamma} = \tilde{\delta}' \circ \tilde{\delta}$ of the form

$$\tilde{\gamma}\colon \quad L \xrightarrow{\tilde{\delta}} L' \xrightarrow{\tilde{\delta}'} L$$

is surjective. Any surjective mapping of a finite set to itself is injective. The sets L is finite, hence, the composition $\tilde{\gamma}$ is injective, so it is bijective. From this we conclude that $\tilde{\delta}$, being

the first component of $\tilde{\gamma}$, is an injective mapping from L to L'. In a similar way one can show that the composition $\tilde{\gamma'} = \tilde{\delta} \circ \tilde{\delta'}$ is bijective and its first component, namely $\tilde{\delta'}$, is an injective mapping from L' to L. We obtained that $\tilde{\delta}$ and $\tilde{\delta'}$ are injective mappings from the finite set L to the finite set L' and back, respectively. One consequence from this fact is that the sets L and L' have the same cardinality, namely $|L| = |L'|$, and the queries are over the same relational schemas, i.e. the sets of relational names occur in the bodies of Q and Q' are the same. Another consequence is that $\tilde{\delta}$ and $\tilde{\delta'}$ are surjective, moreover, they are bijective.

Observe that if the extension of a substitution (to sets of atoms) is a surjective mapping from one set of atoms to another, then the substitution itself is a surjective mapping from the set of terms occurring in one set of atoms to the set of terms occurring in another. The extension $\tilde{\gamma} = \tilde{\delta'} \circ \tilde{\delta}$ is surjective, hence, the substitution $\gamma = \delta' \circ \delta$ is a surjective mapping from the set of terms $\mathit{Terms}(Q)$ to itself. Similarly, the substitution $\gamma' = \delta \circ \delta'$ is a surjective mapping from the set of terms $\mathit{Terms}(Q')$ to itself. From the surjectivity of γ and γ', using the same reasons as we used in the previous paragraph, we conclude that γ and γ' are bijective, their components δ and δ' are also bijective and $|\mathit{Terms}(Q)| = |\mathit{Terms}(Q')|$.

Observe that the homomorphism δ is a bijective mapping from $\mathit{Terms}(Q)$ to $\mathit{Terms}(Q')$, such that it is the identity mapping on the set of constants and distinguished variables of Q. Hence the inequality $|\mathit{Vars}(Q')| \leq |\mathit{Vars}(Q)|$ holds (in general, δ can map some non-distinguished variables of Q to constants of Q' that do not appear in Q). Analogously, bijectivity of the homomorphism δ' gives us the inequality $|\mathit{Vars}(Q)| \leq |\mathit{Vars}(Q')|$. Hence, the sets of variables in the queries Q and Q' have the same cardinality, namely $|\mathit{Vars}(Q')| = |\mathit{Vars}(Q)|$. We obtained that the equivalent queries Q and Q' are over the same relational schemas, have the same active domains, i.e., the same constants occur in the queries, and the same number of non-distinguished variables. Hence, they are identical up to renaming of variables. In particular, the homomorphisms δ and δ' are such renamings. $\square$