Containment for Rule-Based Ontology-Mediated Queries

Pablo Barceló
MI Foundational Research on Data &
DCC
University of Chile, Chile
pbarcelo@dcc.uchile.cl

Gerald Berger
Institute for Logic and Computation
TU Wien, Austria
gberger@dbai.tuwien.ac.at

Andreas Pieris
School of Informatics
University of Edinburgh, UK
apieris@inf.ed.ac.uk

ABSTRACT

Many efforts have been dedicated to identifying restrictions on ontologies expressed as tuple-generating dependencies (tgds), a.k.a. existential rules, that lead to the decidability of answering ontology-mediated queries (OMQs). This has given rise to three families of formalisms: guarded, non-recursive, and sticky sets of tgds. We study the containment problem for OMQs expressed in such formalisms, which is a key ingredient for solving static analysis tasks associated with them. Our main contribution is the development of specially tailored techniques for OMQ containment under the classes of tgds stated above. This enables us to obtain sharp complexity bounds for the problems at hand.

ACM Reference Format:

1 INTRODUCTION

Motivation and goals. The novel application of knowledge representation tools for handling incomplete and heterogeneous data is giving rise to a new field, recently coined as knowledge-enriched data management [1]. A crucial problem in this field is ontology-based data access (OBDA) [32], which refers to the utilization of ontologies (i.e., sets of logical sentences) for providing a unified conceptual view of various data sources. Users can then pose their queries solely in the schema provided by the ontology, abstracting away from the specifics of the individual sources. In OBDA, one interprets the ontology \( \Sigma \) and the user query \( q \), which is typically a union of conjunctive queries (UCQ), as two components of one composite query \( Q = (S, \Sigma, q) \), known as ontology-mediated query (OMQ); \( S \) is called the data schema, indicating that \( Q \) will be posed on databases over \( S \) [12]. Therefore, OBDA is often realized as the problem of answering OMQs.

While in this setting description logics (DLs) are often used for modeling ontologies, it is widely accepted that for handling arbitrary arity relations in relational databases it is convenient to use tuple-generating dependencies (tgds), a.k.a. existential rules or Datalog\(^*\) rules; cf. [23]. Several aspects of OMQs in which the ontology is a set of tgds and the actual query is a UCQ (simply called OMQs from now on) have been studied in the data management literature; most notably (a) query evaluation [2, 15–17], i.e., given an OMQ \( Q = (S, \Sigma, q) \), a database \( D \) over \( S \), and a tuple of constants \( \bar{c} \), does \( \bar{c} \) belong to the evaluation of \( q \) over every extension of \( D \) that satisfies \( \Sigma \), or, equivalently, is \( \bar{c} \) a certain answer for \( Q \) over \( D \)? and (b) relative expressiveness [12, 25, 26]: how does the expressiveness of OMQs compare to the one of other query languages?

This work focuses on another crucial task for OMQs; namely, containment: for two OMQs \( Q_1 \) and \( Q_2 \) with data schema \( S \), does \( Q_1(D) \subseteq Q_2(D) \) hold for every (finite) database \( D \) over \( S \) (where \( Q(D) \) denotes the certain answers for \( Q \) over \( D \))? Apart from the traditional applications of containment, such as query optimization or view-based query answering, it has been recently shown that OMQ containment has applications on other important static analysis tasks, namely, distribution over components [9], and UCQ rewritability [10]. Despite its prominence, no work to date has carried out an in-depth investigation of containment for OMQs based on tgds. When considered in its full generality, the OMQ containment problem is undecidable. In order to understand which restrictions on the tgds lead to decidability, we recall the two main reasons that render the general containment problem undecidable:

Undecidability of query evaluation: OMQ evaluation is, in general, undecidable [6], and it can be reduced to OMQ containment. More precisely, OMQ containment is undecidable whenever query evaluation for at least one of the involved languages (i.e., the language of the left-hand or the right-hand side query) is undecidable.

Undecidability of containment for Datalog: decidability of query evaluation does not ensure decidability of query containment. A prime example is Datalog, i.e., the OMQ language based on full tgds. Datalog containment is undecidable [34]; thus, OMQ containment is undecidable if the involved languages extend Datalog.

In view of the above observations, we focus on languages that have a decidable query evaluation, and do not extend Datalog. The main classes of tgds, which give rise to OMQ languages with the desirable properties, can be classified into three main families depending on the underlying restrictions: (i) frontier-guarded tgds [2, 15], which contain inclusion dependencies and linear tgds, (ii) non-recursive sets of tgds [22], and (iii) sticky sets of tgds [17].

While the decidability of containment for the above OMQ languages can be established via translations into query languages with a decidable containment problem, such translations do not lead to optimal complexity upper bounds (details are given below). Thus,
the main goal of this work is to develop specially tailored decision procedures for the containment problem under the OMQ languages in question, and, ideally, obtain precise complexity bounds.

Our contributions. The complexity of OMQ containment for the languages in question is given in Table 1. Using small fonts, we recall the complexity of OMQ evaluation. The complexity of OMQ containment – in small fonts, we recall the complexity of OMQ evaluation.

Our contributions as follows:

Linear, non-recursive and sticky sets of tgds. The OMQ languages based on linear, non-recursive, and sticky sets of tgds share a useful property: they are UCQ rewritable (implying in [23]), that is, an OMQ can be rewritten into a UCQ. This property immediately yields decidability for their associated containment problems, since UCQ containment is decidable [33]. However, the obtained complexity bounds are not optimal, since the UCQ rewritings are unavoidably very large [23]. To obtain more precise bounds, we reduce containment to query evaluation, an idea that is often applied in query containment; see, e.g., [18, 19, 33].

Consider a UCQ rewritable OMQ language \(\mathcal{Q}\). If \(Q_1\) and \(Q_2\) belong to \(\mathcal{Q}\), both with data schema \(S\), then we can establish a small witness property, which states that non-containment of \(Q_1\) in \(Q_2\) can be witnessed via a database over \(S\) whose size is bounded by an integer \(k \geq 0\), the maximal size of a disjunct in a UCQ rewriting of \(Q_1\). For linear tgds, such an integer \(k\) is polynomial, but for non-recursive and sticky sets of tgds it is exponential (implicit in [23]). The above small witness property allows us to devise a simple non-deterministic algorithm, which makes use of query evaluation as a subroutine for checking non-containment of \(Q_1\) in \(Q_2\); guess a database \(D\) over \(S\) of size at most \(k\), and then check if there is a certain answer for \(Q_1\) over \(D\) that is not a certain answer for \(Q_2\) over \(D\). This leads to an optimal upper bounds for OMQs based on linear and sticky sets of tgds; however, the exact complexity of OMQs based on non-recursive sets of tgds remains open:

- For OMQs based on linear tgds, the containment problem is in \(\text{PSpace}\), and in \(\Pi^P_2\) if the arity is fixed. The \(\text{PSpace}\)-hardness is shown by reduction from query evaluation, while the \(\Pi^P_2\)-hardness is implicit in [11].
- For OMQs based on sticky sets of tgds, the problem is in \(\text{coNPtime}\), and in \(\Pi^P_2\) if the arity of the schema is fixed. The \(\text{coNPtime}\)-hardness is shown by exploiting the standard tiling problem for the exponential grid, while the \(\Pi^P_2\)-hardness is inherited from [11].
- Finally, for OMQs based on non-recursive sets of tgds, containment is in \(\text{coNPtime}'\) and hard for \(\text{P}^{\text{NP}}\), even for fixed arity. The lower bound is shown by exploiting a recently introduced tiling problem [21].

Guarded tgds. The OMQ language based on guarded tgds is not UCQ rewritable, which forces us to develop different tools to study its containment problem. Let us remark that guarded OMQs can be rewritten as guarded Datalog queries (by exploiting the translations devised in [3, 26]), for which containment is decidable in \(2\text{ExpTime}\) [13]. But, again, the known rewritings are very large [26], and the reduction of containment for guarded OMQs to containment for guarded Datalog does not yield optimal upper bounds.

To obtain optimal bounds for the problem in question, we exploit two-way alternating parity automata on trees (2WAPA) [20, 35]. We show that if \(Q_1\) and \(Q_2\) are guarded OMQs such that \(Q_1\) is not contained in \(Q_2\), then this is witnessed over a class of “tree-like” databases that can be represented as the set of trees accepted by a 2WAPA \(A\). We then build a 2WAPA \(B\) with exponentially many states that recognizes those trees accepted by \(A\) that represent witnesses to non-containment of \(Q_1\) in \(Q_2\). Hence, \(Q_1\) is contained in \(Q_2\) iff \(B\) accepts no tree. Since the emptiness problem for 2WAPA is feasible in exponential time in the number of states [20], we obtain that containment for guarded OMQs is in \(2\text{ExpTime}\). A matching lower bound, even for fixed arity schemas, follows from [10].

Similar ideas based on 2WAPA have been recently used to show that containment for OMQs based on expressive DLs is in \(2\text{ExpTime}\) [10]. In the DL context, schemas consist only of unary and

<table>
<thead>
<tr>
<th>Linear</th>
<th>\text{PSpace}-c</th>
<th>\Pi^P_2-c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sticky</td>
<td>\text{coNPtime}-c</td>
<td>\Pi^P_2-c</td>
</tr>
<tr>
<td>Non-recursive</td>
<td>in \text{coNPtime}' and \text{P}^{\text{NP}}-hard</td>
<td>in \text{coNPtime}' and \text{P}^{\text{NP}}-hard</td>
</tr>
<tr>
<td>Guarded</td>
<td>\text{2ExpTime}-c</td>
<td>\text{2ExpTime}-c</td>
</tr>
<tr>
<td>Frontier-guarded</td>
<td>\text{2ExpTime}-c</td>
<td>\text{2ExpTime}-c</td>
</tr>
</tbody>
</table>

Table 1: Complexity of OMQ containment – in small fonts, we recall the complexity of OMQ evaluation.
binary relations. Our automata construction, however, is different from the one in [10] for two reasons: (a) we need to deal with higher arity relations, and (b) even for unary and binary relations, our OMQ language allows to express properties that are not expressible by the DL-based OMQ languages studied in [10].

Frontier-guarded tgds. Frontier-guarded tgds generalize guarded tgds, and as a matter of fact the techniques we develop for studying OMQ containment for the latter do not extend in a straightforward manner to the former. Instead, we provide a translation from a frontier-guarded OMQ $Q$ into a guarded OMQ $Q'$ such that $Q$ and $Q'$ are equivalent over acyclic databases. This allows to exploit the machinery developed for guarded OMQs, and show that containment for frontier-guarded OMQs is in $2\text{ExpTime}$. As for guarded OMQs, a matching lower bound is inherited from [10], even for fixed arity schema. Let us stress that the employed translation from frontier-guarded into guarded OMQs does not preserve the query answers over arbitrary databases, but only over acyclic databases. This is not surprising since frontier-guarded OMQs are strictly more expressive than guarded OMQs; see, e.g., [25].

Combining languages. The above complexity results refer to the containment problem relative to a certain OMQ language $\mathcal{O}$, i.e., both queries fall in $\mathcal{O}$. However, it is natural to consider the version of the problem where the involved OMQs fall in different languages. Unsurprisingly, if the left-hand side query is expressed in a UCQ rewriter OMQ language (based on linear, non-recursive, or sticky sets of tgds), we can use the algorithm that relies on the small witness property discussed above, which provides optimal upper bounds for almost all the considered cases (the only exception is the containment of sticky in non-recursive OMQs over schemas of unbounded arity). Things are more interesting if the ontology of the left-hand side query is expressed using guarded or frontier-guarded tgds, while the ontology of the right-hand side query is not (frontier-)guarded. By using automata techniques, we show that containment of (frontier-)guarded non-recursive OMQs is in $3\text{ExpTime}$, while containment of (frontier-)guarded in sticky OMQs is in $2\text{ExpTime}$. We establish matching lower bounds, even over schemas of fixed arity, by refining techniques from [19].

Organization. Preliminaries are given in Section 2. In Section 3 we introduce the OMQ containment problem. Containment for UCQ rewriter OMQs is studied in Section 4, for guarded OMQs in Section 5, and for frontier-guarded OMQs in Section 6. We consider the case where the involved queries fall in different languages in Section 7. Finally, we conclude in Section 8.

2 PRELIMINARIES

Databases and conjunctive queries. Let $C$, $N$, and $V$ be disjoint countably infinite sets of constants, (labeled) nulls, and (regular) variables (used in queries and dependencies), respectively. A schema $\Sigma$ is a finite set of relation symbols (or predicates) with associated arity. We write $R/n$ to denote that $R$ has arity $n$. A term is either a constant, null, or variable. An atom over $\Sigma$ is an expression of the form $R(\bar{v})$, where $R \in \Sigma$ is of arity $n \geq 0$ and $\bar{v}$ is an $n$-tuple of terms. A fact is an atom whose arguments consist only of constants. An instance over $\Sigma$ is a (possibly infinite) set of atoms over $\Sigma$ that contain constants and nulls, while a database over $\Sigma$ is a finite set of facts over $\Sigma$. We may call an instance and a database over $\Sigma$ an $\Sigma$-instance and $\Sigma$-database, respectively. The active domain of an instance $I$, denoted $\text{dom}(I)$, is the set of all terms occurring in $I$.

A conjunctive query (CQ) over $\Sigma$ is a formula of the form:

\[ q(\bar{x}) := \exists y_1 \cdots \exists y_m (R_1(\bar{v}_1) \land \cdots \land R_m(\bar{v}_m)), \]

where each $R_i(\bar{v}_i)$ ($1 \leq i \leq m$) is an atom without nulls over $\Sigma$, each variable mentioned in the $\bar{v}_i$'s appears either in $\bar{x}$ or $\bar{y}_i$, and $\bar{x}$ are the free variables of $q$. If $\bar{x}$ is empty, then $q$ is a Boolean CQ. As usual, the evaluation of CQs is defined in terms of homomorphisms. Let $I$ be an instance and $q(\bar{x})$ a CQ of the form (1). A homomorphism from $q$ to $I$ is a mapping $h$, which is the identity on $C$, from the terms that appear in $q$ to the set of constants and nulls $C \cup N$ such that $R_i(h(\bar{v}_i)) \in I$, for each $1 \leq i \leq m$. The evaluation of $q(\bar{x})$ over $I$, denoted $q(I)$, is the set of all tuples $h(\bar{v})$ of constants such that $h$ is a homomorphism from $q$ to $I$. We denote by $\text{CQ}$ the class of conjunctive queries. A union of conjunctive queries (UCQ) over $\Sigma$ is a formula of the form $q(\bar{x}) := q_1(\bar{x}) \lor \cdots \lor q_n(\bar{x})$, where each $q_i(\bar{x})$ is a CQ of the form (1). The evaluation of $q(\bar{x})$ over $I$, denoted $q(I)$, is the set of tuples $\bigcup_{1 \leq i \leq n} q_i(I)$. We denote by $\text{UCQ}$ the class of union of conjunctive queries.

Tgds and the chase procedure. A tuple-generating dependency (tgd) is a first-order sentence of the form:

\[ \forall \bar{x} \exists \bar{y} \left( \psi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z}) \right), \]

where $\phi$ and $\psi$ are conjunctions of atoms without nulls. For brevity, we write this tgd as $\phi(\bar{x}, \bar{y}) \rightarrow \psi(\bar{x}, \bar{z})$ and use comma instead of $\land$ for conjoining atoms. Notice that $\phi$ can be empty, in which case the tgd is called fact tgd and is written as $\top \rightarrow \psi(\bar{x}, \bar{z})$. We assume that each variable in $\bar{x}$ is mentioned in some atom of $\psi$. We call $\phi$ and $\psi$ the body and head of the tgd, respectively. An instance $I$ over $\Sigma$ satisfies the tgd in (2) if the following holds: whenever there is a homomorphism $h$ from $\phi$ to $I$, then $h$ can be extended to a homomorphism $h'$ from $\psi$ to $I$. We say that an instance $I$ satisfies a set $\Sigma$ of tgds, denoted $I \models \Sigma$, if $I$ satisfies every tgd in $\Sigma$. We denote by $\text{TGD}$ the class of (finite) sets of tgds.

The chase is a useful algorithmic tool when reasoning with tgds [15, 22, 29, 31]. We start by defining a single chase step. Let $I$ be an instance over a schema $\Sigma$ and $\tau = \phi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})$ a tgd over $\Sigma$. We say that $\tau$ is applicable w.r.t. $I$ if there exists a tuple $(\bar{a}, \bar{b})$ of terms in $I$ such that $\phi(\bar{a}, \bar{b}) \subseteq I$. In this case, the result of applying $\tau$ over $I$ with $(\bar{a}, \bar{b})$ is the instance $I'$ that extends $I$ with every atom in $\psi(\bar{z}, \bar{a})$, where $\bar{a}$ is the tuple obtained by simultaneously replacing each variable $\bar{z} \in \bar{z}$ with a fresh distinct null not occurring in $I$. For a single chase step we write $I \xrightarrow{\tau,(\bar{a},\bar{b})} I'$.

Let us assume now that $I$ is an instance and $\Sigma$ a finite set of tgds. A chase sequence for $I$ under $\Sigma$ is a sequence:

\[ I_0 \xrightarrow{\tau_0,\bar{a}_0} I_1 \xrightarrow{\tau_1,\bar{a}_1} I_2 \cdots \]

of chase steps such that: (1) $I_0 = I$; (2) for each $i \geq 0$, $\tau_i$ is a tgd in $\Sigma$; and (3) $\bigcup_{1 \leq i \leq n} I_i \models \Sigma$. We call $\bigcup_{1 \leq i \leq n} I_i$ the result of this chase sequence, which always exists. Although the result of a chase sequence is not unique (up to isomorphism), each such result is equally useful for our purposes, since it can be homomorphically embedded into every other result. Henceforth, we denote by chase($I, \Sigma$) the result of an arbitrary chase sequence for $I$ under $\Sigma$. 
Ontology-mediated queries. An ontology-mediated query (OMQ) is a triple \((S, \Sigma, q)\), where \(S\) is a schema, \(\Sigma\) is a set of tgds (the ontology), and \(q\) is a \((U)CQ\) over \(S \cup \text{sch}(\Sigma)\) (and possibly other predicates), with \(\text{sch}(\Sigma)\) the set of predicates occurring in \(\Sigma\). Notice that the set of tgds can introduce predicates not in \(S\), which allows us to enrich the schema of the UCQ \(q\). Moreover, the tgds can modify the content of a predicate \(R \in S\), or, in other words, \(R\) can appear in the head of a tgd of \(\Sigma\). We have explicitly included \(S\) in the specification of the OMQ to emphasize that it will be evaluated over \(S\)-databases, even though \(\Sigma\) and \(q\) might use additional relational symbols. We call \(S\) the data schema.

The semantics of an OMQ is given in terms of certain answers. The certain answers to a UCQ \(q(\bar{x})\) w.r.t. a database \(D\) and a set \(\Sigma\) of tgds is the set of tuples:

\[
\text{cert}(q, D, \Sigma) = \bigcap_{D \subseteq \tau, I \subseteq \Sigma} \{ \tilde{c} \in \text{dom}(I)^{|\bar{x}|} | \tilde{c} \in q(I) \}.
\]

Consider an OMQ \(Q = (S, \Sigma, q)\). The evaluation of \(Q\) over an \(S\)-database \(D\), denoted \(Q(D)\), is defined as \(\text{cert}(q, D, \Sigma)\). It is well-known that \(\text{cert}(q, D, \Sigma) = q(\text{chase}(D, \Sigma))\) (see, e.g., [15]), which immediately implies that \(Q(D) = q(\text{chase}(D, \Sigma))\).

**Ontology-mediated query languages.** We write \((C, Q)\) for the OMQ language that consists of all OMQs of the form \((S, \Sigma, q)\), where \(\Sigma\) falls in the class \(C\) of tgds, i.e., \(C \subseteq \text{TGD}\) (concrete classes of tgds are discussed below), and the query \(q\) falls in \(Q \in \{CQ, UCCQ\}\). A problem that is quite important for our work is OMQ evaluation, defined as follows:

**PROBLEM:** Eval\((C, Q)\)

**INPUT:** An OMQ \(Q = (S, \Sigma, q(\bar{x})) \in (C, Q)\), an \(S\)-database \(D\), and \(\tilde{c} \in \text{dom}(D)^{|\bar{x}|}\).

**QUESTION:** Is \(\tilde{c} \in Q(D)\)?

It is well-known that \(\text{Eval}(\text{TGD}, CQ)\) is undecidable; implicit in [6]. This has led to a flurry of activity for identifying syntactic restrictions on sets of tgds that make the latter problem decidable. Such a restriction defines a subclass \(C\) of tgds. The known decidable classes of tgds are classified into three main decidability paradigms, which, in turn, give rise to decidable OMQ languages:

**Guardedness:** A tgd is guarded (frontier-guarded) if it has a body-atom, called guard (frontier-guard), that contains all the body-variables (all the body-variables that appear in the head of a tgd). A guarded tgd is trivially frontier-guarded, but there are frontier-guarded tgds that are not guarded. Although the chase under (frontier-) guarded tgds does not necessarily terminate, the problem of deciding whether a tuple of constants is a certain answer to a UCQ w.r.t. a database and a set of (frontier-) guarded tgds is decidable. This follows from the fact that the result of the chase has bounded treewidth (see, e.g., [2, 15]). Let \(G\) (resp., \(FG\)) be the class of (finite) sets of guarded (resp., frontier-guarded) tgds. Then:

**Proposition 2.1.** [2, 15] \(\text{Eval}(G, (U)CQ)\) is 2ExpTime-complete, and \(\text{ExpTime}\)-complete for fixed arity. Moreover, the problem \(\text{Eval}(FG, (U)CQ)\) is complete for 2ExpTime, even for fixed arity."
3 OMQ CONTAINMENT: THE BASICS

The goal of this work is to study in depth the problem of checking whether an OMQ is contained in another OMQ, both over the same data schema $S$, or, equivalently, whether $Q_1(D) \subseteq Q_2(D)$ for every finite $S$-database $D$. In this case we write $Q_1 \subseteq Q_2$; we write $Q_1 \equiv Q_2$ if $Q_1 \subseteq Q_2$ and $Q_2 \subseteq Q_1$. The OMQ containment problem in question is defined as follows: $Q_1$ and $Q_2$ are OMQ languages $(C, Q)$, where $C$ is a class of tgds (e.g., linear, non-recursive, sticky, etc.), and $Q \in \{CQ, UCQ\}$.

<table>
<thead>
<tr>
<th>PROBLEM</th>
<th>Cont($Q_1, Q_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INPUT</td>
<td>Two OMQs $Q_1 \in Q_1$ and $Q_2 \in Q_2$.</td>
</tr>
<tr>
<td>QUESTION</td>
<td>Is $Q_1 \subseteq Q_2$?</td>
</tr>
</tbody>
</table>

Whenever $Q_1 = Q_2 = \emptyset$, we refer to the containment problem by simply writing Cont($\emptyset$).

In what follows, we establish some simple but fundamental results, which help to better understand the nature of our problem. We first investigate the relationship between evaluation and containment, which in turn allows us to obtain an initial boundary for the decidability of our problem, i.e., we can obtain a positive result only if the evaluation problem for the involved OMQ languages is decidable (e.g., those introduced in the previous section). We then focus on the OMQ languages introduced in Section 2 and observe that, once we fix the class of tgds, it does not make a difference whether we consider CQs or UCQs. In other words, we show that an OMQ in $(C, UCQ)$, where $C \in \{FG, G, \text{LIN}, \text{NR}, S\}$, can be rewritten as an OMQ in $(C, CQ)$. This fact simplifies our later complexity analysis since for establishing upper (resp., lower) bounds it suffices to focus on CQs (resp., UCQs).

3.1 Evaluation vs. Containment

OMQ evaluation and OMQ containment are strongly connected. As we explain below, the former can be easily reduced to the latter. But let us first introduce some auxiliary notation. Consider a database $D$ and a tuple $\bar{x} = (c_1, \ldots, c_n) \in \text{dom}(D)^n$, where $n \geq 0$. We denote by $q_{D,\bar{x}}(\bar{x})$, where $\bar{x} = (x_1, \ldots, x_n)$, the CQ obtained from the conjunction of atoms occurring in $D$ after replacing each constant $c$ with the variable $x_c$. Consider now an OMQ $Q = (S, \Sigma, q(\bar{x})) \in (C, CQ)$, where $C$ is some class of tgds, an $S$-database $D$, and a tuple $\bar{e} \in \text{dom}(D)^{|\bar{x}|}$. It is not difficult to show that:

$$\bar{e} \in Q(D) \iff (\text{sch}(\Sigma), \emptyset, q_{D,\bar{e}}) \subseteq (\text{sch}(\Sigma), \emptyset, q).$$

Let $O_{\emptyset}$ be the OMQ language that consists of all OMQs of the form $(S, \emptyset, q)$, i.e., the set of tgds is empty, where $q$ is a CQ. It is clear that $Q_1 \subseteq O_{\emptyset}$ and $Q_2 \subseteq O_{\emptyset}$. Therefore, for every OMQ language $O = (C, CQ)$, where $C$ is a class of tgds, we immediately get that:

**Proposition 3.1.** Eval($O$) can be reduced in polynomial time into Cont($O_{\emptyset}, O$).

We now show that the problem of evaluation is also reducible to the complement of containment. Let us say that for technical reasons, which will be made clear in a while, we focus our attention on classes $C$ of tgds that are closed under fact tgd extension, i.e., for every set $\Sigma \in C$, a set obtained from $\Sigma$ by adding a (finite) set of fact tgds is still in $C$. Notice that the classes introduced above enjoy this property. Consider now an OMQ $Q = (S, \Sigma, q(\bar{x})) \in (C, CQ)$, where $C$ is some class of tgds, an $S$-database $D$, and a tuple $\bar{e} \in \text{dom}(D)^{|\bar{x}|}$. It is easy to see that:

$$\bar{e} \in Q(D) \iff (S, \Sigma_D, q^*_D) \not\subseteq (S, \emptyset, \exists x P(x)).$$

where $\Sigma_D^*$ is obtained from $\Sigma$ by renaming each predicate $R$ in $\Sigma$ into $R^* \notin S$ and adding the set of fact tgds:

($T \rightarrow R^*(c_1, \ldots, c_k) \mid R(c_1, \ldots, c_k) \in D$),

$q^*_D$ is obtained from $q(\bar{e})$ by renaming each predicate $R$ into $R^* \notin S$, and the predicate $P$ does not occur in $S$. Indeed, the above equivalence holds since $P \notin S$ implies that $Q_2(D) = \emptyset$, for every $S$-database $D$. Since $C$ is closed under fact tgd extension, $Q_1 \subseteq O_{\emptyset}$, while $Q_2 \subseteq O_{\emptyset}$. We write $\text{coCont}(Q_1, Q_2)$ for the complement of Cont($Q_1, Q_2$). Hence, for every OMQ language $O = (C, CQ)$, where $C$ is a class of tgds (closed under fact tgd extension), it holds that:

**Proposition 3.2.** Eval($O$) can be reduced in polynomial time into $\text{coCont}(O_{\emptyset}, O_{\emptyset})$.

By definition, $O_{\emptyset}$ is contained in every OMQ language $(C, CQ)$, where $C$ is a class of tgds. Therefore, as a corollary of Propositions 3.1 and 3.2, we obtain an initial boundary for the decidability of OMQ containment: we can obtain a positive result only if the evaluation problem for the involved OMQ languages is decidable.

**Corollary 3.3.** Cont($O_1, O_2$) is undecidable if Eval($O_1$) is undecidable or Eval($O_2$) is undecidable.

Can we prove the converse of Corollary 3.3, i.e., Cont($O_1, O_2$) is decidable if both Eval($O_1$) and Eval($O_2$) are decidable? The answer to this question is negative since containment of Datalog queries is undecidable [34]. Indeed, Datalog queries can be directly encoded in the OMQ language based on the class $\mathcal{F}$ of full tgds, i.e., tgds without existentially quantified variables. The next result follows:

**Proposition 3.4.** [34] Cont($\mathcal{F}, CQ$) is undecidable.

This result, combined with the fact that Eval($\mathcal{F}$) is decidable (since the chase under full tgds always terminates), implies that the converse of Corollary 3.3 does not hold. Proposition 3.4 also rules out the OMQ languages that are based on classes of tgds that generalize $\mathcal{F}$, e.g., the weak versions of the ones introduced in Section 2, called weakly frontier-guarded [2], weakly guarded [15], weakly acyclic [22], and weakly sticky [17] that guarantee the decidability of OMQ evaluation. The question that comes up concerns the decidability and complexity of containment for the OMQ languages that are based on the non-weak versions of the above classes, i.e., frontier-guarded, guarded, non-recursive, and sticky. This will be the subject of the next three sections.

3.2 From UCQs to CQs

Before we proceed further, let us state the following useful result:

**Proposition 3.5.** Given an OMQ $Q \in (C, UCQ)$, where $C \in \{FG, G, \text{LIN}, \text{NR}, S\}$, we can construct in polynomial time an OMQ $Q' \in (C, CQ)$ such that $Q \equiv Q'$.

The idea of those classes is the same: relax the condition of the class so that only those positions that receive null values during the chase are taken into account.
We are now ready to proceed with the complexity analysis of containment for the OMQ languages in question. Our plan of attack can be summarized as follows:

- We consider, in Section 4, Cont((C, CQ)), for C ∈ {L, NR, S}. These languages enjoy the key property of UCQ rewritability. This property allows us to show the following result: if the containment does not hold, then this is witnessed via a “small” database, which in turn allows us to devise simple guess-and-check algorithms.

- We then proceed, in Section 5, with Cont((G, CQ)). This OMQ language does not enjoy UCQ rewritability, and the task of establishing a small witness property that leads to an optimal upper bound turned out to be challenging. However, non-containment is witnessed via a “tree-like” database, which allows us to devise a decision procedure based on two-way alternating parity automata on finite trees.

- Cont((FG, CQ)) is studied in Section 6. Recall that (FG, CQ) is strictly more expressive than (G, CQ), and thus, we cannot directly apply the machinery for Cont((G, CQ)). Nevertheless, after focusing on acyclic databases, frontier-guarded OMQs can be rewritten as guarded OMQs, which essentially provides a reduction from Cont((FG, CQ)) to Cont((G, CQ)).

- In Section 7, we study the case where the OMQ containment problem involves two different languages. If the left-hand side language is UCQ rewritable, then we can devise a guess-and-check algorithm based on the above small witness property. The interesting case is when the left-hand side language is (FG, CQ), where we employ tree automata techniques.

## 3.3 Plan of Attack

Proposition 3.5 allows us to focus on OMQs that are based on CQs. Assuming that C1, C2 ∈ {FG, G, L, NR, S} and C is a complexity class that is closed under polynomial time reductions, then:

\[
\text{Cont}((C_1, CQ), (C_2, CQ)) \text{ is } C\text{-complete} \iff \text{Cont}((C_1, UCQ), (C_2, UCQ)) \text{ is } C\text{-complete}.
\]

## 4 UCQ Rewritable Languages

We now focus on OMQ languages that enjoy the crucial property of UCQ rewritability. Roughly, an OMQ language \( \mathcal{O} \) is UCQ rewritable if every query in \( \mathcal{O} \) can be equivalently rewritten as a UCQ. The formal definition follows:

**Definition 4.1. (UCQ Rewritability)** We call an OMQ language \((C, CQ)\), where \(C \subseteq TGD\), UCQ rewritable if, for each OMQ \(Q = (S, \Sigma, q(x)) \in (C, CQ)\), we can construct a UCQ \(q'(x)\) such that \(Q(D) = q'(D)\) for every \(S\text{-database } D\).

We proceed to establish our desired small witness property based on UCQ rewritability. By the definition of UCQ rewritability, for each language \(\mathcal{O}\) that is UCQ rewritable, there exists a computable function \(f_\mathcal{O}\) from \(\mathcal{O}\) to the natural numbers such that the following holds: for every OMQ \(Q = (S, \Sigma, q(x)) \in \mathcal{O}\), and UCQ rewriting \(q_1(x) \lor \cdots \lor q_n(x)\) of \(Q\), it is the case that \(\max_{1 \leq i \leq n} |q_i| \leq f_\mathcal{O}(Q)\), where \(|q_i|\) denotes the number of atoms occurring in \(q_i\).

**Proposition 4.2.** Consider a UCQ rewritable language \(\mathcal{O}\), and two OMQs \(Q \in \mathcal{O}\) and \(Q' \in (TGD, CQ)\), both with data schema \(S\). If \(Q \not\subseteq Q'\), then there exists an \(S\text{-database } D\), where \(|D| \leq f_\mathcal{O}(Q)\), such that \(Q(D) \not\subseteq Q'(D)\).

**Proof (sketch).** We assume that \(q(x) = \sqrt{\sum_{i=1}^n q_i(x)}\) is a UCQ rewriting of \(Q\). Since, by hypothesis, \(Q \not\subseteq Q'\), we conclude that \(q \not\subseteq Q'\), which in turn implies that there exists an \(i \in \{1, \ldots, n\}\) such that \(q_i \not\subseteq Q'\). It is easy to show that \(c(x) \not\subseteq Q'(D_q)\), where \(c(x)\) is a tuple of constants obtained by replacing each variable \(x\) in \(\tilde{x}\) with the constant \(c(x)\), and \(D_q\) is the \(S\text{-database}\) obtained from \(q_i\) after replacing each variable \(x\) in \(q_i\) with the constant \(c(x)\). Thus, \(q(x) \not\subseteq Q(D_q)\), we get that \(c(x) \in Q(D_q)\). Therefore, \(Q(D_q) \not\subseteq Q'(D_q)\), and the claim follows since \(|D_q| \leq f_\mathcal{O}(Q)\).

**Theorem 4.3.** \(\text{Cont}(\mathcal{O}_1, \mathcal{O}_2)\) is decidable if \(\mathcal{O}_1\) is UCQ rewritable and \(\text{Eval}(\mathcal{O}_2)\) is decidable. Given \(Q_1 = (S, \Sigma_1, q_1(x)) \in \mathcal{O}_1\) and \(Q_2 = (S, \Sigma_2, q_2(x)) \in \mathcal{O}_2:\)

1. Guess an \(S\text{-database } D\) such that \(|D| \leq f_{\mathcal{O}_1}(Q_1)\), and a tuple \(\tilde{c} \in \text{dom}(D)^{|\Sigma|}\); and
2. Verify that \(\tilde{c} \in Q_1(D)\) and \(\tilde{c} \not\in Q_2(D)\).

We immediately get that:

**Theorem 4.3.** Cont\((\mathcal{O}_1, \mathcal{O}_2)\) is decidable if \(\mathcal{O}_1\) is UCQ rewritable and Eval\((\mathcal{O}_2)\) is decidable.

This generic result shows that Cont\(((C, CQ))\) is decidable for every class \(C \in \{L, NR, S\}\), but it says nothing about complexity. This will be the subject of the rest of the section.

## 4.1 Linearity

The problem of computing UCQ rewritings for OMQs in \((L, CQ)\) has been studied in [23], where a resolution-based procedure, called XRewrite, has been proposed. This rewriting algorithm accepts a query \(Q = (S, \Sigma, q(x)) \in (L, CQ)\) and constructs a UCQ rewriting \(q'(x)\) over \(S\) by starting from \(q\) and exhaustively applying rewriting steps based on resolution. Due to the fact that the set of tgd's is linear, i.e., the tgd-bodies consist of single atoms, during the execution of XRewrite, it is not possible to obtain a UCQ that has more atoms than the original one. Therefore:

**Proposition 4.4.** \(f_{(L, CQ)}((S, \Sigma, q)) \leq |q|\).

Having the above result in place, it can be shown that the algorithm underlying Theorem 4.3 guesses a polynomially sized witness to non-containment, and then calls a \(C\text{-oracle}\) for solving query evaluation under linear OMQs, where \(C\) is PSpace in general, and NP if the arity is fixed; these complexity classes are obtained from Proposition 2.2. Therefore, coCont\(((L, CQ))\) is in PSpace in general, and in \(\Sigma_2^P\) in case of fixed arity. Regarding the lower bounds, Proposition 3.1 allows us to inherit the PSpace-hardness of Eval\((L, CQ)\); this holds even for constant-free tgd's. Unfortunately, in the case of
fixed arity, we can only obtain \( \text{NP} \)-hardness, while Proposition 3.2 allows to obtain \( \text{coNP} \)-hardness. Nevertheless, it is implicit in [11] (see the proof of Theorem 9), where the containment problem for OMQ languages based on description logics is considered, that \( \text{Cont}(\mathcal{L}, \mathcal{CQ}) \) is \( \Pi_2^P \)-hard, even for tgds of the form \( P(x) \rightarrow R(x) \).

**Theorem 4.5.** \( \text{Cont}(\mathcal{L}, \mathcal{CQ}) \) is \( \text{PSPACE-complete} \), and \( \Pi_2^P \)-complete if the arity of the schema is fixed. The lower bounds hold even for tgds without constants.

### 4.2 Non-Recursiveness

Although the OMQ language \( \mathcal{NR}, \mathcal{CQ} \) is not explicitly considered in [23], where the algorithm \( \text{XRewrite} \) is defined, the same algorithm can deal with \( \mathcal{NR}, \mathcal{CQ} \). By analyzing the UCQ rewritings constructed by \( \text{XRewrite} \), whenever the input query falls in \( \mathcal{NR}, \mathcal{CQ} \), we can establish the following result; here, \( \text{body}(\tau) \) denotes the body of the tgd \( \tau \).

**Proposition 4.6.** It holds that:

\[
\begin{align*}
\hat{f}(\mathcal{NR}, \mathcal{CQ})((S, \Sigma, q)) & \leq |q| \cdot \left( \max_{\tau \in \Sigma}(|\text{body}(\tau)|) \right)^{|\text{sch}(\Sigma)|}.
\end{align*}
\]

Proposition 4.6 implies that non-containment for queries that fall in \( \mathcal{NR}, \mathcal{CQ} \) is witnessed via a database of at most exponential size. We show next that this bound is optimal:

**Proposition 4.7.** There are sets of \( \mathcal{NR}, \mathcal{CQ} \) OMQs

\[
\begin{align*}
\{ Q^n_1 \} = (S, \Sigma^n_1, q^n_1) & \text{ for } n > 0 \quad \text{ and } \quad \{ Q^n_2 \} = (S, \Sigma^n_2, q^n_2) & \text{ for } n > 0,
\end{align*}
\]

where \( |\text{sch}(\Sigma^n_1)| = |\text{sch}(\Sigma^n_2)| = n + 2, \) such that for every \( \text{S-database} D, \) if \( Q^n_1(D) \not\subseteq Q^n_2(D) \) then \( |D| \geq 2^{n-1} \).

Let us now focus on the complexity of \( \text{Cont}(\mathcal{NR}, \mathcal{CQ}) \). By naively combining the algorithm underlying Theorem 4.3 and the exponential bound provided by Proposition 4.6, we get that \( \text{coCont}(\mathcal{NR}, \mathcal{CQ}) \) is feasible in non-deterministic exponential time with access to a \( \text{NExpTime} \) oracle; the oracle is needed for solving \( \text{Eval}(\mathcal{NR}, \mathcal{CQ}) \). Nevertheless, this rough upper bound can be significantly improved; in fact, it can be decreased to \( \text{NExpTime}^{\text{NP}} \), which is nearly optimal (more details are given below), by employing a refined version of the algorithm underlying Theorem 4.3. Recall that \( \text{NExpTime}^{\text{NP}} \) forms the second level of the exponential hierarchy, a.k.a. \( \Sigma_2^{\text{EXP}} \), and it collects all the decision problems that can be solved via an alternating exponential time algorithm with two alternations that starts from an existential state, i.e., it can perform a series of existential steps followed by a series of universal steps. The refined version of the algorithm underlying Theorem 4.3 is such an algorithm.

Before giving this algorithm, let us recall a crucial property of non-recursive OMQs. Given a database \( D \), an OMQ \( (S, \Sigma, q(\vec{x})) \) \( \in \mathcal{NR}, \mathcal{CQ} \), and a tuple \( \vec{c} \in \text{dom}(D)^{|\vec{x}|} \), if \( \vec{c} \in Q(D) \) then there exists a finite chase sequence:

\[
\begin{align*}
D & \rightarrow I_0 \rightarrow I_1 \rightarrow \cdots \rightarrow I_n \rightarrow I_{g(D, \Sigma)},
\end{align*}
\]

for \( D \) under \( \Sigma \), where:

\[
\begin{align*}
g(D, \Sigma) & = |D| \cdot \left( \max_{\tau \in \Sigma}(|\text{body}(\tau)|) \right)^{|\text{sch}(\Sigma)|},
\end{align*}
\]

such that \( \vec{c} \in q(I_{g(D, \Sigma)}) \); implicit in [30]. Having this property in place, we can now present our alternating algorithm. Given \( Q_1 = (S, \Sigma_1, q_1(\vec{x})) \) and \( Q_2 = (S, \Sigma_2, q_2(\vec{x})) \):

1. (1) Guess an \( \text{S-database} D \) of size at most \( f(\mathcal{NR}, \mathcal{CQ})(Q_1) \), and a tuple \( \vec{c} \in \text{dom}(D)^{|\vec{x}|} \).
2. (2) Guess a chase sequence

\[
\begin{align*}
D & \rightarrow I_0 \rightarrow I_1 \rightarrow \cdots \rightarrow I_n \rightarrow I_{g(D, \Sigma)}
\end{align*}
\]

for \( D \) under \( \Sigma_1 \).
3. (3) Guess a mapping \( h \), which is the identity on \( C \), from the variables in \( q_1 \) to \( \text{dom}(I_{g(D, \Sigma)}(\vec{x})) \).
4. (4) If \( h \) is a homomorphism from \( q_1 \) to \( I_{g(D, \Sigma)}(\vec{x}) \) such that \( h(\vec{x}) = \vec{c} \), then proceed; otherwise, reject.
5. (5) Universally select each chase sequence

\[
\begin{align*}
D & \rightarrow I_0 \rightarrow I_1 \rightarrow \cdots \rightarrow I_n \rightarrow I_{g(D, \Sigma)}
\end{align*}
\]

for \( D \) under \( \Sigma_2 \).
6. (6) Universally select each mapping \( h \), which is the identity on \( C \), from the variables in \( q_2 \) to \( \text{dom}(I_{g(D, \Sigma)}(\vec{x})) \).
7. (7) If \( h \) is a homomorphism from \( q_2 \) to \( I_{g(D, \Sigma)}(\vec{x}) \) such that \( h(\vec{x}) = \vec{c} \), then reject; otherwise, accept.

The above algorithm is an alternating exponential time algorithm with two alternations that starts from an existential state. Moreover, it accepts iff \( Q_1 \subseteq Q_2 \), and the desired upper bound follows.

It is not known whether our problem is \( \text{coNExpTime}^{\text{NP}} \)-complete. Nevertheless, we provide a nearly matching lower bound; in fact, \( \text{PExp} \)-hardness. More details on how the above complexity classes are related are discussed below. Let us now explain how \( \text{PExp} \)-hardness is obtained. To this end, we exploit a tiling problem that has been recently introduced in [21]. Roughly speaking, an instance of this tiling problem is a triple \( (m, T_1, T_2) \), where \( m \) is an integer in unary representation, and \( T_1, T_2 \) are standard tiling problems for the \( (2^n \times 2^n) \)-grid. The question is whether, for every initial condition \( w \) of length \( m, T_1 \) has no solution with \( w \) or \( T_2 \) has some solution with \( w \). The initial condition \( w \) simply fixes the first \( m \) tiles of the first row of the grid. We construct in polynomial time two \( (\mathcal{NR}, \mathcal{CQ}) \) queries \( Q_1 \) and \( Q_2 \) such that \( (m, T_1, T_2) \) has a solution iff \( Q_1 \not\subseteq Q_2 \). The idea is to force every input database to store an initial condition \( w \) of length \( m \), and then encode the problem whether \( T_i \) has a solution with \( w \) into \( Q_i \) for each \( i \in \{1, 2\} \). Then:

**Theorem 4.8.** \( \text{Cont}(\mathcal{NR}, \mathcal{CQ}) \) is in \( \text{coNExpTime}^{\text{NP}} \), and \( \text{PExp} \)-hard. The lower bound holds even if the arity of the schema is fixed and the tgds are without constants.

### \text{NExpTime}^{\text{NP}} \text{ vs.} \text{PExp} \text{.}

It is known that \( \text{NExpTime}^{\text{NP}} \) is a delicate class: if we restrict its oracle access too much, it collapses to \( \text{PExp} \) [28]. For example, following the notation of [28], \( \text{PExp} \) coincides with \( \text{NExpTime}^{\text{NP}}(\text{polylog}, \text{exp}) \), where only polynomially many oracle calls are allowed throughout the computation tree of the Turing machine. Also, \( \text{PExp} \) coincides with \( \text{NExpTime}^{\text{NP}}(\text{polylog}(\text{exp}), \text{exp}) \), where only polynomially many oracle calls are allowed on each path of the computation tree, and exponentially many calls with a “yes” answer throughout the computation tree of the Turing machine. The above results support our claim that \( \text{PExp} \) is a nearly matching lower bound for \( \text{Cont}(\mathcal{NR}, \mathcal{CQ}) \).
4.3 Stickiness

We now focus on OMQs that fall in \( \langle S, CQ \rangle \). As shown in [23], given a query \( \langle S, q \rangle \), there exists an execution of XRewrite that constructs a UCQ rewriting \( q_i(\bar{s}) \land \cdots \lor q_n(\bar{s}) \) over \( S \) with the following property: for each \( i \in \{1, \ldots, n\} \), if a variable \( v \) occurs in \( q_i \) in more than one atom, then \( v \) already occurs in \( q \).  This property has been used in [23] to bound the number of atoms that can appear in a single CQ \( q_i \). We write \( T(q) \) for the set of terms (constants and variables) occurring in \( q \); \( C(\Sigma) \) for the set of constants occurring in \( \Sigma \); and \( ar(S) \) for the maximum arity over all predicates of \( S \).

**Proposition 4.9.** It holds that
\[
|f_{\langle S, CQ \rangle}(\langle S, \Sigma, q \rangle)| \leq |S| \cdot (|T(q)| + |C(\Sigma)| + 1)^{ar(S)}.
\]

Proposition 4.9 implies that non-containment for \( \langle S, CQ \rangle \) queries is witnessed via a database of at most exponential size. As for \( \langle \mathbb{NR}, CQ \rangle \) queries, we can show that this bound is optimal; for a set \( S \) of tgds, we write \( |\Sigma| \) for the number of symbols occurring in \( \Sigma \).

**Proposition 4.10.** There exists a set of \( \langle S, CQ \rangle \) OMQs:
\[
\{Q^n = \langle S/n, \Sigma^*, q(\bar{s}) \rangle \}_{n \geq 0}, \text{ where } |\Sigma^n| \in O(n^2),
\]
such that for every \( Q = \langle S, \Sigma^*, q(\bar{s}) \rangle \) \( \in \langle TGD, CQ \rangle \) and \( S \)-database \( D \), if \( Q^n(D) \not\subseteq Q(D) \) then \( |D| \geq 2^n - 2 \).

We now study the complexity of Cont\( \langle S, CQ \rangle \). Let us first look at schemes of unbounded arity. Proposition 4.9 implies that the algorithm underlying Theorem 4.3 runs in exponential time assuming access to a C-oracle, where C is a complexity class powerful enough for solving Eval\( \langle S, CQ \rangle \) and its complement. But, since Eval\( \langle S, CQ \rangle \) is in \( \text{NP} \)-hard, recall that a set of Boolean CQs. In other words, we study the problem for Boolean CQs in \( \text{EXPTIME} \), and thus, the oracle call is not really needed. From this discussion, we conclude that coCont\( \langle S, CQ \rangle \) is in \( \text{EXPTIME} \). A matching lower bound is obtained by a reduction from the standard tiling problem for the \((2^n \times 2^n)\)-grid. In fact, the same lower bound has been recently established in [9]; however, our result is stronger as it shows that the problem remains hard even if the right-hand-side query is a linear OMQ of a simple form – this is also discussed in Section 7, where containment of queries that fall in different OMQ languages is studied. Regarding schemas of fixed arity, Proposition 4.9 provides a witness for non-containment of polynomial size, which implies that the algorithm underlying Theorem 4.3 runs in polynomial time with access to an NP-oracle. Therefore, coEval\( \langle S, CQ \rangle \) is in \( \Sigma_2^p \), while a matching lower bound is implicit in [11].

**Theorem 4.11.** Cont\( \langle S, CQ \rangle \) is \( \text{cO(nEXP-Time-compl)} \), even if the set of tgds uses only two constants. In the case of fixed arity, it is \( \Pi_2^p \)-complete, even for constant-free tgds.

5 GUARDEDNESS

We proceed with the problem of containment for guarded OMQs, and we establish the following result:

**Theorem 5.1.** Cont\( \langle G, CQ \rangle \) is \( \text{2EXPTIME-complete} \). The lower bound holds even if the arity of the schema is fixed, and the tgds are without constants.

The lower bound is immediately inherited from [10], where it is shown that containment for OMQs based on the description logic \( \mathcal{ELI} \) is \( \text{2EXPTIME-hard} \). Recall that a set of \( \mathcal{ELI} \) axioms can be equivalently rewritten as a constant-free set of guarded tgds using only unary and binary predicates, which implies the lower bound stated in Theorem 5.1. However, we cannot immediately inherit the desired upper bound since the DL-based OMQ languages considered in [10] are either weaker than or incomparable to \( \langle G, CQ \rangle \). Nevertheless, the technique developed in [10] was extremely useful for our analysis. Actually, our automata-based procedure exploits a combination of ideas from [10, 27]. The rest of this section is devoted to providing a high-level explanation of this procedure.

For clarity, we focus on constant-free tgds and CQs, but all the results can be extended to the general case at the price of more involved definitions and proofs. Moreover, for simplicity, we focus on Boolean CQs. In other words, we study the problem for \( \langle G, BCQ \rangle \), where \( BCQ \) denotes the class of Boolean CQs. This does not affect the generality of our proof since it is known that Cont\( \langle G, BCQ \rangle \) can be reduced in polynomial time to Cont\( \langle G, CQ \rangle \) [10].

A first glimpse. As said, \( G, CQ \) is not UCQ rewritable and, therefore, we cannot employ Proposition 4.2 in order to establish a small witness property as in Section 4. We have tried, by following a different route, to establish a small witness property for \( \langle G, CQ \rangle \), which can then be used for obtaining an optimal upper bound for Cont\( \langle G, BCQ \rangle \), but it turned out to be a difficult task. Nevertheless, we can show a tree witness property, which states that non-containment for \( \langle G, BCQ \rangle \) is witnessed via a tree-like database. This allows us to devise a procedure based on alternating tree automata. Summing up, the proof for the \( \text{2EXPTIME} \) membership of \( \langle G, CQ \rangle \) proceeds in three steps:

1. Establish a tree witness property;
2. Encode the tree-like witnesses as trees that can be accepted by an alternating tree automaton; and
3. Construct an automaton that decides Cont\( \langle G, CQ \rangle \); in fact, we reduce Cont\( \langle G, CQ \rangle \) into emptiness for two-way alternating parity automata on finite trees.

Each one of the above three steps is discussed in more detail in the following three sections. Let us say that our automata-based approach provides a small witness property for \( \langle G, CQ \rangle \). We obtain that non-containment is witnessed via a triple-exponentially-sized database; details are given below. However, we do not know whether this is optimal.

5.1 Tree Witness Property

From the above informal discussion, it is clear that tree-like databases are crucial for our analysis. Let us make this notion more precise using guarded tree decompositions. A tree decomposition of a database \( D \) is a labeled rooted tree \( T = (V, E, \lambda) \), where \( \lambda : V \to 2^{\text{dom}(D)} \), such that: (i) for each atom \( R(t_1, \ldots, t_n) \in D \), there exists \( v \in V \) such that \( \lambda(v) \supseteq \{t_1, \ldots, t_n\} \), and (ii) for every term \( t \in \text{dom}(D) \), the set \( \{v \in V \mid t \in \lambda(v)\} \) induces a connected subtree of \( T \). The tree decomposition \( T \) is called \( [U] \)-guarded, where \( U \subseteq V \), if, for every node \( v \in V \setminus U \), there exists an atom \( R(t_1, \ldots, t_n) \in D \) such that \( \lambda(v) \subseteq \{t_1, \ldots, t_n\} \). We write root\( T \) for the root node of \( T \), and \( D_T(v) \), where \( v \in V \), for the subset of \( D \) induced by \( \lambda(v) \). We are now ready to formalize the notion of the tree-like database:
Definition 5.2. An S-database D is a C-tree, where C ⊆ D, if there is a tree decomposition T of D such that:

1. \( D_T(\text{root}(T)) = C \) and
2. T is \([[\text{root}(T)]]\)-guarded.

Roughly, whenever a database D is a C-tree, C is the cyclic part of D, while the rest of D is tree-like. For deciding Cont\((\mathcal{G}, \mathcal{B}CQ)\) it suffices to focus on databases that are C-trees and \(|\text{dom}(C)|\) depends only on the left-hand side OMQ. Recall that for a schema S we write \( \text{ar}(S) \) for the maximum arity over all predicates of S. Then:

Proposition 5.3. Let \( Q_i = (V, \Sigma_i, q_i) \in (\mathcal{G}, \mathcal{B}CQ) \), for \( i \in \{1, 2\} \). The following are equivalent:

1. \( Q_1 \subseteq Q_2 \).
2. \( Q_1(D) \subseteq Q_2(D) \), for every C-tree S-database D such that \(|\text{dom}(C)| \leq |(\text{ar}(S) \cup \text{sch}(\Sigma_i)) \cdot |q_i|)|.

The fact that (1) \( \Rightarrow \) (2) holds trivially, while (2) \( \Rightarrow \) (1) is shown by using a variant of the notion of guarded unraveling and compactness. Let us clarify that the above result does not provide a decision procedure for Cont\((\mathcal{G}, \mathcal{B}CQ)\), since we have to consider infinitely many databases that are C-trees with \(|\text{dom}(C)| \leq |(\text{ar}(S) \cup \text{sch}(\Sigma_i)) \cdot |q_i|)|.

5.2 Encoding Tree-like Databases

The treewidth of a database D is the minimum width among all the tree decompositions \( T = (V, E, \lambda) \) of D, while the width of \( T \) is defined as \( \max_{v \in V} \{ |\lambda(v)| \} - 1 \). It is generally known that a database D whose treewidth is bounded by an integer \( k \) can be encoded into a tree over a finite alphabet of double-exponential size in \( k \) that can be accepted by an alternating tree automaton; see, e.g., [7].

Consider an alphabet \( \Gamma \), and let \( \mathbb{N}^\ast \) be the set of finite sequences of natural numbers, including the empty sequence. A \( \Gamma \)-labeled tree is a pair \( L = (N, \lambda) \), where \( N \subseteq \mathbb{N}^\ast \) is closed under prefixes, and \( \lambda : N \rightarrow \Gamma \) is the labeling function. The elements of N identify the nodes of L. It can be shown that D and a tree decomposition T of D with width \( k \) can be encoded as a \( \Gamma \)-labeled tree L, where \( \Gamma \) is an alphabet of double-exponential size in \( k \), such that each node of T corresponds to exactly one node of L and vice versa.

Consider now a C-tree S-database D, and let T be the tree decomposition that witnesses that D is a C-tree. The width of T is at most \( k = (|\text{dom}(C)| + \text{ar}(S) - 1) \), and thus, the treewidth of D is bounded by \( k \). Hence, from the above discussion, D and T can be encoded as a \( \Gamma \)-labeled tree, where \( \Gamma \) is of double-exponential size in \( k \). In general, given an S-database D that is a C-tree due to the tree decomposition T, we show that D and T can be encoded as a \( \Gamma_{S,1} \)-labeled tree, with \(|\text{dom}(C)| \leq l \) and \( |\Gamma_{S,1}| \) being double-exponential in \( \text{ar}(S) \) and exponential in \(|S| \) and \( l \).

Although every C-tree S-database D can be encoded as a \( \Gamma_{S,1} \)-labeled tree, the other direction does not hold. In other words, it is not true that every \( \Gamma_{S,1} \)-labeled tree encodes a C-tree S-database D and its corresponding tree decomposition. In view of this fact, we need the additional notion of consistency. A \( \Gamma_{S,1} \)-labeled tree is called consistent if it satisfies certain syntactic properties – we do not give these properties here since they are not vital in order to understand the high-level idea of the proof. Now, given a consistent \( \Gamma_{S,1} \)-labeled tree L, we can show that L can be decoded into an S-database [L] that is a C-tree with \(|\text{dom}(C)| \leq l \). From the above discussion and Proposition 5.3, we obtain:

Lemma 5.4. Let \( Q_i = (S, \Sigma_i, q_i) \in (\mathcal{G}, \mathcal{B}CQ) \), for \( i \in \{1, 2\} \). The following are equivalent:

1. \( Q_1 \subseteq Q_2 \).
2. \( Q_1([L]) \subseteq Q_2([L]) \), for every consistent \( \Gamma_{S,1} \)-labeled tree L, where \( l = (\text{ar}(S) \cup \text{sch}(\Sigma_i)) \cdot |q_i|) \).

5.3 Constructing Tree Automata

Having the above result in place, we can now proceed with our automata-based procedure. We use two-way alternating parity automata (2WAPA) that run on finite labeled trees. Two-way alternating automata process the input tree while branching in an alternating fashion to successor states, and thereby moving either down or up the input tree. Our goal is to reduce Cont\((\mathcal{G}, \mathcal{B}CQ)\) to the emptiness problem for 2WAPA. As usual, given a 2WAPA \( \mathcal{A} \), we denote by \( \mathcal{L}(\mathcal{A}) \) the language of \( \mathcal{A} \), i.e., the set of labeled trees it accepts. The emptiness problem is defined as follows: given a 2WAPA \( \mathcal{A} \), does \( \mathcal{L}(\mathcal{A}) = \emptyset \)? Thus, given \( Q_1, Q_2 \in (\mathcal{G}, \mathcal{B}CQ) \), we need to construct a 2WAPA \( \mathcal{A} \) such that \( Q_1 \subseteq Q_2 \) if \( \mathcal{L}(\mathcal{A}) \neq \emptyset \). Deciding whether \( \mathcal{L}(\mathcal{A}) \) is empty is feasible in exponential time in the number of states, and in polynomial time in the size of the input alphabet [20]. Therefore, we should construct \( \mathcal{A} \) in double-exponential time, while the number of states must be at most exponential.

We first need a way to check consistency of labeled trees. It is not difficult to devise an automaton for this task.

Lemma 5.5. Consider a schema \( S \) and an integer \( l > 0 \). There is a 2WAPA \( \mathcal{E}_{S,1} \) that accepts a \( \Gamma_{S,1} \)-labeled tree \( L \) if \( l \) is consistent. The number of states of \( \mathcal{E}_{S,1} \) is logarithmic in the size of \( \Gamma_{S,1} \). Furthermore, \( \mathcal{E}_{S,1} \) can be constructed in polynomial time in the size of \( \Gamma_{S,1} \).

Now, the crucial task is, given an OMQ \( Q \in (\mathcal{G}, \mathcal{B}CQ) \), to devise an automaton that accepts labeled trees which correspond to databases that make \( Q \) true.

Lemma 5.6. Let \( Q = (S, \Sigma, q) \in (\mathcal{G}, \mathcal{B}CQ) \). There is a 2WAPA \( \mathcal{A}_{Q,l} \), where \( l > 0 \), that accepts a consistent \( \Gamma_{S,1} \)-labeled tree \( L \) iff \( \mathcal{L}(\mathcal{A}_{Q,l}) \neq \emptyset \). \( \mathcal{A}_{Q,l} \) has exponentially many states in \(||Q|| \) and \( l \), and it can be constructed in double-exponential time in ||Q|| and \( l \).

The intuition underlying \( \mathcal{A}_{Q,l} \) can be described as follows. \( \mathcal{A}_{Q,l} \) tries to identify all the possible ways the CQ query can be mapped to chase\((D, E)\), for any C-tree S-database D such that \(|\text{dom}(C)| \leq l \). It then arrives at possible ways how the input tree can satisfy Q. These “possible ways” correspond to squid decompositions, a notion introduced in [15] which indicates which part of the query is mapped to the cyclic part C of D, and which to the tree-like part of D. The automaton exhaustively checks all squid decompositions by traversing the input tree and, at the same time, explores possible ways how to match the single parts of the squid decomposition at hand. The automaton finally accepts if it finds a squid decomposition that can be mapped to chase\((D, E)\).

Having the above automata in place, we can proceed with our main technical result, which shows that Cont\((\mathcal{G}, \mathcal{B}CQ)\) can be reduced to the emptiness problem for 2WAPA. But let us first recall some key results about 2WAPA, which are essential for our final construction. It is well-known that languages accepted by 2WAPA
are closed under intersection and complement. Given two 2WAPAs $\mathcal{A}_1$ and $\mathcal{A}_2$, we write $\mathcal{A}_1 \cap \mathcal{A}_2$ for a 2WAPA, which can be constructed in polynomial time, that accepts the language $L(\mathcal{A}_1) \cap L(\mathcal{A}_2)$. Moreover, for a 2WAPA $\mathcal{A}$, we write $\mathcal{A}^{-}$ for a 2WAPA, which is also constructible in polynomial time, that accepts the complement of $L(\mathcal{A})$.

We can now show the following:

**Proposition 5.7.** Consider $\mathcal{Q}_1, \mathcal{Q}_2 \in (\mathcal{G}, \mathcal{BCQ})$. We can construct in double-exponential time a 2WAPA $\mathcal{A}$, which has exponentially many states, such that

$$\mathcal{Q}_1 \subseteq \mathcal{Q}_2 \iff L(\mathcal{A}) = \emptyset.$$ 

Proof (sketch). Let $\mathcal{Q}_i = (S, \Sigma_i, q_i)$, for $i \in \{1, 2\}$, and $l = (ar(S \cup \text{sch}(\Sigma_i)) \cdot |q_i|)$. Then $\mathcal{A}$ is defined as

$$(\xi_{S, l} \cap \mathcal{A}_{Q_1, l}) \cap \mathcal{A}_{Q_2, l}.$$ 

Since $\mathcal{A}_{S, l}$ has double-exponential size, Lemmas 5.5 and 5.6 imply that $\mathcal{A}$ can be constructed in double-exponential time, while it has exponentially many states. Lemma 5.4 implies that $\mathcal{Q}_1 \subseteq \mathcal{Q}_2$ iff $L(\mathcal{A}) = \emptyset$, and the claim follows. \hfill $\Box$

Proposition 5.7 implies that $\text{Cont}((\mathcal{G}, \mathcal{BCQ}))$ is in $\text{2ExpTime}$, and Theorem 5.1 follows. The above proposition provides a small witness property for $\text{Cont}((\mathcal{G}, \mathcal{BCQ}))$. In particular, if $\mathcal{Q}_1 \nsubseteq \mathcal{Q}_2$, then this is witnessed via a database $[L]$, where $L$ is a tree accepted by the automaton $\mathcal{A}$ in Proposition 5.7. Since $\mathcal{A}$ has exponentially many states, we can conclude that the trees accepted by $\mathcal{A}$ have size at most triple-exponential. This is because $\mathcal{A}$ can be transformed into a non-deterministic tree automaton with double-exponential many states, which in turn accepts trees of size at most triple-exponential. Therefore, $[L]$ is a triple-exponentially-sized database. It is open whether this is an optimal upper bound.

6 FRONTIER-GUARDEDNESS

We proceed to show that Theorem 5.1 can be extended to OMQs based on frontier-guarded tgds:

**Theorem 6.1.** $\text{Cont}((\mathcal{G}, \mathcal{CQ}))$ is complete for $\text{2ExpTime}$. The lower bound holds even if the arity of the schema is fixed, and the tgds are without constants.

As for $\text{Cont}((\mathcal{G}, \mathcal{CQ}))$, the lower bound is inherited from [10]. The rest of this section is devoted to establish the desired upper bound. As in Section 5, we focus on constant-free tgds and constant-free BCQs, but the result can be extended to the general case. In fact, in order to simplify our analysis even more, let us observe that for containment purposes under OMQs based on frontier-guarded tgds, it suffices to focus on Boolean atomic queries, i.e., BCQs consisting of a single atom; we refer to this class of queries as $\mathcal{BACQ}$. The reason for this is because a BCQ can be seen as a frontier-guarded tgd. More precisely, an OMQ $(\mathcal{S}, \Sigma, q) \in (\mathcal{FG}, \mathcal{BACQ})$ can be equivalently rewritten as the OMQ $(\mathcal{S}, \Sigma \cup \{q \rightarrow \text{Ans}\}, \text{Ans}) \in (\mathcal{FG}, \mathcal{BACQ})$, where each variable in $q \rightarrow \text{Ans}$ is interpreted as a universally quantified variable. From the above discussion, it suffices to show that $\text{Cont}((\mathcal{FG}, \mathcal{BACQ}))$ is in $\text{2ExpTime}$.

Our goal is to provide a reduction from $\text{Cont}((\mathcal{FG}, \mathcal{BACQ}))$ to $\text{Cont}((\mathcal{G}, \mathcal{BAQ}))$, and then apply Theorem 5.1. The main ingredients of our reduction are the following:

1. A query $Q \in (\mathcal{FG}, \mathcal{BAQ})$ can be rewritten as a query $Q' \in (\mathcal{G}, \mathcal{BAQ})$ in such a way that $Q$ and $Q'$ are equivalent over acyclic databases, i.e., databases that have a $[\mathcal{G}]$-guarded tree decomposition.
2. We observe that for $(\mathcal{G}, \mathcal{BAQ})$ we can characterize satisfiability via acyclic databases. In other words, if there exists a database that satisfies a $(\mathcal{G}, \mathcal{BAQ})$ query $Q$, then $Q$ is satisfied by an acyclic database.

Let us make the above statements more formal. The translation of $(\mathcal{FG}, \mathcal{BAQ})$ into $(\mathcal{G}, \mathcal{BAQ})$ relies on the notion of treefication (see, e.g., [4, 5]), and is inspired by a construction given in [5] that translates guarded negation fixed point sentences into guarded fixed point sentences. Our goal is to transform a frontier-guarded tgd into a set of guarded tgds by treefying the body of the former. In fact, the treefication procedure will first transform a tgd-body, which is essentially a CQ, to a set of strictly acyclic CQs, i.e., CQs that are acyclic and have an atom that contains its free variables. Then each strictly acyclic query will give rise to linearly many guarded tgds. Let us now recall treefications.

Consider a CQ $q(x)$ over a schema $S$. The $T$-treefication of $q(x)$, where $T \supseteq S$, is the set $\mathcal{A}_T^T$ of all strictly acyclic CQs $q'(x)$ over $T$ of size at most $3|q|$ such that $(i) q' \subseteq q$, and $(ii)$ is minimal, i.e., by removing an atom would render into a CQ that is not strictly acyclic or $q' \nsubseteq q$. The set $\mathcal{A}_T^T$ can be seen as the UCO $\mathcal{A}_T^T$ defined as the disjunction of all CQs contained in $\mathcal{A}_T^T$. Notice that the query $q(x)$ is in general not equivalent to its $T$-treefication. However, $q(x)$ and $\mathcal{A}_T^T(x)$ are equivalent over acyclic T-databases [4, 5].

We are now ready to explain how a frontier-guarded OMQ is transformed into a guarded OMQ. Consider a frontier-guarded tgd $t : \phi(x, y) \rightarrow \exists z \psi(x, z)$ and a schema $T$. Let $f_T^\mathcal{T}(\tau)$, where $\mathcal{C}$ is a predicate not in $T$, be the set of tgds

$$\left\{ q(x) \rightarrow \exists z \psi(x, z) \mid q(x) \in \mathcal{A}_T^T(\{C\}) \right\}.$$ 

Notice that the tgds in $f_T^\mathcal{T}(\tau)$ may not be guarded. However, by construction, their bodies are strictly acyclic CQs, and this allows us to rewrite each tgd in $f_T^\mathcal{T}(\tau)$ into linearly many guarded tgds, which we denote by $g^\mathcal{T}(\tau)$. Given an OMQ $Q = (S, \Sigma, q) \in (\mathcal{FG}, \mathcal{BAQ})$, let

$$g^\mathcal{T}(\tau) = \left( \bigcup_{\tau \in \Sigma} S_{\tau \cdot \text{sch}(\Sigma)}^\mathcal{T}(\tau), q \right) \in (\mathcal{G}, \mathcal{BAQ}),$$

where $\mathcal{C}$ is an auxiliary predicate not in $S \cup \text{sch}(\Sigma)$. This completes the translation from frontier-guarded to guarded OMQs. We can show the following crucial lemma, which actually formalizes the first intuitive statement given above. Given a schema $S$ and a predicate $\mathcal{C}/n \in S$, for brevity, we write $\mathcal{S}_C$ for $S \cup \{C\}$. Given an S-database $D$, let $D_C$ be the $S_C$-database $D \cup \{C(t) \mid t \in \text{dom}(D)^n\}$. By the width of an OMQ $Q$, written $\text{width}(Q)$, we mean the maximum number of variables in the body of a tgd of $Q$.

**Lemma 6.2.** Let $Q = (S, \Sigma, q) \in (\mathcal{FG}, \mathcal{BAQ})$, and $Q' = g^\mathcal{T}(\tau)$, where $C \notin S$ has arity at least $\text{width}(Q)$. Then:

1. For each acyclic $\mathcal{S}_C$-database $D$, $Q(D) = Q'(D)$.
2. For each $\mathcal{S}$-database $D$, $Q(D) \neq \emptyset \Rightarrow Q'(D) \neq \emptyset$.

Let us now formalize the second intuitive statement given above. Actually, the next result is implicit in the proof of Proposition 5.3,
which establishes that non-containment for \((G, CQ)\) is witnessed via a tree-like database. We write \(I \rightarrow D\) for the fact that the instance \(I\) can be mapped via a homomorphism to the database \(D\).

**Lemma 6.3.** Consider an \(S\)-database \(D\), and an OMQ \(Q = (S, Σ, q) ∈ (G, BAQ)\). If \(Q(D) \neq ∅\), then there is a finite acyclic \(S\)-instance \(I\) such that \(Q(I) \neq ∅\) and \(I \rightarrow D\).

Having the above lemmas in place, it is easy to show that \(g_c(\cdot)\) provides a reduction from \(\text{Cont}((PG, BAQ))\) to \(\text{Cont}((G, BAQ))\), if the arity of \(C\) is sufficiently large.

**Proposition 6.4.** Let \(Q_i = (S, Σ_i, q_i) ∈ (G, BAQ)\), for \(i \in \{1, 2\}\), and consider a predicate \(C \not∈ (S \cup \text{sch}(Σ_1) \cup \text{sch}(Σ_2))\) that has arity \(\max_{i \in \{1, 2\}}|\text{width}(Q_i)|\). Then,

\[
Q_1 \subseteq Q_2 ⇔ g_c(Q_1) \subseteq g_c(Q_2).
\]

Proof (sketch). Let \(Q_i' = g_c(Q_i)\), for \(i \in \{1, 2\}\). Assume that \(Q_1 \not\subseteq Q_2\). This implies that there exists an \(S\)-database \(D\) such that \(Q_1(D) \neq ∅\) and \(Q_2(D) = ∅\). By Lemma 6.3, there exists a finite acyclic \(S\)-instance \(I\) such that \(Q_1(I) \neq ∅\) and \(I \rightarrow D\). Since \(Q_2(D) = Q_2(D) = ∅\), and \(Q_2\) is closed under homomorphisms, \(Q_2(I) = ∅\). Consequently, by Lemma 6.2, \(Q_2'(I) = ∅\), which implies that \(Q_1' \not\subseteq Q_2'\). The other direction can be shown analogously.

The above proposition provides the desired reduction from \(\text{Cont}((PG, BAQ))\) to \(\text{Cont}((G, BAQ))\), which allows us to apply the algorithm for \(\text{Cont}((G, CQ))\), devised in Section 5. However, it should not be overlooked that this reduction takes exponential time due to the treeification procedure. In fact, for a CQ \(q\), \(|A_c|^w \leq |T|^{|Q(q)|}(|q|w)|O(q)|w\), where \(w\) is the maximum arity over all predicates of \(T\) [4, 5]. Nevertheless, since the reduction provided by Proposition 6.4 increases the arity of the schema only polynomially, while the algorithm for \(\text{Cont}((G, BAQ))\) provided by Theorem 5.1 is double-exponential only on the arity of the underlying schema, we obtain that \(\text{Cont}((PG, BAQ))\) is feasible in double-exponential time, as needed.

We conclude this section by noticing that, as for guarded OMQs, we get a small witness property for \(\text{Cont}((PG, CQ))\), which states that non-containment is witnessed via a triple-exponentially-sized database. More precisely, \(Q_1 \not\subseteq Q_2\) implies \(g_c(Q_1) \not\subseteq g_c(Q_2)\), and we can show that the latter non-containment is witnessed via a triple-exponentially-sized acyclic database \(D\). Since, by Lemma 6.2, \(Q_1\) and \(g_c(Q_1)\), for \(i \in \{1, 2\}\), are equivalent over acyclic databases, \(D\) is a witness for \(Q_1 \not\subseteq Q_2\).

### 7 COMBINING LANGUAGES

In the previous three sections, we studied the containment problem relative to a language \(G\), i.e., both OMQs fall in \(G\). However, it is natural to consider the version of the problem where the involved OMQs fall in different languages. This is the goal of this section. Our analysis proceeds by considering the two cases where the left-hand side (LHS) query falls in a UCQ rewritable OMQ language, or it is guarded. Notice that the two cases where the LHS query is guarded or frontier-guarded behave in the same way. Thus, for brevity, we only focus on the former case.

#### 7.1 The LHS Query is UCQ Rewritable

As an immediate corollary of Theorem 4.3 we obtain the following result: \(\text{Cont}((C_1, CQ), (C_2, CQ))\), for \(C_1 \neq C_2\), \(C_1 \in \{L, NR, S\}\) and \(C_2 \in \{L, NR, S, PG, G\}\) is decidable. By exploiting the algorithm underlying Theorem 4.3, we establish optimal upper bounds for all the problems at hand with the only exception of \(\text{Cont}((S, CQ), (NR, CQ))\). For the latter, we obtain a co\(\text{NP}\)-hard upper bound, by providing a similar analysis as for \(\text{Cont}(( NR, CQ))\), while a \(\text{NP}\)-hard lower bound is inherited from query evaluation by exploiting Proposiiton 3.1. It is rather tedious to go through all the containment problems in question and explain in detail how the exact upper bounds are obtained.\(^3\)

Regarding the matching lower bounds, in most of the cases they are inherited from query evaluation by exploiting Propositions 3.1 and 3.2. There are, however, some exceptions:

- \(\text{Cont}((S, CQ), (L, CQ))\) in the case of unbounded arity, where the problem is co\(\text{NP}\)-hard, even for sets of tgd's that use only two constants. This is shown by a reduction from the standard tiling problem for the exponential grid \(2^n \times 2^n\).
- \(\text{Cont}((L, CQ), (S, CQ))\) and \(\text{Cont}((S, CQ), (L, CQ))\) in the case of bounded arity, where both problems are \(\Pi^p_2\)-hard even for constant-free tgd's; implicit in [11].

#### 7.2 The LHS Query is Guarded

We proceed with the case where the LHS query is guarded, and we show the following result:

**Theorem 7.1.** The problem \(\text{Cont}((G, CQ), (C, CQ))\) is \(\mathcal{C}\)-complete, where:

\[
\mathcal{C} = \begin{cases}
2\text{ExpTime}, & \text{if } C ∈ \{L, S\}, \\
3\text{ExpTime}, & \text{if } C = \text{NR}.
\end{cases}
\]

The lower bounds hold even if the arity of the schema is fixed. Moreover, for \(C = \mathcal{L}\) (resp., \(C ∈ \{\text{NR, S}\}\)) it holds even for tgd's with one constant (resp., without constants).

**Upper bounds.** The \(2\text{ExpTime}\) membership when \(C = \mathcal{L}\) is an immediate corollary of Theorem 5.1. This is not true when \(C ∈ \{\text{NR, S}\}\) since the right-hand side query is not guarded. But in this case, since \((\text{NR}, CQ)\) and \((S, CQ)\) are UCQ rewritable, one can rewrite the right-hand side query as a UCQ, and then apply the machinery developed in Section 5 for solving \(\text{Cont}((G, CQ))\). More precisely, given OMQs \(Q_1 \in (G, CQ)\) and \(Q_2 \in (C, CQ)\), where \(C ∈ \{\text{NR, S}\}\), \(Q_1 \subseteq Q_2\) iff \(Q_1 \subseteq q\), where \(q\) is a UCQ rewriting of \(Q_2\). Thus, an immediate decision procedure, which exploits the algorithm XRewrite, is the following:

1. Let \(q = \text{XRewrite}(Q_2)\);
2. For each \(q' \in q\), if \(Q_1 \subseteq q'\), then proceed; otherwise, reject; and
3. (Accept.

\(^3\)There are twenty-four different cases obtained by considering all the possible pairs \((G_1, G_2)\) of OMQ languages, where \(G_1 \neq G_2\) and \(G_1\) is UCQ rewritable, and the two cases whether the arity of the schema is fixed or not.
The above procedure runs in triple-exponential time. The first step is feasible in double-exponential time [23]. Now, for a single CQ $q' \in q$ (which is a guarded OMQ with an empty set of tgds) the check whether $Q_1 \subseteq q'$ can be done by using the machinery developed in Section 5, which reduces our problem to checking whether the language of a 2WAPA $\mathcal{H}$ is empty. However, it should not be forgotten that $q'$ is of exponential size, and thus, the automaton $\mathcal{H}$ has double-exponentially many states. This in turn implies that checking whether $L(\mathcal{H}) = \varnothing$ is in 3ExpTime, as claimed.

Although the above algorithm establishes an optimal upper bound for non-recursive OMQs, a more refined analysis is needed for sticky OMQs. In fact, we need a more refined complexity analysis for the problem $\text{Cont}((G, \mathbb{C}Q), UCQ)$, that is, to decide whether a guarded OMQ is contained in a UCQ. To this end, we provide an automata construction different from the one employed in Section 5, which allows us to establish a refined complexity upper bound for the problem in question. Consider a $(G, \mathbb{C}Q)$ query $Q$, and a UCQ $q = q_1 \lor \cdots \lor q_n$. As usual, we write $||Q||$ and $||q_i||$ for the number of symbols that occur in $Q$ and $q_i$, respectively, and we write $\text{var}_{\geq 2}(q_i)$ for the set of variables that appear in more than one atom of $q_i$. By exploiting our new automata-based procedure, we show that the problem of checking if $Q \subseteq q$ is feasible in double-exponential time in $(||Q|| + \max_{1 \leq i \leq n}||\text{var}_{\geq 2}(q_i)||)$, exponential time in $\max_{1 \leq i \leq n}||q_i||$, and polynomial time in $n$.

This result allows us to show that the above procedure establishes 2ExpTime-membership when the right-hand side OMQ is sticky. But first we need to recall the following key properties of the UCQ rewriting $q = \text{XRewrite}(Q_2)$, constructed during the first step of the algorithm:

1. $q$ consists of double-exponentially many CQs,
2. each CQ of $q$ is of exponential size, and
3. for each $q' \in q$, $\text{var}_{\geq 2}(q')$ is a subset of the variables of the original CQ that appears in $Q_2$.

By combining these key properties with the complexity analysis performed above, it is now straightforward to show that $\text{Cont}((G, \mathbb{C}Q), (\mathbb{S}, \mathbb{C}Q))$ is in 2ExpTime.

**Lower Bounds.** We establish matching lower bounds by refining techniques from [19], where it is shown that containment of Datalog in UCQ is 2ExpTime-complete, while containment of Datalog in non-recursive Datalog is 3ExpTime-complete; the lower bounds hold for fixed-arity predicates, and constant-free rules. Interestingly, the LHS query can be transformed into a Datalog query such that each rule has a body-atom that contains all the variables, i.e., is guarded. This is achieved by increasing the arity of some predicates in order to have enough positions for all the body-variables. However, for each rule, the number of unguarded variables that we need to guard is constant, and thus, the arity of the schema remains constant. We conclude that $\text{Cont}((G, \mathbb{C}Q), (\mathbb{N}, \mathbb{C}Q))$ is 3ExpTime-hard. Moreover, containment of guarded OMQs in UCQs is 2ExpTime-hard, which in turn allows us to show, by exploiting the construction underlying Proposition 3.5, that $\text{Cont}((G, \mathbb{C}Q), (L, \mathbb{C}Q))$ is 2ExpTime-hard, even if the set of linear tgds uses only one constant, while $\text{Cont}((G, \mathbb{C}Q), (\mathbb{S}, \mathbb{C}Q))$ is 2ExpTime-hard, even for tgds without constants.

**8 CONCLUSIONS**

We have concentrated on the fundamental problem of containment for OMQ languages based on the main decidable classes of tgds, and we have developed specially tailored techniques that allow us to obtain a relatively complete picture for the complexity of the problem at hand. Our main conclusion is that for most of the OMQ languages in question, the containment problem is harder (under widely accepted complexity assumptions) than query evaluation.

**Acknowledgements.** Barceló is funded by the Millennium Institute for Foundational Research on Data and Fondecyt grant 1170109. Berger is funded by the Austrian Science Fund (FWF), project number W1255-N23 and DOC fellowship of the Austrian Academy of Sciences. Pieris is funded by the EPSRC Programme Grant EP/M025268/ “VADA: Value Added Data Systems - Principles and Architecture”.

**REFERENCES**

89–124.


