Reverse Engineering Queries in Ontology-Enriched Systems: The Case of Expressive Horn Description Logic Ontologies

Víctor Gutiérrez-Basulto
Cardiff University, UK
gutierrezbasultov@cardiff.ac.uk

Jean Christoph Jung
University of Bremen, Germany
KU Leuven, Belgium
jeanjung@uni-bremen.de

Leif Sabellek
University of Bremen, Germany
sabellek@uni-bremen.de

Abstract

We introduce the query-by-example (QBE) paradigm for query answering in the presence of ontologies. Intuitively, QBE permits non-expert users to explore the data by providing examples of the information they (do not) want, which the system then generalizes into a query. Formally, we study the following question: given a knowledge base and sets of positive and negative examples, is there a query that returns all positive but none of the negative examples? We focus on description logic knowledge bases with ontologies formulated in Horn-$\mathcal{ALCI}$ and (unions of) conjunctive queries. Our main contributions are characterizations, algorithms and tight complexity bounds for QBE.

1 Introduction

In recent times, ontology-enriched systems (OES) have risen as a prominent technology for data management. The appeal of OES comes from the fact that the ontology provides rich schema information or background knowledge which enriches the answers of queries. The success of this paradigm has led not only to the development of a vast amount of foundational results, but also of optimized systems used in real-life scenarios, see e.g., [Rodriguez-Muro *et al.*, 2013; Kharlamov *et al.*, 2015; Calvanese *et al.*, 2016; Hovland *et al.*, 2017] and references therein. For instance, the OES Ontop is currently being used to access exploration data generated by the petroleum company Statoil [Kharlamov *et al.*, 2015]. In these OES, users access the data through queries usually formulated in powerful query languages such as conjunctive or path queries. Unfortunately, in real life, casual non-expert users are often not able to specify queries using these formalisms (e.g., Statoil geologists [Hovland *et al.*, 2017]), clearly hampering the usability of OES.

In relational databases (witnessing the same problem), an alternative approach for querying was proposed to alleviate this problem: *query-by-example (QBE)*, where roughly, users give positive and negative examples which the system should reverse-engineer into a query conforming with the examples [Zloof, 1975]. Because of 'big data', this querying paradigm has lately gained new interest since even expert users might find it useful to explore the data in this way. As a result, QBE has been investigated for different query languages and data representations, e.g., conjunctive queries over relational data [Tran *et al.*, 2014; ten Cate and Dalmau, 2015; Bonifati *et al.*, 2016; Barceló and Romero, 2017], SPARQL queries over RDF data [Arenas *et al.*, 2016], and path queries over graph databases [Bonifati *et al.*, 2015].

The goal of this paper is two-fold. First, we aim at initiating research on the QBE approach to querying in the context of ontology-enriched systems. We mainly focus on establishing foundational results for QBE over OES with the ontology formulated in description logics (DLs). Formally, we introduce and study the following problem $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ for an ontology language $\mathcal{L}$ and some query language $\mathcal{Q}$: given an $\mathcal{L}$-knowledge base and sets of positive and negative examples, decide whether there is a query $q \in \mathcal{Q}$ such that all positive examples are certain answers to q over $\mathcal{K}$, and none of the negative is. As query language $\mathcal{Q}$, we consider (unions of) conjunctive queries, (U)CQs. We allow for a restricted signature Σ, which is a common feature in many OES. As a simple example, consider the knowledge base consisting of

$$\mathcal{T} = \{\mathsf{Human} \sqsubseteq \mathsf{Vertebrate}, \mathsf{Vertebrate} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{Spine}\},$$
$$\mathcal{A} = \{\mathsf{Human}(ax), \mathsf{hasPart}(an, sp), \mathsf{Spine}(sp), \mathsf{Bug}(bug)\}.$$

If the positive examples are ax, an and the negative example is bug, then $q(x) = \exists y\, \mathsf{hasPart}(x, y) \wedge \mathsf{Spine}(y)$ is a witness query. However, there is no witnessing query for the positive examples an, bug if ax is to be avoided.

The second aim is to continue bridging the gap between DL and machine learning research. Indeed, QBE over knowledge bases can be viewed as an instantiation of the *inductive logic programming (ILP)* framework [Nienhuys-Cheng and de Wolf, 1997]: the background knowledge is given by a DL knowledge base and the learning goal are single rules for $\mathcal{Q} = \text{CQ}$ and sets of rules with the same head for $\mathcal{Q} = \text{UCQ}$, respectively. In this area, the work closest to ours is perhaps [Kietz, 2002].

Our main contributions are characterizations, algorithms, and complexity bounds for $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ for $\mathcal{L}$ an expressive Horn DL $\mathcal{L} \in \{\text{Horn-}\mathcal{ALCI}, \text{Horn-}\mathcal{ALC}\}$ and $\mathcal{Q} \in \{\text{CQ}, \text{UCQ}\}$. In Section 3, we start with providing natural model-theoretic characterizations for $\mathsf{QBE}(\text{Horn-}\mathcal{ALCI}, \mathcal{Q})$ for $\mathcal{Q} \in \{\text{CQ}, \text{UCQ}\}$ by lifting characterizations known from the relational database setting [ten Cate and Dalmau, 2015] by replacing the database with the universal model of the knowledge base. Unfortunately, our characterizations do not give

immediate rise to a decision procedure because the universal model is typically infinite. In Section 4, we exploit the regularity of universal models and provide decision procedures running in 2-ExpTime and coNExpTime for Horn-$\mathcal{ALCI}$ and Horn-$\mathcal{ALC}$, respectively. Having these, we prove matching lower bounds, the most challenging one being a 2-ExpTime-lower bound for $\mathsf{QBE}(\text{Horn-}\mathcal{ALCI}, \mathcal{Q})$, $\mathcal{Q} \in \{\text{CQ}, \text{UCQ}\}$. Interestingly, some results depend on restricting the signature, so we consider also the variant QBE_f of QBE with unrestricted signature. The following table summarizes our results.

$\mathcal{L} \rightarrow$	Horn-$\mathcal{ALCI}$	Horn-$\mathcal{ALC}$
$\mathsf{QBE}(\mathcal{L}, \text{CQ})$	2-ExpTime	coNExpTime
$\mathsf{QBE}(\mathcal{L}, \text{UCQ})$	2-ExpTime	ExpTime
$\mathsf{QBE}_f(\mathcal{L}, \text{CQ})$	2-ExpTime	coNExpTime
$\mathsf{QBE}_f(\mathcal{L}, \text{UCQ})$	ExpTime	ExpTime

We obtain the same results for the variant QDEF of QBE, the problem to decide whether some $q \in \mathcal{Q}$ returns *precisely* the positive examples. In Section 5, we investigate the *size of witness queries*. This is of course vital for practical purposes since at the end the user is interested in obtaining a (witness) query to further explore the data. We particularly show that they can be double exponentially large, which is in contrast to the relational database setting. In Section 6, we discuss related work and lay out directions for future work.

An extended version with appendix can be found under www.informatik.uni-bremen.de/tdki/research/papers.html.

2 Preliminaries

Syntax. We introduce the DL Horn-$\mathcal{ALCI}$ [Krötzsch *et al.*, 2013]. Let $\mathsf{N_C}, \mathsf{N_R}, \mathsf{N_I}$ be infinite disjoint sets of *concept, role,* and *individual names*, respectively. The syntax of *Horn-$\mathcal{ALCI}$ concepts* C, D is given by the grammar:

$$B, B' ::= \top \mid \bot \mid A \mid B \sqcap B' \mid B \sqcup B' \mid \exists r.B$$
$$C, D ::= \top \mid \bot \mid A \mid \neg A \mid C \sqcap D \mid \neg B \sqcup C \mid \exists r.C \mid \forall r.C$$

where $A \in \mathsf{N_C}$ and $r \in \{s, s^- \mid s \in \mathsf{N_R}\}$ is a *role*. Concepts of the form B are called *basic concepts* and roles of the form r^- *inverse roles.* We identify r^- with $s \in \mathsf{N_R}$ if $r = s^-$.

A *Horn-$\mathcal{ALCI}$ TBox (ontology)* $\mathcal{T}$ is a finite set of *concept inclusions (CIs)* $B \sqsubseteq C$, with B a basic concept and C a Horn-$\mathcal{ALCI}$ concept. An *ABox* $\mathcal{A}$ is a finite set of *concept* and *role assertions* of the form $A(a)$ and $r(a, b)$, where $A \in \mathsf{N_C}$, $r \in \mathsf{N_R}$ and $a, b \in \mathsf{N_I}$. We write $\mathsf{ind}(\mathcal{A})$ for the set of individuals in $\mathcal{A}$. A *Horn-$\mathcal{ALCI}$ knowledge base (KB)* $\mathcal{K}$ is a pair $(\mathcal{T}, \mathcal{A})$ of a Horn-$\mathcal{ALCI}$ TBox $\mathcal{T}$ and an ABox $\mathcal{A}$. The fragment Horn-$\mathcal{ALC}$ is obtained by disallowing inverse roles; $\mathcal{ELI}$ is the fragment allowing only concept inclusions $C \sqsubseteq D$ with $C, D ::= \top \mid A \mid C \sqcap D \mid \exists r.C$.

Semantics. The semantics is defined in terms of interpretations $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, consisting of a non-empty *domain* $\Delta^{\mathcal{I}}$ and an *interpretation function* $\cdot^{\mathcal{I}}$ mapping concept names to subsets of the domain and role names to binary relations over the domain. Further, we adopt the *standard name assumption*, i.e., $a^{\mathcal{I}} = a$ for all $a \in \mathsf{N_I}$. The interpretation of complex concepts $C^{\mathcal{I}}$ is defined in the usual way [Baader *et al.*, 2017]. An interpretation $\mathcal{I}$ is a *model of a TBox* $\mathcal{T}$ if $B^{\mathcal{I}} \subseteq C^{\mathcal{I}}$ for all CIs $B \sqsubseteq C \in \mathcal{T}$; and it is a *model of an ABox* $\mathcal{A}$ if $(a, b) \in r^{\mathcal{I}}$ for all $r(a, b) \in \mathcal{A}$ and $a \in A^{\mathcal{I}}$ for all $A(a) \in \mathcal{A}$. We call a KB $(\mathcal{T}, \mathcal{A})$ *consistent* if $\mathcal{T}$ and $\mathcal{A}$ have a common model.

Queries. A *conjunctive query (CQ)* is an expression of the form $q(\mathbf{x}) = \exists \mathbf{y}\, \varphi(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are tuples of variables and $\varphi(\mathbf{x}, \mathbf{y})$ is a conjunction of *atoms* of the form $A(v)$ or $r(v, w)$ with $A \in \mathsf{N_C}$, $r \in \mathsf{N_R}$, and $v, w \in \mathbf{x} \cup \mathbf{y}$. We call $\mathbf{x}$ *answer variables* and $\mathbf{y}$ *quantified variables* of q. A *union of conjunctive queries (UCQ)* is an expression of the form $q(\mathbf{x}) = q_1(\mathbf{x}) \vee \ldots \vee q_n(\mathbf{x})$, where each $q_i(\mathbf{x})$ is a CQ with answer variables $\mathbf{x}$. A *match* of a CQ q in an interpretation $\mathcal{I}$ is a function $\pi : \mathbf{x} \cup \mathbf{y} \to \Delta^{\mathcal{I}}$ such that $\pi(v) \in A^{\mathcal{I}}$ for every atom $A(v)$ of q and $(\pi(v), \pi(w)) \in r^{\mathcal{I}}$ for every atom $r(v, w)$ of q. We write $\mathcal{I} \models q(a_1, \ldots, a_n)$ if there is a match of q in $\mathcal{I}$ with $\pi(x_i) = a_i$, for all $i \leq n$. A tuple $\mathbf{a}$ of elements from $\mathsf{ind}(\mathcal{A})$ is a *certain answer* to q over a KB $(\mathcal{T}, \mathcal{A})$, written $\mathcal{T}, \mathcal{A} \models q(\mathbf{a})$, if $\mathcal{I} \models q(\mathbf{a})$ for all models $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$.

A *signature* Σ is a set of concept and role names. For a given signature Σ and a query language $\mathcal{Q}$, we denote with $\mathcal{Q}_\Sigma$ the set of all queries in $\mathcal{Q}$ that use only names from Σ. Given an ABox $\mathcal{A}$, S^+ and S^- denote n-ary relations over $\mathsf{ind}(\mathcal{A})$, called *positive* and *negative examples over* $\mathcal{A}$, resp.

Reasoning Problems. We study the following decision problem for some ontology language $\mathcal{L}$ and query language $\mathcal{Q}$:

Problem:	*Query-by-Example* $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$
Input:	$(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma)$ with $(\mathcal{T}, \mathcal{A})$ an $\mathcal{L}$- KB, S^+ and S^- examples over $\mathcal{A}$ and Σ a signature
Question:	Is there a query $q(\mathbf{x}) \in \mathcal{Q}_\Sigma$ such that • $\mathcal{T}, \mathcal{A} \models q(\mathbf{a})$ for all $\mathbf{a} \in S^+$, and • $\mathcal{T}, \mathcal{A} \not\models q(\mathbf{b})$, for all $\mathbf{b} \in S^-$?

A closely related problem is the *query definability problem* $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$ which takes as input a tuple $(\mathcal{T}, \mathcal{A}, S^+, \Sigma)$ and asks whether there is a query $q(\mathbf{x}) \in \mathcal{Q}_\Sigma$ such that an n-tuple $\mathbf{a}$ is a certain answer if, and only if $\mathbf{a} \in S^+$. If a tuple $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma)$ is a yes-instance of $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$, then we call the query $q(\mathbf{x})$ a *witness*. We further define the variant $\mathsf{QBE}_f(\mathcal{L}, \mathcal{Q})$ (*f* standing for *full*) as the problem of deciding for a given tuple $(\mathcal{T}, \mathcal{A}, S^+, S^-)$ whether $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma^*) \in \mathsf{QBE}(\mathcal{L}, \mathcal{Q})$, where Σ^* is the set of *all* concept and role names occurring in $(\mathcal{T}, \mathcal{A})$; $\mathsf{QDEF}_f(\mathcal{L})$ is defined analogously. Besides the decision problems, we will also be interested in the size of witness queries (if they exist).

We remark that allowing individual names in witness queries might be desirable in some applications, where the user knows some 'special' individuals which are relevant for her query. We show that our choice of forbidding them is without loss of generality. Let $\mathsf{QBE}_c(\mathcal{L}, \mathcal{Q})$ be the variant of $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ that takes another input $I \subseteq \mathsf{ind}(\mathcal{A})$ and allows the witness query to use constants from I. We then have:

Lemma 1. *$\mathsf{QBE}_c(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}_c(\mathcal{L}, \mathcal{Q})$ reduce in polynomial time to $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$, resp., for $\mathcal{Q} \in \{CQ, UCQ\}$, for any $\mathcal{L}$.*

Throughout the paper, we will assume that the input knowledge base $(\mathcal{T}, \mathcal{A})$ is consistent and that S^+ is not empty. Both conditions can be effectively checked and if one of them isn't satisfied the reasoning problems become easier, see appendix.

Moreover, we assume that all TBoxes $\mathcal{T}$ are in *$\mathcal{ELI}_\bot$-normal form*, that is, CIs in $\mathcal{T}$ take one of the following forms:

$$\top \sqsubseteq A \quad A \sqsubseteq \bot \quad A \sqcap A' \sqsubseteq B \quad A \sqsubseteq \exists r.B \quad \exists r.A \sqsubseteq B$$

where A, A', B range over concept names and r ranges over roles. It has been shown that every Horn-$\mathcal{ALCI}$ TBox $\mathcal{T}$ can be transformed in polynomial time to an $\mathcal{ELI}_\bot$ TBox $\mathcal{T}'$ such that $\mathcal{T}'$ is a conservative extension of $\mathcal{T}$ [Bienvenu *et al.*, 2016], and it is easily verified that then $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma) \in \mathsf{QBE}(\text{Horn-}\mathcal{ALCI}, \mathcal{Q})$ iff $(\mathcal{T}', \mathcal{A}, S^+, S^-, \Sigma) \in \mathsf{QBE}(\mathcal{ELI}_\bot, \mathcal{Q})$, for $\mathcal{Q} \in \{\text{CQ}, \text{UCQ}\}$.

3 Model-Theoretic Characterizations

In this section, we provide model-theoretic characterizations of QBE and QDEF, setting the foundations for the development of our decision procedures later on. We presume the standard notion of Σ-*homomorphisms* between interpretations (cf. appendix) and write $\mathcal{I} \to_\Sigma \mathcal{J}$ if there is a homomorphism restricted to the signature Σ from $\mathcal{I}$ to $\mathcal{J}$, and $(\mathcal{I}, \mathbf{a}) \to_\Sigma (\mathcal{J}, \mathbf{b})$ if there is such homomorphism that additionally maps the tuple $\mathbf{a}$ from $\Delta^\mathcal{I}$ to $\mathbf{b}$ from $\Delta^\mathcal{J}$. We drop the Σ in case it comprises all relevant names.

Our characterization is based on the notion of direct products. Let $\mathcal{I}, \mathcal{J}$ be interpretations. The *direct product $\mathcal{I} \otimes \mathcal{J}$ of $\mathcal{I}$ and $\mathcal{J}$* is the interpretation defined by $\Delta^{\mathcal{I}\otimes\mathcal{J}} = \Delta^\mathcal{I} \times \Delta^\mathcal{J}$, $A^{\mathcal{I}\otimes\mathcal{J}} = A^\mathcal{I} \times A^\mathcal{J}$, and

$$r^{\mathcal{I}\otimes\mathcal{J}} = \{((a_1, b_1), (a_2, b_2)) \mid (a_1, a_2) \in r^\mathcal{I}, (b_1, b_2) \in r^\mathcal{J}\},$$

for all concept names A and role names r. The product $(\mathcal{I}, \mathbf{a}) \otimes (\mathcal{J}, \mathbf{b})$ is defined as $(\mathcal{I} \otimes \mathcal{J}, \mathbf{a} \otimes \mathbf{b})$, where $(a_1, \ldots, a_n) \otimes (b_1, \ldots, b_n) = ((a_1, b_1), \ldots, (a_n, b_n))$. Given Σ, a product $\Pi_{i=1}^n (\mathcal{I}_i, \mathbf{a}_i) = (\mathcal{I}_1, \mathbf{a}_1) \otimes \ldots \otimes (\mathcal{I}_n, \mathbf{a}_n)$ is called Σ-*safe* if every element of the tuple $\mathbf{a}_1 \otimes \ldots \otimes \mathbf{a}_n$ appears in the extension of some concept or role name from Σ in $\Pi_{i=1}^n (\mathcal{I}_i, \mathbf{a}_i)$; again, we drop Σ in case it is trivial.

Let us recall the characterization for QBE with CQs over relational databases [ten Cate and Dalmau, 2015; Barceló and Romero, 2017]. For the sake of simplicity, we state it here in our terminology, that is, consider ABoxes instead of databases. Given an ABox $\mathcal{A}$ and sets S^+, S^- of examples over $\mathcal{A}$, there is a CQ distinguishing S^+ and S^- iff

1. $\Pi_{\mathbf{a}\in S^+} (\mathcal{I}_\mathcal{A}, \mathbf{a})$ is safe, and
2. $\Pi_{\mathbf{a}\in S^+} (\mathcal{I}_\mathcal{A}, \mathbf{a}) \not\to (\mathcal{I}_\mathcal{A}, \mathbf{b})$ for every $\mathbf{b} \in S^-$,

where $\mathcal{I}_\mathcal{A}$ is $\mathcal{A}$ viewed as an interpretation. The intuition behind this characterization is as follows: the constructed product can be viewed as CQ with answer variables $\Pi_{\mathbf{a}\in S^+} \mathbf{a}$; in fact, this CQ is the *least general generalization* of the positive examples. Condition 1 ensures that it is a well-defined CQ by requiring all answer variables to actually appear, and Condition 2 ensures that no negative examples are returned.

We argue, however, that this simple characterization does not apply to the case with ontologies. In fact, the example from the introduction does not satisfy Condition 1, but there exists a witness query. We lift the characterization to take into account non-empty TBoxes using *universal interpretations*.

Universal Interpretations. Let $(\mathcal{T}, \mathcal{A})$ be a consistent Horn-$\mathcal{ALCI}$ KB and $\mathcal{T}$ in $\mathcal{ELI}_\bot$-normal form. A *type for $\mathcal{T}$* is a subset t of the concept names in $\mathcal{T}$ such that $\mathcal{T} \models \bigsqcap t \sqsubseteq A$ implies $A \in t$ for all concept names A. When $a \in \mathsf{ind}(\mathcal{A})$, t, t' are types for $\mathcal{T}$, and r is a role, we write

- $a \rightsquigarrow_r^{\mathcal{T},\mathcal{A}} t$ if $\mathcal{T}, \mathcal{A} \models \exists r. \bigsqcap t(a)$ and t is maximal with this condition, and
- $t \rightsquigarrow_r^\mathcal{T} t'$ if $\mathcal{T} \models \bigsqcap t \sqsubseteq \exists r. \bigsqcap t'$ and t' is maximal with this condition.

A *path for $\mathcal{A}$ and $\mathcal{T}$* is a finite sequence $\pi = a r_0 t_1 \cdots t_{n-1} r_{n-1} t_n$, $n \geq 0$, with $a \in \mathsf{ind}(\mathcal{A})$, $r_0, \ldots, r_{n-1}$ roles, and $t_1, \ldots, t_n$ types for $\mathcal{T}$ such that

(i) $a \rightsquigarrow_{r_0}^{\mathcal{T},\mathcal{A}} t_1$ and (ii) $t_i \rightsquigarrow_{r_i}^\mathcal{T} t_{i+1}$ for every $1 \leq i < n$.

We use $\mathsf{tail}(\pi)$ to denote the last element of a path π. Let Paths be the set of all paths for $\mathcal{A}$ and $\mathcal{T}$. The *universal model* $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ of $(\mathcal{T}, \mathcal{A})$ is defined as follows:

$$\begin{aligned}
\Delta^{\mathcal{U}_{\mathcal{T},\mathcal{A}}} &= \mathsf{Paths} \\
A^{\mathcal{U}_{\mathcal{T},\mathcal{A}}} &= \{a \in \mathsf{ind}(\mathcal{A}) \mid \mathcal{T}, \mathcal{A} \models A(a)\} \cup \\
&\quad \{\pi \in \mathsf{Paths} \setminus \mathsf{ind}(\mathcal{A}) \mid A \in \mathsf{tail}(\pi)\} \\
r^{\mathcal{U}_{\mathcal{T},\mathcal{A}}} &= \{(a, b) \in \mathsf{ind}(\mathcal{A})^2 \mid r(a, b) \in \mathcal{A}\} \cup \\
&\quad \{(\pi, \pi r t) \mid \pi r t \in \mathsf{Paths}\} \cup \\
&\quad \{(\pi r^- t, \pi) \mid \pi r^- t \in \mathsf{Paths}\}
\end{aligned}$$

It is well-known that $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ is *universal* in the sense that $\mathcal{T}, \mathcal{A} \models q(\mathbf{a})$ iff $\mathcal{U}_{\mathcal{T},\mathcal{A}} \models q(\mathbf{a})$ for every UCQ $q(\mathbf{x})$ and every tuple $\mathbf{a}$ of individuals [Bienvenu and Ortiz, 2015].

We state now our characterization for $\mathcal{Q} = \text{CQ}$.

Theorem 1. *For every Horn-$\mathcal{ALCI}$ KB $(\mathcal{T}, \mathcal{A})$, all n-ary relations S^+ and S^- over $\mathsf{ind}(\mathcal{A})$, and signatures Σ, we have:*

- $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma) \in \mathsf{QBE}(\textit{Horn-}\mathcal{ALCI}, \textit{CQ})$ *iff*
  1. $\Pi_{\mathbf{a}\in S^+} (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a})$ *is Σ-safe, and*
  2. $\Pi_{\mathbf{a}\in S^+} (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \not\to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ *for all* $\mathbf{b} \in S^-$.
- $(\mathcal{T}, \mathcal{A}, S^+, \Sigma) \in \mathsf{QDEF}(\textit{Horn-}\mathcal{ALCI}, \textit{CQ})$ *iff*
  1.' $\Pi_{\mathbf{a}\in S^+} (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a})$ *is Σ-safe, and*
  2.' $\Pi_{\mathbf{a}\in S^+} (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \not\to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ *for all* $\mathbf{b} \in \mathsf{ind}(\mathcal{A})^n \setminus S^+$.

Thus, the characterization is the same as in the database setting with $\mathcal{I}_\mathcal{A}$ replaced by $\mathcal{U}_{\mathcal{T},\mathcal{A}}$. Note that $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ is possibly infinite, so the product is, in contrast to the database case, *not* the witness. In fact, the proof for direction ($\Leftarrow$) merely shows that *there is* a witness, but in a non-constructive way based on the finite outdegree of $\mathcal{U}_{\mathcal{T},\mathcal{A}}$. Hence, Theorem 1 does not give immediate bounds on the size of witness queries.

In case of UCQs the additional expressive power leaves us with a simpler characterization, the product is compensated for by the use of disjunction in the query language and is thus not necessary anymore.

Theorem 2. *For every Horn-$\mathcal{ALCI}$ KB $(\mathcal{T}, \mathcal{A})$, all n-ary relations S^+ and S^- over $\mathsf{ind}(\mathcal{A})$, and signatures Σ, we have:*

- $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma) \in \mathsf{QBE}(\textit{Horn-}\mathcal{ALCI}, \textit{UCQ})$ *iff* $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a})$ *is Σ-safe and* $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \not\to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$, *for all* $\mathbf{a} \in S^+$ *and* $\mathbf{b} \in S^-$.
- $(\mathcal{T}, \mathcal{A}, S^+, \Sigma) \in \mathsf{QDEF}(\textit{Horn-}\mathcal{ALCI}, \textit{UCQ})$ *iff* $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a})$ *is Σ-safe and* $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \not\to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ *for all* $\mathbf{a} \in S^+$ *and* $\mathbf{b} \in \mathsf{ind}(\mathcal{A})^n \setminus S^+$.

4 Complexity of QBE and QDEF

Based on the characterizations in Theorems 1 and 2, we now pinpoint the precise complexity for the introduced decision problems. We start with observing that Σ-safety (in both theorems) can be checked in exponential time by computing first $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ up to depth 1, computing the product (only in case of Theorem 1), and directly checking the condition.

Lemma 3. *Σ-safety can be decided in* ExpTime.

For Conditions 2 and 2' of Theorem 1 it is sufficient to give an algorithm for deciding $\Pi_{\mathbf{a}\in S^+}(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ for some $\mathbf{b}$; this algorithm can also be used for the homomorphism checks in Theorem 2 by treating the elements $\mathbf{a} \in S^+$ individually. Note that there is no immediate decision procedure, as the involved interpretations $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ are typically infinite. We can, however, exploit regularity of $\mathcal{U}_{\mathcal{T},\mathcal{A}}$.

Let us fix an input $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma)$ with $k = |S^+|$, and denote with $\mathcal{U}^k_{\mathcal{T},\mathcal{A}}$ the product $\Pi^k_{i=1}\mathcal{U}_{\mathcal{T},\mathcal{A}}$. Observe first that $\mathcal{U}^k_{\mathcal{T},\mathcal{A}}$ might be disconnected and that for our purposes it suffices to consider the substructure $\mathcal{P}$ of $\mathcal{U}^k_{\mathcal{T},\mathcal{A}}$ containing all elements from $\mathsf{ind}(\mathcal{A})^k$ and everything that is reachable from there; thus, the domain $\Delta^\mathcal{P}$ of $\mathcal{P}$ is the smallest set such that:

- $\mathsf{ind}(\mathcal{A})^k \subseteq \Delta^\mathcal{P}$, and whenever $\mathbf{p} \in \Delta^\mathcal{P}$ and $(\mathbf{p}, \mathbf{p}') \in r^{\mathcal{U}^k_{\mathcal{T},\mathcal{A}}}$ or $(\mathbf{p}', \mathbf{p}) \in r^{\mathcal{U}^k_{\mathcal{T},\mathcal{A}}}$, then also $\mathbf{p}' \in \Delta^\mathcal{P}$.

It is easy to show that for $\mathbf{a}^* = \Pi_{\mathbf{a}\in S^+}\mathbf{a}$, we have:

Lemma 4. *For every $\mathbf{b} \in S^-$, we have $(\mathcal{U}^k_{\mathcal{T},\mathcal{A}}, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ iff $(\mathcal{P}, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$.*

For what follows, it is convenient to characterize $r^\mathcal{P}$ in terms of (tuples of) *types*, similar to the definition of $\mathcal{U}_{\mathcal{T},\mathcal{A}}$. For doing so, let TP be the set of all types for $\mathcal{T}$ and $\Delta = \mathsf{ind}(\mathcal{A}) \cup \mathsf{TP}$. Then define, for each role r, a binary relation $\hookrightarrow_r^{\mathcal{T},\mathcal{A}}$ on Δ^k by taking $\mathbf{c} \hookrightarrow_r^{\mathcal{T},\mathcal{A}} \mathbf{d}$ iff $\mathbf{c} = (c_1, \dots, c_k)$ and $\mathbf{d} = (d_1, \dots, d_k)$ and for each $1 \le i \le k$ we have:

- if $c_i, d_i \in \mathsf{ind}(\mathcal{A})$, then $r(c_i, d_i) \in \mathcal{A}$ or $r^-(d_i, c_i) \in \mathcal{A}$;
- if $c_i \in \mathsf{ind}(\mathcal{A}), d_i \in \mathsf{TP}$, then $c_i \rightsquigarrow_r^{\mathcal{T},\mathcal{A}} d_i$;
- if $c_i, d_i \in \mathsf{TP}$, then $c_i \rightsquigarrow_r^\mathcal{T} d_i$ or $d_i \rightsquigarrow_{r^-}^\mathcal{T} c_i$.

For $\mathbf{p} = (\pi_1, \dots, \pi_k) \in \Delta^\mathcal{P}$, denote with $\mathsf{tail}(\mathbf{p})$ the tuple $(\mathsf{tail}(\pi_1), \dots, \mathsf{tail}(\pi_n))$. It should be clear that we have

- $(\mathbf{p}, \mathbf{p}') \in r^\mathcal{P}$ iff $\mathsf{tail}(\mathbf{p}) \hookrightarrow_r^{\mathcal{T},\mathcal{A}} \mathsf{tail}(\mathbf{p}')$.

We give a characterization for $(\mathcal{P}, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$, which will be the basis of our decision procedure. Intuitively, we decompose $\mathcal{P}$ into the non-tree-shaped part with domain $N = \mathsf{ind}(\mathcal{A})^k$ and the tree-shaped subinterpretations below each $\mathbf{a} \in N$ (which are characterized alone by their roots, similar to $\mathcal{U}_{\mathcal{T},\mathcal{A}}$). The latter have to be decomposed again because they might be mapped to different parts of $\mathcal{U}_{\mathcal{T},\mathcal{A}}$. We denote with $\mathcal{P}_\mathbf{c}$ for $\mathbf{c} \in \Delta^k$ the sub-interpretation of $\mathcal{P}$ rooted at some $\mathbf{c}$. Moreover, we use the notation $\mathcal{U}_{\mathcal{T},t}$ for a type t for $\mathcal{T}$ as an abbreviation for $\mathcal{U}_{\mathcal{T},\{A(a_t)|A\in t\}}$ and denote with a_t its root. Given some Σ-role r, a tuple $\mathbf{c} \in \Delta^k$, a set $T \subseteq \Delta^k$, and a type $t \in \mathsf{TP}$, we write $(r, \mathbf{c}, T, t) \in \mathsf{PHom}$, for *partial homomorphism*, if there is a partial function $g : \Delta^{\mathcal{P}_\mathbf{c}} \to \Delta^{\mathcal{U}_{\mathcal{T},t}}$ satisfying the following conditions:

- g is a homomorphism on its domain;
- $g(\mathbf{c}) = \pi$ for some $\pi = a_t r t' \in \Delta^{\mathcal{U}_{\mathcal{T},t}}$;
- if $g(\mathbf{p})$ is defined and $(\mathbf{p}, \mathbf{p}') \in s^{\mathcal{P}_\mathbf{c}}$ for a Σ-role s, then either $g(\mathbf{p}) = a_t$ and $\mathsf{tail}(\mathbf{p}) \in T$ or $g(\mathbf{p}')$ is defined.

Intuitively, $(r, \mathbf{c}, T, t)$ belongs to PHom if there is a homomorphism from $\mathcal{P}_\mathbf{c}$ to the subtree rooted at some r-successor of the root a_t of $\mathcal{U}_{\mathcal{T},t}$ given that some parts of $\mathcal{P}_\mathbf{c}$ can be 'delayed' to T when they map to a_t. The component T is necessary because of the 'bidirectional nature' of Horn-$\mathcal{ALCI}$. In general, a homomorphism $(\mathcal{P}, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ does not map subtrees of $\mathcal{P}$ to subtrees in $\mathcal{U}_{\mathcal{T},\mathcal{A}}$, and T is used to synchronize between different subtrees of $\mathcal{U}_{\mathcal{T},\mathcal{A}}$, see the characterization below. For $\mathbf{c} \in \Delta^k, t \in \mathsf{TP}$ we write $\mathbf{c} \to_\Sigma t$ if there is a Σ-homomorphism from an element of type $\mathbf{c}$ to an element of type t. We further denote with $\mathsf{tp}_{\mathcal{U}_{\mathcal{T},\mathcal{A}}}(\pi)$ the type of π in $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ and with $\mathcal{P}|_N$ the restriction of $\mathcal{P}$ to domain N. We establish the following characterization.

Lemma 5. *$(\mathcal{P}, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ iff there is a Σ-homomorphism $h : (\mathcal{P}|_N, \mathbf{a}^*) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ and a labeling $T(\pi) \subseteq \Delta^k$ for every $\pi \in \mathsf{range}(h) \cup \mathsf{ind}(\mathcal{A})$ such that:*

1. *for every $\mathbf{p} \in N$, we have $\mathbf{p} \in T(h(\mathbf{p}))$;*
2. *for every $\mathbf{c} \in T(\pi)$, we have $\mathbf{c} \to_\Sigma \mathsf{tp}_{\mathcal{U}_{\mathcal{T},\mathcal{A}}}(\pi)$;*
3. *for every $\pi \in \mathsf{range}(h) \cup \mathsf{ind}(\mathcal{A})$, every $\mathbf{c} \in T(\pi)$, and every $\mathbf{d}$ with $\mathbf{c} \hookrightarrow_r^{\mathcal{T},\mathcal{A}} \mathbf{d}$ one of the following is true:*
   (a) *there is some $\pi' \in \mathsf{range}(h) \cup \mathsf{ind}(\mathcal{A})$ such that $(\pi, \pi') \in r^{\mathcal{U}_{\mathcal{T},\mathcal{A}}}$ and $\mathbf{d} \in T(\pi')$, or*
   (b) *$(r, \mathbf{d}, T(\pi), \mathsf{tp}_{\mathcal{U}_{\mathcal{T},\mathcal{A}}}(\pi)) \in \mathsf{PHom}$.*

4.1 Horn-$\mathcal{ALCI}$

We now devise a decision procedure for the criterion in Lemma 5. First observe that there are only double exponentially many mappings h and T, since $\mathbf{a}^*$ is forced to be mapped to $\mathbf{b}$, parts disconnected from $\mathbf{a}^*$ can be neglected, and $\mathcal{U}_{\mathcal{T},\mathcal{A}}$ has bounded outdegree. We can thus enumerate all possible such mappings. Conditions 1, 2, and 3(a) can be checked in double exponential time using straightforward algorithms. For Condition 3(b), we devise a mosaic-based decision procedure similar to an algorithm in [Jung *et al.*, 2017].

A mosaic represents the neighborhood of an element in $\Delta^{\mathcal{U}_{\mathcal{T},\mathcal{A}}}$ of some type $t \in \mathsf{TP}$, together with 'types' of elements from $\Delta^\mathcal{P}$ which can be mapped there. Formally, a *mosaic* is a tuple $M = (t, T, r_0, t_0, T_0, \dots, r_n, t_n, T_n)$ with $n \le |\mathcal{T}|$, $t, t_i \in \mathsf{TP}$, $T, T_i \subseteq \Delta^k \setminus N$, and r_i Σ-roles such that:

(i) $t_0 \rightsquigarrow_{r_0} t$ and $t \rightsquigarrow_{r_i} t_i$ for all $1 \le i \le n$;

(ii) if $t \rightsquigarrow_r t'$, there is an $1 \le i \le n$ with $(r_i, t_i) = (r, t')$;

(iii) $\mathbf{t} \in T$ implies $\mathbf{t} \to_\Sigma t$;

(iv) for every $\mathbf{t} \in T$ and every $\mathbf{t} \hookrightarrow_r^{\mathcal{T},\mathcal{A}} \mathbf{t}'$ for some Σ-role r, we either have $r = r_0^-$ and $\mathbf{t}' \in T_0$ or $r = r_i$ and $\mathbf{t}' \in T_i$, for some $1 \le i \le n$.

Intuitively, t_0 is the predecessor type of t and $t_1, \dots, t_n$ are the successors via roles r_i. Condition (iv) ensures that successors of types $\mathbf{t} \in T$ mapped to t can be mapped to either t_0 or some t_i. Given a set $\hat{T} \subseteq \Delta^k \setminus N$, a *root mosaic for $\hat{T}$* is a tuple $M = (t, T, r_0, t_0, T_0, \dots, r_n, t_n, T_n)$ satisfying (i)–(iii)

above, $T_0 = \emptyset$, and the variant (iv') of (iv) which is obtained by replacing '$\mathbf{t} \in T$' with '$\mathbf{t} \in T \setminus \hat{T}$'.

We define a *mosaic elimination* procedure as follows. Define a sequence of sets of mosaics by starting with $\mathfrak{M}_0$ as the set of all mosaics and root mosaics, and obtain $\mathfrak{M}_{i+1}$ from $\mathfrak{M}_i$ by removing all (root) mosaics $M = (t, T, r_0, t_0, T_0, \ldots, r_n, t_n, T_n)$ from $\mathfrak{M}_i$ violating the following compatibility condition:

(E) for every $1 \leq j \leq n$, there is an $M' = (t', T', r'_0, t'_0, T'_0, \ldots, r'_m, t'_m, T'_m) \in \mathfrak{M}_i$ such that $t = t'_0$, $r_j = r'_0$, $t_j = t'$, $T'_0 \subseteq T$, and $T_j \subseteq T'$.

Let $\hat{\mathfrak{M}}$ be where the sequence $\mathfrak{M}_0 \supseteq \mathfrak{M}_1 \supseteq \ldots$ stabilizes.

Lemma 6. *$(\hat{r}, \hat{\mathbf{t}}, \hat{T}, \hat{t}) \in \mathsf{PHom}$ iff $\hat{\mathfrak{M}}$ contains a root mosaic $M = (t, T, r_0, t_0, T_0, \ldots, r_n, t_n, T_n)$ for $\hat{T}$ with $t = \hat{t}$ and a mosaic $M' = (t', T', r'_0, t'_0, T'_0, \ldots, r'_m, t'_m, T'_m)$ with $\hat{\mathbf{t}} \in T'$ and $r_0 = \hat{r}$ such that $t = t'_0$, $r_i = r'_0$, $t_i = t'$, $T'_0 \subseteq T$, and $T_i \subseteq T'$ for some i.*

It remains to discuss the running time of our procedure. The set Δ has size $|\mathcal{A}| + 2^{|\mathcal{T}|}$, thus Δ^k is of size $N_{\mathcal{A},\mathcal{T},k} := (|\mathcal{A}| + 2^{|\mathcal{T}|})^k$. Hence, the number of mosaics is bounded by $2^{N_{\mathcal{A},\mathcal{T},k}}$. In each round of the elimination procedure at least one mosaic is removed, thus the procedure terminates after double exponentially many steps. Finally, the checks in (E) can be implemented in exponential time. We thus conclude:

Corollary 7. *For $\mathcal{L} =$ Horn-$\mathcal{ALCI}$ and $\mathcal{Q} \in \{CQ, UCQ\}$, $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$ are in* 2-ExpTime.

We show next a matching lower bound.

Lemma 8. *For $\mathcal{L} = \mathcal{ELI}$ and $\mathcal{Q} \in \{CQ, UCQ\}$, $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$ are* 2-ExpTime*-hard.*

We reduce the word problem for exponential space bounded alternating Turing machines (ATM) which is 2-ExpTime-hard [Chandra *et al.*, 1981]. Given an ATM M and a word w, we construct a TBox $\mathcal{T}$ such that M accepts w iff $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, a) \to_\Sigma (\mathcal{U}_{\mathcal{T},\mathcal{A}}, b)$ for $\mathcal{A} = \{A(a), B(b)\}$ and some signature Σ. We assume without loss of generality that instead of halting in the accepting state, M enters an infinite loop of special states without changing the tape anymore. We sketch the main idea by describing the universal model.

Below a, $\mathcal{T}$ enforces the infinite tree that is obtained by repeatedly glueing the pattern in Figure 1(a) to its leaves (as indicated by $\circ$; only Σ-symbols depicted). Note that this pattern (without the outgoing path labeled with α_0, α_1) is in fact the basic one of a computation tree of an ATM: a universal configuration of length 2^n (labeled with U) followed by a branch into two existential configurations of the same length (labeled with E_1, E_2). We call this the *skeleton tree*. Apart from the skeleton tree, for *every* possible choice of α_0, α_1, a path of the shape depicted in the right starts from *every* node of the tree. There, α_0 and α_1 range over all possible triples containing the content of three consecutive tape cells, e.g., $\langle a, b, c\rangle$ or $\langle a, (q, b), c\rangle$.

Below b, $\mathcal{T}$ enforces an infinite tree as illustrated in Figure 1(b), having the following properties:

- It starts with a path of length 2^n labeled with the initial configuration, encoded using triples.

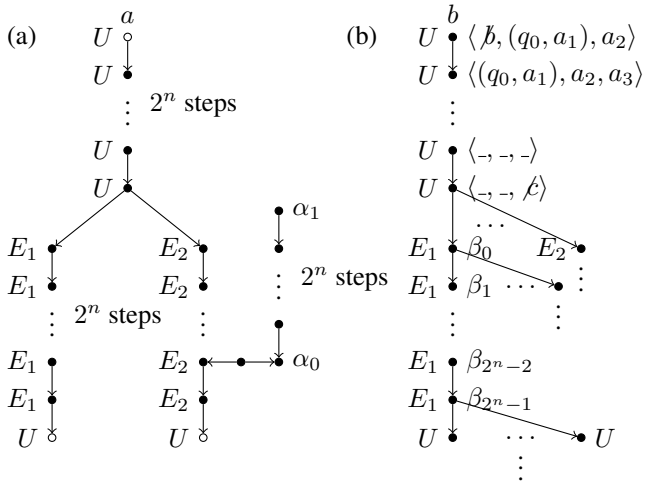


Figure 1: Parts of the universal model enforced in Lemma 8.

- All other nodes are labeled with a pair $\beta = (\alpha, \alpha')$ with α, α' triples as described above. In this case, α (resp., α') is intended to represent the content of the tape cell in the current (resp., previous) configuration.
- Every path of length 2^n (between E_x and U or U and E_x), e.g., $\beta_0, \ldots, \beta_{2^n-1}$ in Figure 1, corresponds to the description of a valid configuration of M, and a possible predecessor configuration.
- This is continued infinitely, always switching between universal and existential configurations, as depicted.

It is instructive to consider some homomorphism h_0 of the skeleton tree below a into the tree below b. Informally, h_0 can be thought of as a labeling of the skeleton (and thus of the computation tree) with actual configurations. It remains to ensure that the transition relation of M is obeyed, which is done as follows. Every node in the tree below b has also outgoing paths of the same shape as the one depicted in the left side (for the sake of clarity and space they are not depicted in Fig. 1). However, it has only such paths for every α, α' except the label $\beta = \alpha_0, \alpha_1$ at the current node. Let now v be a node in the skeleton tree and assume its image $v' = h_0(v)$ has label $\beta = (\alpha, \alpha')$. Clearly, h_0 can be extended for all outgoing paths except the one labeled with $\alpha_0 = \alpha$ and $\alpha_1 = \alpha'$. Additionally, by construction, the end of this path can only be mapped to the corresponding cell in the previous configuration. The homomorphism condition ensures that the computation tree obeys the transition relation.

Summarizing, from Lemma 8 and Corollary 7 we obtain:

Theorem 9. *For $\mathcal{L} =$ Horn-$\mathcal{ALCI}$ and $\mathcal{Q} \in \{CQ, UCQ\}$, $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$ are* 2-ExpTime*-complete.*

We remark that the hardness proof crucially relies on Σ. For CQs, it is adapted to the unrestricted signature case by adding an assertion $A'(a')$ to the ABox, and enforcing below a' a tree identical (up to Σ-homomorphisms) to the tree below a, by using fresh copies of non-Σ concept names. The product of the trees below a and a' is then as in the proof of Lemma 8.

This approach does not apply to UCQs. In fact, the problem becomes easier with unrestricted signature. To see the reason

for this complexity drop, note that, for a TBox $\mathcal{T}$ in normal form, we have $(\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{a}) \to (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$ iff $(\mathcal{U}_{\mathcal{T},\mathcal{A}}|_{\mathsf{ind}(\mathcal{A})}, \mathbf{a}) \to (\mathcal{U}_{\mathcal{T},\mathcal{A}}, \mathbf{b})$. This can be straightforwardly decided in exponential time, see the appendix.

Theorem 10. *For $\mathcal{L}$ = Horn-$\mathcal{ALCI}$, $\mathsf{QBE}_f(\mathcal{L}, UCQ)$ and $\mathsf{QDEF}_f(\mathcal{L}, UCQ)$ are* ExpTime*-complete.*

4.2 Horn-$\mathcal{ALC}$

For Horn-$\mathcal{ALC}$, note that the characterizations in Theorems 1 and 2 and Lemma 5 are still valid since Horn-$\mathcal{ALC}$ is a fragment of Horn-$\mathcal{ALCI}$. However, the absence of inverse roles simplifies the decision procedure of PHom which is the bottleneck for Horn-$\mathcal{ALCI}$. Indeed, both $\mathcal{P}_\mathbf{c}$ and the anonymous parts of $\mathcal{U}_{\mathcal{T},t}$ are directed regular trees where all roles point away from the root, so the set T can be ignored and we can decide PHom by applying standard techniques for regular trees.

Lemma 11. *Given $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma)$ with $(\mathcal{T}, \mathcal{A})$ a Horn-$\mathcal{ALC}$ KB, the relation* PHom *can be decided in* ExpTime*.*

Applying this Lemma, we obtain a coNExpTime upper bound for deciding QBE from the algorithm devised in the previous section. A matching lower bound is inherited from the database setting [ten Cate and Dalmau, 2015]. For UCQs, a careful analysis of Lemma 5 yields an ExpTime upper bound; the matching lower bound is obtained by a reduction from subsumption in Horn-$\mathcal{ALC}$ [Krötzsch *et al.*, 2013].

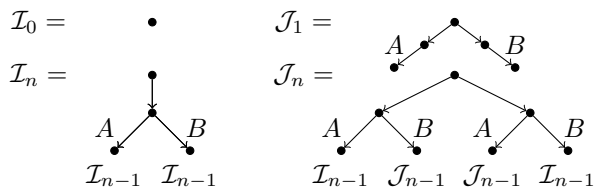
Theorem 12. *For $\mathcal{L}$ = Horn-$\mathcal{ALC}$, $\mathsf{QBE}(\mathcal{L}, \mathcal{Q})$ and $\mathsf{QDEF}(\mathcal{L}, \mathcal{Q})$ are* coNExpTime*-complete if $\mathcal{Q} = CQ$ and* ExpTime*-complete if $\mathcal{Q} = UCQ$. All results also hold with unrestricted signature.*

5 Size of Witness Queries

We finally investigate the size of witness queries. We first establish the following double exponential lower bound.

Lemma 13. *There is a family of Horn-$\mathcal{ALC}$ knowledge bases $(\mathcal{T}_n, \mathcal{A}_n)_{n\geq 1}$, sets of examples S^+ and S^-, a signature Σ, and a polynomial $p(n)$ such that, for all $n \geq 1$, $|\mathcal{T}_n \cup \mathcal{A}_n| \leq p(n)$, $(\mathcal{T}_n, \mathcal{A}_n, S^+, S^-, \Sigma) \in \mathsf{QBE}$(Horn-$\mathcal{ALC}$, (U)CQ) and every (U)CQ witnessing this is of size $\Omega(2^{2^n})$.*

The main idea for the lower bound is to give Horn-$\mathcal{ALC}$ knowledge bases $(\mathcal{T}_n, \mathcal{A}_n)$ over two individuals a, b such that in $\mathcal{U}_{\mathcal{T}_n,\mathcal{A}_n}$ the trees below a and b are Σ-homomorphically equivalent to $\mathcal{I}_{2^n}$ and $\mathcal{J}_{2^n}$, respectively, where $\mathcal{I}_n, \mathcal{J}_n$ are given by the following recursive 'definitions':



It can be shown that $(\mathcal{I}_n, a) \not\to_\Sigma (\mathcal{J}_n, b)$, but that $(\mathcal{I}', a) \to_\Sigma (\mathcal{J}_n, b)$ for any sub-interpretation $\mathcal{I}'$ of $\mathcal{I}_n$. Thus, the smallest Σ-(U)CQ distinguishing between a and b in $(\mathcal{T}_n, \mathcal{A}_n)$ is $(\mathcal{I}_{2^n}, a)$ viewed as CQ, whose size is $\Omega(2^{2^n})$.

For Horn-$\mathcal{ALC}$, a matching upper bound is obtained by an analysis of Lemma 5 and the observations made in the previous Section. For Horn-$\mathcal{ALCI}$, we obtain a four-fold exponential upper bound on the size of the witness query by viewing the check for PHom as a reachability game on pushdown systems and apply known results from there [Kupferman *et al.*, 2010; Carayol and Hague, 2014]. We leave the exact sizes for future work.

Theorem 14. *If $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma) \in \mathsf{QBE}(\mathcal{L}, CQ)$, there is a witness query of at most double (resp., four-fold) exponential size if $\mathcal{L}$ = Horn-$\mathcal{ALC}$ (resp., $\mathcal{L}$ = Horn-$\mathcal{ALCI}$).*

6 Discussion and Future Work

Our investigation opens a new whole research avenue towards improving the usability of ontology-enriched systems. From the theoretical perspective, the most natural next step is to broaden our understanding to different ontology and query languages. Given the state of the art of OES, we are particularly interested in 'lightweight' DLs, such as *DL-Lite* and $\mathcal{EL}$; our results already provide a solid basis for these logics. For non-Horn or Datalog$^\pm$ ontologies it will be more challenging – a good starting point for non-Horn DLs might be [Botoeva *et al.*, 2016b]. As for the query language, we will study regular path queries. From the practical perspective, it suggests itself to develop systems for QBE over KBs which not only implement reverse-engineering algorithms, but also allow interaction with the user, as done e.g., in [Bonifati *et al.*, 2014; Diaz *et al.*, 2016]. Given the high complexity of QBE, it will be also important to design heuristics [Tran *et al.*, 2014; Mottin *et al.*, 2016] or approximations [Barceló and Romero, 2017], as for relational databases. We note that some approximations considered by Barceló and Romero [2017] do not directly lead to better complexity in the context of OES. For example, 2-ExpTime-hardness in Lemma 8 already holds for *tree-shaped* CQs. Another possible approximation is bounding the size of the witness queries.

Related within DL research is the study of *query conservative extensions (QCE)*, where the question is whether two given ontologies or two knowledge bases can be distinguished by a query (without providing examples). Indeed, in the context of QCE, characterizations based on homomorphisms and universal models have been devised and inverse roles also tend to increase the complexity, see [Botoeva *et al.*, 2016a] for a recent survey, and references therein. We are, however, not aware of any direct reductions between QBE and QCE.

Within the broader context of machine learning, we believe that our results lay the foundations for questions related to *learnability* of queries, see [Cohen and Page, 1995] for an overview. In this line, one could investigate an ILP inspired variant: if an instance $(\mathcal{T}, \mathcal{A}, S^+, S^-, \Sigma)$ of QBE does *not* have a witness, is there an extension $\mathcal{T}' \supseteq \mathcal{T}$ such that there is a witness? In the context of active learning, one would be interested in learning a (conjunctive) query with membership and/or equivalence queries over a DL knowledge base. Finally, it would be interesting to extend the recently introduced framework of learning concepts over background structures of small degree and having only *local access* to the data [Grohe and Ritzert, 2017] with an ontology.

Acknowledgments

The authors were funded by EU's Horizon 2020 programme under the Marie Skłodowska-Curie grant 663830, the Research Foundation - Flanders project G.0428.15, and ERC consolidator grant 647289 CODA, respectively.

References

[Arenas *et al.*, 2016] Marcelo Arenas, Gonzalo I. Diaz, and Egor V. Kostylev. Reverse engineering SPARQL queries. In *Proc. of WWW-16*, pages 239–249, 2016.

[Baader *et al.*, 2017] Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. *An Introduction to Description Logics*. Cambride University Press, 2017.

[Barceló and Romero, 2017] Pablo Barceló and Miguel Romero. The complexity of reverse engineering problems for conjunctive queries. In *Proc. of ICDT-17*, 2017.

[Bienvenu and Ortiz, 2015] Meghyn Bienvenu and Magdalena Ortiz. Ontology-mediated query answering with data-tractable description logics. In *Proc. of RW-15*, 2015.

[Bienvenu *et al.*, 2016] Meghyn Bienvenu, Peter Hansen, Carsten Lutz, and Frank Wolter. First order-rewritability and containment of conjunctive queries in Horn description logics. In *Proc. of IJCAI-16*, pages 965–971, 2016.

[Bonifati *et al.*, 2014] Angela Bonifati, Radu Ciucanu, and Slawek Staworko. Interactive inference of join queries. In *Proc. of EDBT-14*, pages 451–462, 2014.

[Bonifati *et al.*, 2015] Angela Bonifati, Radu Ciucanu, and Aurélien Lemay. Learning path queries on graph databases. In *Proc. EDBT-15*, pages 109–120, 2015.

[Bonifati *et al.*, 2016] Angela Bonifati, Radu Ciucanu, and Slawek Staworko. Learning join queries from user examples. *ACM Trans. Database Syst.*, 40(4):24:1–24:38, 2016.

[Botoeva *et al.*, 2016a] Elena Botoeva, Boris Konev, Carsten Lutz, Vladislav Ryzhikov, Frank Wolter, and Michael Zakharyaschev. Inseparability and conservative extensions of description logic ontologies: A survey. In *Proc. of RW-16*, 2016.

[Botoeva *et al.*, 2016b] Elena Botoeva, Carsten Lutz, Vladislav Ryzhikov, Frank Wolter, and Michael Zakharyaschev. Query-based entailment and inseparability for ALC ontologies. In *Proc. of IJCAI-16*, 2016.

[Calvanese *et al.*, 2016] Diego Calvanese, Pietro Liuzzo, Alessandro Mosca, José Remesal, Martin Rezk, and Guillem Rull. Ontology-based data integration in epnet: Production and distribution of food during the roman empire. *Eng. Appl. of AI*, 51:212–229, 2016.

[Carayol and Hague, 2014] Arnaud Carayol and Matthew Hague. Regular strategies in pushdown reachability games. In *Proc. of Reachability Problems, RP-14*, 2014.

[Chandra *et al.*, 1981] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *J. ACM*, 28(1):114–133, January 1981.

[Cohen and Page, 1995] William W. Cohen and C. David Page. Polynomial learnability and inductive logic programming: Methods and results. *New Generation Comput.*, 13(3&4):369–409, 1995.

[Diaz *et al.*, 2016] Gonzalo I. Diaz, Marcelo Arenas, and Michael Benedikt. Sparqlbye: Querying RDF data by example. *PVLDB*, 9(13):1533–1536, 2016.

[Grohe and Ritzert, 2017] Martin Grohe and Martin Ritzert. Learning first-order definable concepts over structures of small degree. In *Proc. of LICS-17*, pages 1–12, 2017.

[Hovland *et al.*, 2017] Dag Hovland, Roman Kontchakov, Martin G. Skjæveland, Arild Waaler, and Michael Zakharyaschev. Ontology-based data access to Slegge. In *Proc. of ISWC-17*, pages 120–129, 2017.

[Jung *et al.*, 2017] Jean Christoph Jung, Carsten Lutz, Mauricio Martel, and Thomas Schneider. Query conservative extensions in horn description logics with inverse roles. In *Proc. of IJCAI-17*, 2017.

[Kharlamov *et al.*, 2015] Evgeny Kharlamov, Dag Hovland, Ernesto Jiménez-Ruiz, Davide Lanti, Hallstein Lie, Christoph Pinkel, Martín Rezk, Martin G. Skjæveland, Evgenij Thorstensen, Guohui Xiao, Dmitriy Zheleznyakov, and Ian Horrocks. Ontology based access to exploration data at Statoil. In *Proc. of ISWC-15*, pages 93–112, 2015.

[Kietz, 2002] Jörg-Uwe Kietz. Learnability of description logic programs. In *Inductive Logic Programming, 12th International Conference, ILP 2002, Sydney, Australia, July 9-11, 2002. Revised Papers*, pages 117–132, 2002.

[Krötzsch *et al.*, 2013] Markus Krötzsch, Sebastian Rudolph, and Pascal Hitzler. Complexities of horn description logics. *ACM Trans. Comput. Logic*, 14(1):2:1–2:36, 2013.

[Kupferman *et al.*, 2010] Orna Kupferman, Nir Piterman, and Moshe Y. Vardi. An automata-theoretic approach to infinite-state systems. In *Time for Verification, Essays in Memory of Amir Pnueli*, pages 202–259, 2010.

[Mottin *et al.*, 2016] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. Exemplar queries: a new way of searching. *VLDB J.*, 25(6):741–765, 2016.

[Nienhuys-Cheng and de Wolf, 1997] Shan-Hwei Nienhuys-Cheng and Ronald de Wolf, editors. *Foundations of Inductive Logic Programming*, volume 1228 of *Lecture Notes in Computer Science*. Springer, 1997.

[Rodriguez-Muro *et al.*, 2013] Mariano Rodriguez-Muro, Roman Kontchakov, and Michael Zakharyaschev. Ontology-based data access: Ontop of databases. In *Proc. of ISWC-13*, pages 558–573, 2013.

[ten Cate and Dalmau, 2015] Balder ten Cate and Víctor Dalmau. The product homomorphism problem and applications. In *Proc. of ICDT-15*, pages 161–176, 2015.

[Tran *et al.*, 2014] Quoc Trung Tran, Chee Yong Chan, and Srinivasan Parthasarathy. Query reverse engineering. *VLDB J.*, 23(5):721–746, 2014.

[Zloof, 1975] Moshé M. Zloof. Query-by-example: the invocation and definition of tables and forms. In *Proc. of VLDB-75*, pages 1–24, 1975.