Query Containment in Description Logics Reconsidered

Meghyn Bienvenu
Laboratoire de Recherche en Informatique
CNRS & Université Paris Sud, France

Carsten Lutz
Fachbereich Informatik
Universität Bremen, Germany

Frank Wolter
Department of Computer Science
University of Liverpool, UK

Abstract

While query answering in the presence of description logic (DL) ontologies is a well-studied problem, questions of static analysis such as query containment and query optimization have received less attention. In this paper, we study a rather general version of query containment that, unlike the classical version, cannot be reduced to query answering. First, we allow a restriction to be placed on the vocabulary used in the instance data, which can result in shorter equivalent queries; and second, we allow each query its own ontology rather than assuming a single ontology for both queries, which is crucial in applications to versioning and modularity. We also study global minimization of queries in the presence of DL ontologies, which is more subtle than for classical databases as minimal queries need not be isomorphic.

1 Introduction

In ontology-based data access (OBDA), an ontology is used to improve query answering over instance data in various ways, for example by providing a semantics for the data vocabulary, by enriching the query vocabulary, and by translating between data and query vocabularies when they diverge. In the past decade, this paradigm has received significant attention, with a focus on using description logics (DLs) as the ontology language. In particular, answering conjunctive queries (CQs) and the simpler instance queries (IQs) in DL-based OBDA has been extensively studied, so that today, various algorithmic approaches are known, and the computational complexity is well-understood. On the one hand, for expressive DLs such as $\mathcal{ALC}$ and $\mathcal{SHIQ}$, CQ answering is typically EXPTIME- or 2-EXPTIME-complete for combined complexity and NP-complete for data complexity (Calvanese, De Giacomo, and Lenzerini 1998; Hustadt, Motik, and Sattler 2005; Glimm et al. 2008; Lutz 2008; Ortiz, Rudolph, and Simkus 2011). On the other hand, for so-called lightweight DLs such as DL-Lite and $\mathcal{EL}$, CQ answering is in PTIME for data complexity and can be implemented efficiently using off-the-shelf relational database systems (Calvanese et al. 2007; Lutz, Toman, and Wolter 2009; Pérez-Urbina, Horrocks, and Motik 2009; Kontchakov et al. 2010; Calvanese et al. 2011).

While query answering in DLs has been studied intensively, little attention has been paid to the query containment problem, which consists in deciding, given a DL ontology (TBox) $\mathcal{T}$ and two queries q_1 and q_2 of same arity, whether for every data instance (ABox), the answers to q_1 given $\mathcal{T}$ are a subset of the answers to q_2 given $\mathcal{T}$. This is in contrast to relational databases, where query containment is a crucial and widely studied problem due to the central role it plays in query optimization (Abiteboul, Hull, and Vianu 1995). In particular, Chandra and Merlin observed in a classical paper that minimal CQs are unique up to isomorphism, which also means that the unique minimal CQ for a given CQ q can be produced by the following simple procedure: start with q and repeatedly remove atoms that are redundant in the sense that dropping them preserves equivalence; the order in which atoms are dropped is irrelevant and the only non-trivial part is checking equivalence, implemented as two query containment checks (Chandra and Merlin 1977).

Clearly, query optimization is important also in OBDA. For example, in the combined approach to CQ answering presented in (Lutz, Toman, and Wolter 2009; Kontchakov et al. 2010), the CQ is passed virtually unchanged to a relational database system for execution, and thus prior optimization improves performance. The relative lack of interest in OBDA query containment is somewhat surprising and seems to stem mainly from the fact that, for most query languages including CQs and IQs, the problem can be polynomially reduced to query answering and vice versa; thus, algorithms and complexity results transfer (a notable exception are regular path queries, whose containment problem was recently studied in a DL context in (Calvanese, Ortiz, and Simkus 2011)). The aim of this paper is to reconsider CQ- and IQ-containment in DL-based OBDA by (i) proposing a generalized version of containment that enables novel applications and cannot be polynomially reduced to query answering, (ii) giving algorithms and complexity results for this problem, with a focus on lightweight DLs of the DL-Lite and $\mathcal{EL}$ families, and (iii) showing that while naive Chandra-Merlin-minimization as described above fails in the presence of ontologies, by applying slightly refined strategies one can still achieve strong guarantees for the produced minimal queries.

Regarding (i), we generalize OBDA query containment in two directions. First, we pick up the observation of (Baader et al. 2010) that, when an ontology is used to enrich the query vocabulary with symbols that do not occur in the data, then it is useful to carefully distinguish between the data vocabulary and the ontology/query vocabulary. Specifically, we use the data vocabulary Σ as an additional input to query containment, which is then refined to quantify only over Σ-ABoxes. While this sometimes increases the complexity of containment, it can lead to signficantly smaller queries in query minimization.

The second generalization is to associate a separate ontology $\mathcal{T}_i$ with each query q_i instead of assuming a single ontology $\mathcal{T}$ for both queries q_1 and q_2. This is natural from a traditional database perspective, where the ontology would likely be viewed as a component of the query rather than as an independent object. It also enables applications to *ontology versioning* and *ontology modules*. In the former, a typical scenario is that a new version $\mathcal{T}_{\mathsf{new}}$ of a reference ontology $\mathcal{T}_{\mathsf{ref}}$ has to be adopted in an existing application; for example, $\mathcal{T}_{\mathsf{ref}}$ could be the medical terminology SNOMED CT or the National Cancer Institute Ontology NCI (IHTSDO 2008; Sioutos et al. 2006), both widely used and frequently updated. To verify that the update does not affect the application, the user wants to check, for each relevant query q, whether q given $\mathcal{T}_{\mathsf{new}}$ is equivalent to q given $\mathcal{T}_{\mathsf{ref}}$ (see Section 2 for more details). Applications to ontology modules are in a similar spirit: assume that a large ontology $\mathcal{T}$ is replaced with a smaller module $\mathcal{T}' \subseteq \mathcal{T}$ to speed up query processing. When $\mathcal{T}'$ was generated manually or using a technique that does not guarantee preservation of query answers, then the user wants to check for each relevant query q, whether q given $\mathcal{T}$ is equivalent to q given $\mathcal{T}'$.

Regarding (ii), we consider the complexity of generalized query containment both for CQs and IQs, and for a variety of DLs from the DL-Lite-, $\mathcal{EL}$-, and $\mathcal{ALC}$-families. An interesting first observation is that the two proposed extensions of query containment are intimately related. In fact, containment with an ABox signature Σ *and* two TBoxes can, in most cases, be polynomially reduced to containment with an ABox signature but only one TBox, and to containment without an ABox signature but with two TBoxes. Another relevant observation is that query emptiness, as studied in (Baader et al. 2010), is a special case of our version of query containment, and thus lower complexity bounds carry over. In particular, this means that containment is undecidable in $\mathcal{ALCF}$, the extension of $\mathcal{ALC}$ with functional roles, both for IQs and CQs. For weaker DLs, we exhibit a rich complexity landscape that ranges from PTIME (for IQ-containment in DL-Lite$_{\mathsf{core}}$ and IQ-containment w.r.t. acyclic $\mathcal{EL}$-TBoxes) via Π_2^p-completeness (for CQ-containment in DL-Lite$_{\mathsf{core}}$ and DL-Lite$_{\mathsf{horn}}$) and PSPACE-completeness (for CQ-containment w.r.t. acyclic $\mathcal{EL}$-TBoxes) to EXPTIME-completeness (for IQ-containment and CQ-containment w.r.t. general $\mathcal{EL}$-TBoxes and $\mathcal{EL}_\bot$-TBoxes). We also show decidability of IQ-containment in $\mathcal{ALC}$, with a $\mathrm{P}^{\mathrm{NExp}}$ upper bound. The precise complexity remains open, and so does the decidability of CQ-containment in $\mathcal{ALC}$.

Regarding (iii), we develop strategies for minimizing queries in the presence of ontologies formulated in DL-Lite and $\mathcal{EL}$. We show that, by adopting a suitable minimization strategy, the uniqueness of minimal queries can be regained in DL-Lite, and the resulting queries have an optimal relational structure in the sense that this structure can be found as a subquery in any equivalent query. For $\mathcal{EL}$, we show how to produce an equivalent acyclic query whenever it exists. In this part, we work with the classical notion of query containment instead of with the generalized one.

Throughout the paper, we mostly confine ourselves to proof sketches and (without further notice) defer full proof details to the long version, which is made available at http://www.informatik.uni-bremen.de/∼clu/papers/

2 Preliminaries

We use standard notation for the syntax and semantics of DLs, please see (Baader et al. 2003) for details. As usual, $\mathsf{N_C}$, $\mathsf{N_R}$, and $\mathsf{N_I}$ denote countably infinite sets of concept names, role names, and individual names, C, D denote (potentially) composite concepts, A, B concept names, r, s role names, and a, b individual names. We consider the following three families of DLs.

The DL-Lite family. The basic member is DL-Lite$_{\mathsf{core}}$, where *TBoxes* are finite sets of *concept inclusions (CIs)* of the forms

$$B_1 \sqsubseteq B_2 \quad \text{and} \quad B_1 \sqcap B_2 \sqsubseteq \bot$$

with B_1 and B_2 concepts of the form $\exists r$, $\exists r^-$, $\top$, $\bot$, or A. In the extension DL-Lite$_{\mathsf{horn}}$, we additionally allow conjunction, thus obtaining CIs of the forms

$$B_1 \sqcap \cdots \sqcap B_n \sqsubseteq B \quad \text{and} \quad B_1 \sqcap \cdots \sqcap B_n \sqsubseteq \bot$$

cf. (Calvanese et al. 2007; Artale et al. 2009).

The $\mathcal{EL}$-family. Its basic member $\mathcal{EL}$ offers the concept constructors $\top$, $C \sqcap D$, and $\exists r.C$. The extension of $\mathcal{EL}$ with the bottom concept $\bot$ is denoted $\mathcal{EL}_\bot$. In both cases, a *TBox* is a finite set of CIs $C \sqsubseteq D$ with C and D (potentially) compound concepts. We use *concept definitions* $A \equiv C$ in TBoxes as abbreviations for two CIs $A \sqsubseteq C$ and $C \sqsubseteq A$. See for example (Baader, Brandt, and Lutz 2005) for more information on the $\mathcal{EL}$-family of DLs.

The $\mathcal{ALC}$-family. The basic member $\mathcal{ALC}$ offers the concept constructors $\neg C, C \sqcap D$, and $\exists r.C$. $\mathcal{ALCI}$ is the extension of $\mathcal{ALC}$ with the $\exists r^-.C$ constructor, where r^- denotes an *inverse role*, and $\mathcal{ALCF}$ the extension with functional roles. Sometimes we mention $\mathcal{ALCFI}$, which is the union of $\mathcal{ALCI}$ and $\mathcal{ALCF}$ and contains all DLs studied in this paper as a fragment. In $\mathcal{ALC}$ and its extensions, a *TBox* is again a finite set of CIs $C \sqsubseteq D$. See (Baader et al. 2003) for more details.

In any of these DLs, data is stored in an *ABox*, which is a finite set of *concept assertions* $A(a)$ and *role assertions* $r(a, b)$. We use $\mathsf{Ind}(\mathcal{A})$ to denote the set of individual names used in the ABox $\mathcal{A}$.

The semantics of DLs is based on interpretations $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ as usual, see (Baader et al. 2003). An interpretation is a *model* of a TBox $\mathcal{T}$ (resp. ABox $\mathcal{A}$) if it satisfies all concept inclusions in $\mathcal{T}$ (resp. assertions in $\mathcal{A}$), where satisfac-

tion is defined in the standard way. An ABox $\mathcal{A}$ is *consistent* w.r.t. a TBox $\mathcal{T}$ if $\mathcal{A}$ and $\mathcal{T}$ have a common model.

Instance queries (IQs) take the form $A(x)$ and *conjunctive queries (CQs)* take the form $\exists\vec{x}.\varphi(\vec{x},\vec{y})$, where $x,\vec{x},\vec{y}$ denote (tuples of) variables taken from a set $\mathsf{N_V}$ and φ is a conjunction of atoms of the form $A(t)$ and $r(t,t')$ with t,t' *terms*, i.e., individual names or variables from $\vec{x}\cup\vec{y}$. We call the variables in $\vec{y}$ the *answer variables* and those in $\vec{x}$ the *quantified* variables. The *arity* of a CQ is the number of answer variables and $\mathsf{term}(q)$ denotes the set of terms used in q. In what follows, we sometimes slightly abuse notation and use CQ to denote the class of all conjunctive queries and likewise for IQ. Whenever convenient, we treat a CQ (and even an IQ) as a *set* of atoms.

Let $\mathcal{I}$ be an interpretation and q an (instance or conjunctive) query with answer variables $x_1,\dots,x_k$. For $\vec{a} = a_1,\dots,a_k \in \mathsf{N_I}$, an *$\vec{a}$-match* for q in $\mathcal{I}$ is a mapping $\pi : \mathsf{term}(q) \to \Delta^{\mathcal{I}}$ such that $\pi(x_i) = a_i^{\mathcal{I}}$ for $1 \le i \le k$, $\pi(a) = a^{\mathcal{I}}$ for all $a \in \mathsf{term}(q)\cap\mathsf{N_I}$, $\pi(t) \in A^{\mathcal{I}}$ for all $A(t) \in q$, and $(\pi(t_1),\pi(t_2)) \in r^{\mathcal{I}}$ for all $r(t_1,t_2) \in q$. We write $\mathcal{I} \models q[\vec{a}]$ if there is an $\vec{a}$-match of q in $\mathcal{I}$. For a TBox $\mathcal{T}$ and an ABox $\mathcal{A}$, we write $\mathcal{T},\mathcal{A} \models q[\vec{a}]$ if $\mathcal{I} \models q[\vec{a}]$ for all models $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$. In this case, $\vec{a}$ is a *certain answer* to q w.r.t. $\mathcal{A}$ and $\mathcal{T}$. We use $\mathsf{cert}_{\mathcal{T}}(q,\mathcal{A})$ to denote the set of all certain answers to q w.r.t. $\mathcal{A}$ and $\mathcal{T}$.

We use the term *predicate* to refer to a concept name or role name and *signature* to refer to a set of predicates. Then $\mathsf{sig}(q)$ denotes the set of predicates used in the query q, and similarly $\mathsf{sig}(\mathcal{T})$ (resp. $\mathsf{sig}(\mathcal{A})$) refers to the signature of a TBox $\mathcal{T}$ (resp. ABox $\mathcal{A}$). Given a signature Σ, a *Σ-ABox* is an ABox using predicates from Σ only.

In the context of query answering in DLs, it is sometimes useful to adopt the unique name assumption (UNA), which requires that $a^{\mathcal{I}} \neq b^{\mathcal{I}}$ for all interpretations $\mathcal{I}$ and all $a,b \in \mathsf{N_I}$ with $a \neq b$. The results obtained in this paper do not depend on the UNA. In fact, it is well-known that in all DLs studied here (with the exception of $\mathcal{ALCF}$), query answers with and without UNA coincide.

The following definition provides the general perspective on query containment in the presence of DL TBoxes that we propose in this paper.

Definition 1. Let $\mathcal{T}_1,\mathcal{T}_2$ be TBoxes, q_1,q_2 CQs with the same arity, and Σ an ABox signature. Then $(\mathcal{T}_1,q_1)$ *is contained in* $(\mathcal{T}_2,q_2)$ *w.r.t.* Σ, written $(\mathcal{T}_1,q_1) \subseteq_\Sigma (\mathcal{T}_2,q_2)$, if for all Σ-ABoxes $\mathcal{A}$ that are consistent w.r.t. $\mathcal{T}_1$ and $\mathcal{T}_2$, we have $\mathsf{cert}_{\mathcal{T}_1}(q_1,\mathcal{A}) \subseteq \mathsf{cert}_{\mathcal{T}_2}(q_2,\mathcal{A})$.

As discussed in the introduction, this definition generalizes the traditional view of query containment in DLs in two directions: by admitting two distinct TBoxes for the two queries and by allowing a restriction to be placed on the ABox signature. If there is only a single TBox $\mathcal{T}$, we write $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$ instead of $(\mathcal{T},q_1) \subseteq_\Sigma (\mathcal{T},q_2)$. When $\Sigma = \mathsf{N_R}\cup\mathsf{N_C}$, we say that the ABox signature Σ is *full* and simply omit it in the subscript of "$\subseteq$". We say that q_1 *and* q_2 *are equivalent w.r.t.* Σ *and* $\mathcal{T}$, written $q_1 \equiv_{\mathcal{T},\Sigma} q_2$, if $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$ and $q_2 \subseteq_{\mathcal{T},\Sigma} q_1$. Again, Σ is omitted if it is full.

To illustrate Definition 1, consider the following TBox $\mathcal{T}$, a slightly simplified fragment of the SNOMED CT ontology:

$$\begin{aligned}
\mathsf{LabTest} &\sqsubseteq \mathsf{LabProc}\sqcap\mathsf{EvalProc}\\
\mathsf{VenipunctBloodTest} &\sqsubseteq \exists\mathsf{focus}.\mathsf{LabTest}\\
\mathsf{VenipunctBloodTest} &\equiv \exists\mathsf{purpose}.\mathsf{BloodSmpl}\sqcap\\
&\qquad\mathsf{Venipuncture}
\end{aligned}$$

Assume that ABoxes provide data about venipunctures and their purpose, thus the ABox signature Σ contains the concept names $\mathsf{Venipuncture}$ and $\mathsf{BloodSmpl}$ and the role name $\mathsf{purpose}$, but no other symbols from $\mathcal{T}$. Let

$$q(x) = \exists y.(\mathsf{focus}(x,y)\wedge\mathsf{EvalProc}(y)).$$

Then $q(x) \equiv_{\mathcal{T},\Sigma} \mathsf{VenipunctBloodTest}(x)$. Note that, when Σ is full, $q(x)$ is not equivalent to any query with only one atom.

To illustrate the use of multiple TBoxes, we come back to ontology versioning, already mentioned as a relevant application in the introduction; note that versioning is an active research area with a wide variety of approaches, ranging from purely syntactic to fully semantic, logic-based methods (Noy and Musen 2002; Klein et al. 2002; Jimnez-Ruiz et al. 2011; Gonçalves, Parsia, and Sattler 2011; Konev, Walther, and Wolter 2008). Assume that the above TBox $\mathcal{T}$ is updated to the new version $\mathcal{T}'$ in which the first CI is replaced with $\mathsf{LabTest} \sqsubseteq \mathsf{LabProc}$ (this modification corresponds to an update that was made in the 'real' SNOMED CT). To check whether the above query $q(x)$ is unaffected by the update, the user checks whether $(\mathcal{T},q) \subseteq_\Sigma (\mathcal{T}',q)$. In fact, this is not the case, as witnessed by the ABox

$$\mathcal{A} = \{\mathsf{Venipuncture}(a),\mathsf{purpose}(a,b),\mathsf{BloodSmpl}(b)\}$$

where $a \in \mathsf{cert}_{\mathcal{T}}(q,\mathcal{A})$, but $a \notin \mathsf{cert}_{\mathcal{T}'}(q,\mathcal{A})$.

The main reasoning problems studied in this paper are as follows.

Definition 2. Let $Q \in \{\mathrm{CQ},\mathrm{IQ}\}$ and let $\mathcal{L}$ be any of the DLs introduced above. Deciding

1. *Q-containment in $\mathcal{L}$* means to determine, given $q_1,q_2 \in Q$, $\mathcal{L}$-TBoxes $\mathcal{T}_1,\mathcal{T}_2$, and an ABox signature Σ, whether $(\mathcal{T}_1,q_1) \subseteq_\Sigma (\mathcal{T}_2,q_2)$.
2. *single TBox Q-containment in $\mathcal{L}$* means to determine, given $q_1,q_2 \in Q$, an $\mathcal{L}$-TBox $\mathcal{T}$, and an ABox signature Σ, whether $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$.
3. *full signature Q-containment in $\mathcal{L}$* means to determine, given $q_1,q_2 \in Q$ and $\mathcal{L}$-TBoxes $\mathcal{T}_1,\mathcal{T}_2$, whether $(\mathcal{T}_1,q_1) \subseteq (\mathcal{T}_2,q_2)$.
4. *full signature single TBox Q-containment in $\mathcal{L}$* means to determine, given $q_1,q_2 \in Q$ and an $\mathcal{L}$-TBox $\mathcal{T}$, whether $q_1 \subseteq_{\mathcal{T}} q_2$.

Note that Point 4 is the traditional query containment problem in DLs. The first three problems are closely related. In fact, the following lemma shows that, in the presence of an ABox signature, we can eliminate a second TBox.

Theorem 3. *Let $\mathcal{L} \in \{\mathcal{EL},\mathcal{EL}_\bot,\mathcal{ALC},\mathcal{ALCI},\mathcal{ALCF}\}$ and $Q \in \{$CQ, IQ$\}$. Then Q-containment in $\mathcal{L}$ can be polynomially reduced to single TBox Q-containment in $\mathcal{L}$.*

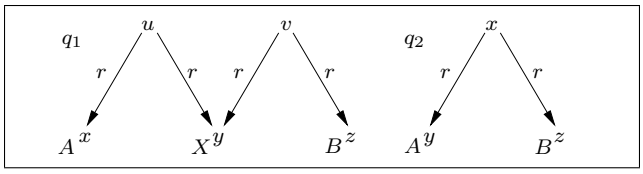

Figure 1: Query containment with restricted ABox signature

Proof. We sketch the proof for $\mathcal{EL}_\bot$ and CQs. Let $\mathcal{T}_1, \mathcal{T}_2$ be $\mathcal{EL}_\bot$-TBoxes, q_1, q_2 CQs of the same arity, and Σ an ABox signature. To simulate $\mathcal{T}_1$ and $\mathcal{T}_2$ using a single TBox, we duplicate the vocabulary. For simplicity, assume first that Σ only contains concept names. We introduce fresh concept names $s_1(A)$ and $s_2(A)$ for every concept name A and fresh role names $s_1(r)$, $s_2(r)$ for every role name r, and extend s_1, s_2 to concepts and CQs in the obvious way; for example, $s_1(C)$ is C with every $X \in \mathsf{N_R} \cup \mathsf{N_C}$ replaced with $s_1(X)$. Let $\mathcal{T}_i' = \{s_i(C) \sqsubseteq s_i(D) \mid C \sqsubseteq D \in \mathcal{T}_i\}$, for $i = 1, 2$ and

$$\mathcal{T} \quad = \quad \mathcal{T}_1' \cup \mathcal{T}_2' \cup \{A \sqsubseteq s_1(A) \sqcap s_2(A) \mid A \in \Sigma \cap \mathsf{N_C}\}.$$

We then have $(\mathcal{T}_1, q_1) \subseteq_\Sigma (\mathcal{T}_2, q_2)$ iff $s_1(q_1) \subseteq_{\mathcal{T},\Sigma} s_2(q_2)$.

The general case, where Σ may contain role names, requires some technical tricks. Since an assertion $r(a, b)$ in the ABox cannot be 'copied' into two assertions $s_1(r)(a, b)$ and $s_2(r)(a, b)$ as in the last component of $\mathcal{T}$, we leave role names from Σ untouched and do not replace them with fresh symbols when defining the TBoxes $\mathcal{T}_i'$. This is unsound as it enables undesired interaction between $\mathcal{T}_1$ and $\mathcal{T}_2$; we rectify this problem by introducing additional concept names E_1 and E_2 that represent the 'active domains' of $\mathcal{T}_1$ and $\mathcal{T}_2$, and syntactically relativize all concepts in $\mathcal{T}_i$ to E_i in the definition of $\mathcal{T}_i'$. ❑

The following result shows that, in the presence of two TBoxes, we can eliminate the ABox signature.

Theorem 4. *For $\mathcal{L} \in \{\mathcal{EL}_\bot, \text{DL-Lite}_{\mathsf{core}}, \text{DL-Lite}_{\mathsf{horn}}, \mathcal{ALC}, \mathcal{ALCI}, \mathcal{ALCF}\}$ and $Q \in \{\text{CQ}, \text{IQ}\}$, Q-containment in $\mathcal{L}$ can be polynomially reduced to full signature Q-containment in $\mathcal{L}$.*

Proof. Here, we only illustrate the proof idea using an example. Assume that we are interested in deciding CQ containment in the presence of the TBox $\mathcal{T} = \{A \sqsubseteq \exists r.A\}$ and with the ABox signature restricted to $\Sigma = \{A\}$. Set

$$\mathcal{T}' = \{A \sqsubseteq \exists r'.A, \exists r.\top \sqsubseteq \bot\}$$

Then, for any two CQs q_1, q_2 of the same arity using the symbols A and r only, $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$ iff $(\mathcal{T}, q_1) \subseteq (\mathcal{T}', q_2')$, where q_2' is obtained from q_2 by replacing any occurrence of r by r'. The equivalence holds since for any ABox $\mathcal{A}$ using the non-Σ-symbol r, $\mathcal{A}$ is not consistent w.r.t. $\mathcal{T}'$. ❑

The next example illustrates a central reason for why restricting the ABox signature (or admitting two distinct TBoxes) can make query containment harder.

Example 5. *Restricting the signature of ABoxes or admitting two TBoxes can introduce a form of disjunction. This is illustrated by the queries q_1 and q_2 in Figure 1, where all variables are quantified, together with the very simple DL-Lite$_{\mathsf{core}}$-TBox $\{A \sqsubseteq X, B \sqsubseteq X\}$ and ABox signature $\Sigma = \{A, B, r\}$. Note that, by definition of $\mathcal{T}$ and Σ, an ABox can only 'enforce' the concept name X by an assertion $A(a)$* or *an assertion $B(a)$, which is the mentioned disjunction and results in the fact that $q_1 \equiv_{\mathcal{T},\Sigma} q_2$.*

Some lower bounds in this paper are inherited from query emptiness, a reasoning problem that is defined as follows: given a query q, TBox $\mathcal{T}$, and ABox signature Σ, decide whether there exists a Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$ such that $\mathsf{cert}_\mathcal{T}(q, \mathcal{A}) \neq \emptyset$. The following lemma shows that query emptiness can be polynomially reduced to single TBox query containment, both for IQs and CQs and for any DL studied in this paper.

Lemma 6. *Let q be a CQ, $\mathcal{T}$ be an $\mathcal{ALCFI}$-TBox, Σ be an ABox signature, A be a concept name that does not occur in q, Σ, and $\mathcal{T}$, and q_A be any query with the same arity as q that uses A. Then there exists a Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$ with $\mathsf{cert}_\mathcal{T}(q, \mathcal{A}) \neq \emptyset$ iff $q \not\subseteq_{\mathcal{T},\Sigma} q_A$.*

3 Containment in DL-Lite

For query containment in the DL-Lite family, we find a difference in complexity depending on whether we consider IQs or CQs. We show IQ-containment to be tractable for DL-Lite$_{\mathsf{core}}$ and co-NP-complete for DL-Lite$_{\mathsf{horn}}$, whereas CQ-containment is Π_2^p-complete for both logics. The lower bounds are proven for single TBox containment, and hence also hold for full signature containment by Theorem 4 .

We begin by the tractability result for IQs in DL-Lite$_{\mathsf{core}}$.

Theorem 7. *IQ-containment in DL-Lite$_{\mathsf{core}}$ is in* PTIME.

Proof. It can be proved that $(\mathcal{T}_1, A(x)) \subseteq_\Sigma (\mathcal{T}_2, B(x))$ if and only if for every $C \in \mathsf{N_C} \cap \Sigma \cup \{\exists r, \exists r^- \mid r \in \mathsf{N_R} \cap \Sigma\}$ we have that $\mathcal{T}_1 \models C \sqsubseteq A$ implies $\mathcal{T}_2 \models C \sqsubseteq B$. The latter property can be verified in polynomial time (Calvanese et al. 2007). ❑

We now turn to DL-Lite$_{\mathsf{horn}}$, showing IQ-containment to be coNP-complete and giving a Π_2^p upper bound for CQ-containment.

Theorem 8. *In DL-Lite$_{\mathsf{horn}}$, IQ-containment is* coNP*-complete and CQ-containment is in Π_2^p.*

Proof. We start with instance queries. When $(\mathcal{T}_1, A(x)) \not\subseteq_\Sigma (\mathcal{T}_2, B(x))$, then there is a Σ-ABox $\mathcal{A}$ and an $a \in \mathsf{Ind}(\mathcal{A})$ with $\mathcal{T}_1, \mathcal{A} \models A(a)$ and $\mathcal{T}_2, \mathcal{A} \not\models B(a)$. For each role $r \in \Sigma$, choose an $a_r \in \mathsf{Ind}(\mathcal{A})$ with $r(a, a_r) \in \mathcal{A}$, if such exists. Let $\mathcal{A}'$ be the restriction of $\mathcal{A}$ to the individuals $\{a\} \cup \{a_r \mid r \in \Sigma\}$. Then $\mathcal{T}_1, \mathcal{A}' \models A(a)$ and $\mathcal{T}_2, \mathcal{A}' \not\models B(a)$. To decide instance query non-entailment, we may thus guess an ABox with $|\mathsf{Ind}(\mathcal{A})| \leq |\Sigma| + 1$ and an $a \in \mathsf{Ind}(\mathcal{A})$ and then check in polytime whether $\mathcal{T}_1, \mathcal{A} \models A(a)$ and $\mathcal{T}_2, \mathcal{A} \not\models B(a)$ (Artale et al. 2009). For the lower bound, we use the coNP-hardness of IQ-emptiness (Baader et al. 2010) together with Lemma 6.

We only sketch the proof for CQs. A witness for $(\mathcal{T}_1, q_1) \not\subseteq_\Sigma (\mathcal{T}_2, q_2)$ consists of an ABox $\mathcal{A}$, a tuple of individuals $\vec{a}$ from $\mathsf{Ind}(\mathcal{A})$, and a part of the canonical model for $\mathcal{A}$ and $\mathcal{T}_1$ (see (Kontchakov et al. 2010)) such that q_1 has an $\vec{a}$-match in that part and $\mathcal{T}_2, \mathcal{A} \not\models q_2(\vec{a})$. Similarly

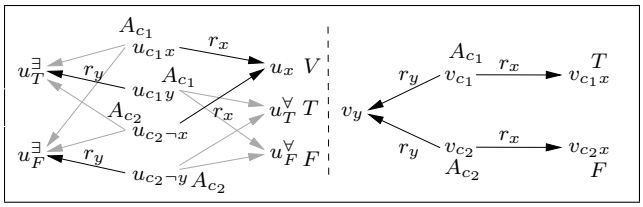

Figure 2: Queries q_1, q_2 for QBF $\forall x \exists y (x \vee y) \wedge (\neg x \vee \neg y)$.

to the case of IQs, we can suppose that the witness ABox $\mathcal{A}$ satisfies $|\mathsf{Ind}(\mathcal{A})| \leq |q| \cdot (|\Sigma| + 1)$. To decide CQ non-containment, we guess such a witness and then verify in coNP that $\mathcal{T}_2, \mathcal{A} \not\models q_2(\vec{a})$. ❑

To complete the picture, a Π_2^p-lower bound for single TBox CQ-containment in DL-Lite$_{\mathsf{core}}$ is proved in the following. Interestingly, it requires only TBox statements of the form $A \sqsubseteq B$ (no roles in the TBox, no disjointness). The construction is inspired by the Π_2^p-lower bound proof for containment of positive existential queries in (Sagiv and Yannakakis 1980).

The proof is by reduction from the validity of Π_2^p-QBFs, i.e., formulas of the form $\forall \vec{x} \exists \vec{y}\, \varphi(\vec{x}, \vec{y})$ with φ a propositional formula in CNF. We call the variables in $\vec{x}$ *universal variables* and the variables in $\vec{y}$ *existential variables* and aim at finding $\mathcal{T}, \Sigma, q_1$, and q_2 such that $\forall \vec{x} \exists \vec{y}\, \varphi(\vec{x}, \vec{y})$ is valid iff $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$. Set

$$\begin{array}{rcl} \mathcal{T} & = & \{T \sqsubseteq V, F \sqsubseteq V\} \\ \Sigma & = & \{T, F\} \cup \{A_c \mid c \text{ clause in } \varphi\} \cup \{r_x \mid x \in \vec{x} \cup \vec{y}\}. \end{array}$$

The query q_1 consists of the following atoms, where all variables are quantified variables:

- $F(u_F^\forall)$, $T(u_T^\forall)$, and $V(u_x)$ for each universal variable x;
- for each clause c in φ and each literal ℓ in c:
  1. $A_c(u_{c\ell})$;
  2. $r_x(u_{c\ell}, u_x)$ if $\ell = x$ or $\ell = \neg x$, with x a universal variable;
  3. $r_y(u_{c\ell}, u_T^\exists)$ if $\ell = y$ is an existential variable;
  4. $r_y(u_{c\ell}, u_F^\exists)$ if $\ell = \neg y$, with y an existential variable;
  5. for each universal variable x different from the variable in ℓ, the atoms $r_x(u_{c\ell}, u_F^\forall)$ and $r_x(u_{c\ell}, u_T^\forall)$;
  6. for each existential variable y different from the variable in ℓ, the atoms $r_y(u_{c\ell}, u_F^\exists)$ and $r_y(u_{c\ell}, u_T^\exists)$.

The query q_2 consists of the following atoms, for each clause c in φ, where again all variables are quantified variables:

- $A_c(v_c)$;
- for each universal variable x that occurs positively in c, the atoms $r_x(v_c, v_{cx})$ and $T(v_{cx})$;
- for each universal variable x that occurs negatively in c, the atoms $r_x(v_c, v_{cx})$ and $F(v_{cx})$;
- for each existential variable y in c, the atom $r_y(v_c, v_y)$.

An example can be found in Figure 2, where the edges from Points 5 and 6 above are drawn in gray to improve readability. Intuitively, the atoms $V(u_x)$ in q_1, together with the $T(v_{cx})$ and $F(v_{cx})$ in q_2 and the fact that $V \notin \Sigma$ ensures all truth assignments to universal variables are considered (cf. Example 5). The truth assignment for the existential variables is selected via the variables v_y of q_2, which intuitively can either be mapped to where $u_T^\exists$ of q_1 is mapped or to where $u_F^\exists$ of q_1 is mapped. Note that each v_{c_i} of q_2 can be mapped to where any of the $u_{c_i\ell}$ of q_1 is mapped, which corresponds to selecting a literal in c_i that is made true.

Theorem 9. *Single TBox CQ-containment in DL-Lite$_{\mathsf{core}}$ is Π_2^p-hard.*

4 Containment in $\mathcal{EL}$, General TBoxes

We study the complexity of IQ- and CQ-containment in $\mathcal{EL}$, concentrating on the most general form of TBox as introduced in Section 2. Our main result is that query containment is ExpTime-complete both for CQs and IQs. The lower bound holds already for single TBox containment and full signature containment, and the upper bound is for $\mathcal{EL}_\bot$.

We start by establishing an ExpTime lower bound for IQ-containment in the single TBox case. By contrast, note that query emptiness can be checked in polynomial time for $\mathcal{EL}$-TBoxes.

Theorem 10. *Single-TBox IQ-containment in $\mathcal{EL}$ is* ExpTime*-hard.*

Proof. The proof is by reduction from instance query emptiness in $\mathcal{EL}_\bot$, shown to be ExpTime-hard in (Baader et al. 2010). Specifically, we establish the following.

Claim. $A(x)$ is Σ-empty w.r.t. $\mathcal{T}$ iff $A(x) \subseteq_{\mathcal{T}',\Sigma} B(x)$ where $B \in \mathsf{N_C}$ is fresh and $\mathcal{T}'$ is obtained from $\mathcal{T}$ by (a) replacing every assertion $C \sqsubseteq \bot$ in $\mathcal{T}$ with $C \sqsubseteq B$ and (b) adding $\exists r.B \sqsubseteq B$ for every role r in $\mathcal{T}$ and Σ. ❑

To obtain an ExpTime lower bound for full signature containment, we show that an analogue of Theorem 4 holds in $\mathcal{EL}$ for instance queries. The proof is similar to that of Theorem 4 and relies on the fact that we can force the second instance query to hold whenever the ABox contains a non-Σ symbol.

Theorem 11. *In $\mathcal{EL}$, IQ-containment can be polynomially reduced to full signature IQ-containment.*

In the remainder of the section, we aim to prove an ExpTime upper bound for CQ-containment in $\mathcal{EL}_\bot$. Because of Theorem 3, it is sufficient to consider single TBox containment. The key insight is that if $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$, then this is witnessed by a *forest-shaped* ABox that consists of a small core whose relational structure is not restricted in any way and a (potentially infinite) tree below each core element. Our decision procedure is based upon a reduction of containment to the existence of a compact description of such a witness ABox.

Tree-shaped queries play an important role in what follows, so we begin by recalling their definition and introducing some relevant notation. We recall that each CQ q can be viewed as a directed graph $G_q = (V_q, E_q)$ with $V_q = \mathsf{term}(q)$ and $E_q = \{(t, t') \mid r(t, t') \in q \text{ for some } r \in \mathsf{N_R}\}$. We call q *tree-shaped* if G_q is a tree and $r(t, t'), s(t, t') \in q$

implies $r = s$. If q is tree-shaped and t is the root of G_q, we call t the *root of* q.

A query q' is *obtained from q by performing fork elimination* if q' is obtained from q by selecting two atoms $r(x, z)$ and $r(y, z)$ with x, y, z quantified variables and $x \neq y$ and then identifying x and y. A query which is obtained from q by repeatedly (but not necessarily exhaustively) performing fork elimination is called a *fork rewriting* of q.

For a CQ q and $t \in \mathsf{term}(q)$, let $\mathsf{Reach}_q(t)$ denote the set of all terms that are reachable from t in the directed graph G_q. For $T \subseteq \mathsf{term}(q)$, we write $q|_T$ to denote for the restriction of q to atoms that contain only terms from T. Define

$$\begin{aligned}
\mathsf{Trees}(q) &:= \{q|_{\mathsf{Reach}_q(x)} \mid x \in \mathsf{term}(q) \text{ a variable} \\
&\qquad \text{and } q|_{\mathsf{Reach}_q(x)} \text{ tree-shaped}\} \\
\mathsf{Trees}^+(q) &:= \{r(t,x) \wedge q' \mid r(t,x) \in q \\
&\qquad \text{and } q' = \emptyset \text{ or } q' \in \mathsf{Trees}(q) \text{ has root } x\} \\
\mathsf{Trees}^*(q) &:= \bigcup_{q' \text{ fork rewriting of } q} \mathsf{Trees}(q') \cup \mathsf{Trees}^+(q').
\end{aligned}$$

The cardinality of $\mathsf{Trees}(q)$ and $\mathsf{Trees}^+(q)$ is clearly polynomial in the size of q, and it follows from results from (Lutz 2008) that this is true of $\mathsf{Trees}^*(q)$ as well.

It is well-known that whenever $\mathcal{T}, \mathcal{A} \not\models q$ for an $\mathcal{EL}_\bot$-TBox $\mathcal{T}$, ABox $\mathcal{A}$, and CQ q, then this is witnessed by a forest-shaped model of $\mathcal{T}$ and $\mathcal{A}$. We now introduce the notion of a match candidate, which intuitively describes a possible match of a CQ in such a model. Let $\mathcal{A}$ be an ABox, q be a CQ with answer variables $x_1, \ldots, x_\ell$, and $\vec{a} = a_1, \ldots, a_\ell \in \mathsf{Ind}(\mathcal{A})^\ell$ be a candidate answer to q in $\mathcal{A}$. An *$\vec{a}$-match candidate for q in $\mathcal{A}$* is a tuple $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ where $p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m$ is a partitioning of q and $f : \mathsf{term}(q) \to \mathsf{Ind}(\mathcal{A})$ maps each term in q to an individual name in $\mathcal{A}$. Let $p_i = r_i(t_i, y_i) \wedge p_i'$ for $1 \leq i \leq n$. We require that the following conditions are satisfied:

1. $p_1, \ldots, p_n \in \mathsf{Trees}^+(q)$ and $\hat{p}_1, \ldots, \hat{p}_m \in \mathsf{Trees}(q)$;
2. $f(x_i) = a_i$ for $1 \leq i \leq \ell$;
3. $f(a) = a$ for all $a \in \mathsf{term}(q) \cap \mathsf{N_I}$;
4. $A(t) \in p_0$ implies $A(f(t)) \in \mathcal{A}$;
5. $r(t, t') \in p_0$ implies $r(f(t), f(t')) \in \mathcal{A}$;
6. the p_i' and the $\hat{p}_i$ contain only quantified variables;
7. if $s \in \{p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m\}$ and $x, y \in \mathsf{term}(s)$, then $f(x) = f(y)$;
8. $\mathsf{term}(p_0) \cup \{t_1, \ldots, t_n\}, \mathsf{term}(p_1'), \ldots, \mathsf{term}(p_n'), \mathsf{term}(\hat{p}_1), \ldots, \mathsf{term}(\hat{p}_m)$ are pairwise disjoint.

Intuitively, the function f is used to map (i) each term in p_0 and (ii) each subquery p_i and $\hat{p}_i$ ($i > 0$) to some individual name. To achieve this, we use the uniformity condition 7 and set $f(p_i) = f(x)$ (resp. $f(\hat{p}_i) = f(x)$) where x is any variable in p_i (resp. $\hat{p}_i$). Now, a match candidate describes a match of q in a forest-shaped model of $\mathcal{T}$ and $\mathcal{A}$ as follows: the atoms in p_0 are mapped to the core of the forest-shaped model with each $t \in \mathsf{term}(p_0)$ being mapped to $f(t)$. Each query p_i is mapped to the tree below $f(p_i)$ in a rooted way (the root is the term t_i), and each query $\hat{p}_i$ is mapped to the tree below $f(\hat{p}_i)$ in a non-rooted way.

We now relate query entailment to the existence of a match candidate. To do this, we represent tree-shaped CQs as concepts expressed in $\mathcal{EL}$ or its extension $\mathcal{EL}^u$. For our purposes, an $\mathcal{EL}^u$ concept takes the form C or $\exists u.C$ where C is an $\mathcal{EL}$-concept and u is the *universal role*, interpreted as $u^\mathcal{I} = \Delta^\mathcal{I} \times \Delta^\mathcal{I}$. To every tree-shaped CQ q and term $t \in \mathsf{term}(q)$, we associate an $\mathcal{EL}$-concept $C_{q,t}$ as follows:

$$C_{q,t} = \bigsqcap_{A(t) \in q} A \sqcap \bigsqcap_{r(t,t') \in q} \exists r.C_{q,t'}.$$

We use C_q to abbreviate $C_{q,t}$ when t is the root of q and C_q^u to denote the $\mathcal{EL}^u$-concept $\exists u.C_q$.

Lemma 12. *Suppose $\mathcal{A}$ is consistent with $\mathcal{T}$, and let $\mathsf{close}(\mathcal{T}, \mathcal{A})$ denote the ABox*

$$\mathcal{A} \cup \{A(a) \mid \mathcal{T}, \mathcal{A} \models A(a), a \in \mathsf{Ind}(\mathcal{A}), A \in \mathsf{N_C}\}.$$

Then $\mathcal{T}, \mathcal{A} \models q(\vec{a})$ iff there exists a fork rewriting q' of q and an $\vec{a}$-match candidate $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ for q' and $\mathsf{close}(\mathcal{T}, \mathcal{A})$ such that $\mathcal{T}, \mathcal{A} \models C(a)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C_{\hat{p}_i}^u \mid f(\hat{p}_i) = a\}$.

According to the preceding lemma, a query is not entailed just in the case that for each match candidate for a fork rewriting, one of the required concept assertions is not entailed. This leads us to define the notion of a spoiler, which describes a possible way of avoiding a query match. We say a map

$$\nu : \mathsf{Ind}(\mathcal{A}) \to 2^{\{C_{q'}, C_{q'}^u \mid q' \in \mathsf{Trees}^*(q)\}}$$

is an *$\vec{a}$-spoiler for q w.r.t. $\mathcal{A}$* if for every fork rewriting q' of q and $\vec{a}$-match candidate $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ for q' in $\mathcal{A}$, there is a p_i such that $C_{p_i} \in \nu(f(p_i))$ or a $\hat{p}_i$ such that $C_{\hat{p}_i}^u \in \nu(f(\hat{p}_i))$.

We are now ready to define compact witnesses, which are the key ingredient to our upper bound. A *compact witness for $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$* is a tuple $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ where $\mathcal{A}_w$ is a $\mathsf{sig}(\mathcal{T})$-ABox with $|\mathsf{Ind}(\mathcal{A}_w)|$ bounded by $|\mathsf{term}(q_1)|$, $\vec{a}_w$ a candidate answer to q_1 and q_2 in $\mathcal{A}_w$, q_1' a fork rewriting of q_1, $\Pi_w = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ an $\vec{a}$-match candidate for q_1' in $\mathcal{A}_w$, and ν_w an $\vec{a}$-spoiler for q_2 w.r.t. $\mathcal{A}_w$ such that the following conditions are satisfied:

1. $\mathcal{A}_w$ is consistent w.r.t. $\mathcal{T}$;
2. if $\mathcal{T}, \mathcal{A}_w \models A(a)$ with $a \in \mathsf{Ind}(\mathcal{A}_w)$ and $A \in \mathsf{N_C}$, then $A(a) \in \mathcal{A}_w$;
3. all role names in $\mathcal{A}_w$ are from Σ;
4. for each $a \in \mathsf{Ind}(\mathcal{A}_w)$, there is a tree-shaped Σ-ABox $\mathcal{A}_a$ with root a which is consistent with $\mathcal{T}$ and such that
   (a) $A(a) \in \mathcal{A}_w$ iff $A(a) \in \mathcal{A}_a$ for all $A \in \mathsf{N_C} \cap \Sigma$;
   (b) $\mathcal{T}, \mathcal{A}_a \models A(a)$ iff $A(a) \in \mathcal{A}_w$ for all $A \in \mathsf{N_C} \setminus \Sigma$;
   (c) $\mathcal{T}, \mathcal{A}_a \models C(a)$ for all concepts C from
   $$\{C_{p_i} \mid f(p_i) = a\} \cup \{C_{\hat{p}_i}^u \mid f(\hat{p}_i) = a\};$$
   (d) $\mathcal{T}, \mathcal{A}_a \not\models C(a)$, for every $C \in \nu_w(a)$.

Given a compact witness $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ for $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$, the desired forest-shaped witness ABox, call it $\mathcal{A}_f$, can be constructed by taking the role assertions in $\mathcal{A}_w$ and attaching the tree-shaped Σ-ABoxes described by condition 4. Note that because of conditions 2 and 4(a-b), the concept assertions in $\mathcal{A}_w$ are precisely the atomic concept assertions concerning individuals in $\mathcal{A}_w$ which are entailed from $\mathcal{T}, \mathcal{A}_f$. Consistency of the whole ABox $\mathcal{A}_f$ follows from conditions 1 and 4. To show entailment of $q_1(\vec{a})$, we use the match candidate Π_w together with condition 4(c), and for non-entailment of $q_2(\vec{a})$, we use the spoiler ν_w together with condition 4(d).

The following theorem gives the reduction announced earlier. It holds only for *normalized TBoxes* whose axioms are of the forms $A \sqsubseteq B$, $A_1 \sqcap A_2 \sqsubseteq B$, $A_1 \sqsubseteq \exists r.A_2$, and $\exists r.A \sqsubseteq B$, where $A, A_1, A_2 \in \mathsf{N_C}$ and $B \in \mathsf{N_C} \cup \{\bot\}$. This is unproblematic since every $\mathcal{EL}_\bot$-TBox can be transformed into a normalized $\mathcal{EL}_\bot$-TBox that is a model conservative extension of $\mathcal{T}$ (Baader, Brandt, and Lutz 2005).

Theorem 13. *If $\mathcal{T}$ is normalized, $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$ iff there is a compact witness for $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$.*

Based on Theorem 13, our decision procedure is as follows: enumerate all tuples $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ that are candidates for compact witnesses for $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$, i.e., $\mathcal{A}_w$ is a $\mathsf{sig}(\mathcal{T})$-ABox with $|\mathsf{Ind}(\mathcal{A}_w)|$ bounded by $|\mathsf{term}(q_1)|$, $\vec{a}_w$ a candidate answer to q_1 and q_2 in $\mathcal{A}_w$, q_1' a fork rewriting of q_1, $\Pi_w = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ an $\vec{a}$-match candidate for q_1' in $\mathcal{A}_w$, and ν_w an $\vec{a}$-spoiler for q_2 w.r.t. $\mathcal{A}_w$. It is not hard to verify that there are only exponentially many such tuples. For each tuple, verify whether it satisfies conditions 1 to 4 of compact witnesses. This can clearly be done in polynomial time for conditions 1 to 3, and thus it remains to deal with condition 4. This is less straightforward, and we use an automaton construction to show the following.

Proposition 14. *Given an $\mathcal{EL}_\bot$-TBox $\mathcal{T}$, an ABox signature Σ, and sets of $\mathcal{EL}^u$-concepts Ψ_1 and Ψ_2, it is in* ExpTime *to decide whether there is a tree-shaped Σ-ABox $\mathcal{A}$ with root a such that*

1. *$\mathcal{A}$ is consistent w.r.t. $\mathcal{T}$;*
2. *$\mathcal{T}, \mathcal{A} \models C(a)$ for all $C \in \Psi_1$;*
3. *$\mathcal{T}, \mathcal{A} \not\models C(a)$ for all $C \in \Psi_2$.*

We thus obtain the desired upper bound:

Theorem 15. *CQ-containment in $\mathcal{EL}_\bot$ is in* ExpTime.

5 Containment in $\mathcal{EL}$, Classical TBoxes

A *classical $\mathcal{EL}$-TBox* is a set $\mathcal{T}$ of concept definitions $A \equiv C$ such that each A is a concept name and no concept name appears on the left-hand side of multiple definitions. $\mathcal{T}$ is called *acyclic* if it is classical and no concept name depends on itself; that is, for any sequence $A_0 \equiv C_0, \ldots, A_n \equiv C_n$ in $\mathcal{T}$ with A_{i+1} a subconcept of C_i, $0 \leq i < n$, A_0 is not a subconcept of C_n. The length of a longest such sequence is called the *definitorial depth* $d(\mathcal{T})$ of an acyclic TBox $\mathcal{T}$.

We show that IQ-containment is tractable for classical $\mathcal{EL}$-TBoxes and CQ-containment is PSpace-complete for acyclic $\mathcal{EL}$-TBoxes. The PSpace lower bound holds already for single TBox CQ-containment and full signature CQ-containment. The exact complexity of CQ-containment for (possibly cyclic) classical $\mathcal{EL}$-TBoxes remains open between PSpace and ExpTime.

Theorem 16. *IQ-containment in $\mathcal{EL}$ with classical TBoxes is in* PTime.

The proof of Theorem 16 applies techniques introduced in (Konev, Walther, and Wolter 2008; Konev et al. 2011) for proving that conservative extensions, Σ-entailment and Σ-inseparability between classical $\mathcal{EL}$-TBoxes are tractable. Here, we consider in more detail CQ-containment for acyclic $\mathcal{EL}$-TBoxes.

Theorem 17. *CQ-containment in $\mathcal{EL}$ with acyclic TBoxes is in* PSpace.

The polynomial reduction of CQ-containment to single TBox CQ-containment given in Theorem 3 can be modified to prove that CQ-containment for acyclic $\mathcal{EL}$-TBoxes is polynomially reducible to single TBox CQ-containment for acyclic $\mathcal{EL}$-TBoxes. Thus, it suffices to prove the PSpace upper bound for single TBox containment. Note that we can w.l.o.g. assume that Σ contains a distinguished role name $r_@$ that does not occur in the TBox and the queries since adding such a name does not impact the result of deciding containment.

We use a variation of the algorithm described in Section 4, and in particular Theorem 13. The central observation is that a PSpace procedure for single TBox CQ-containment w.r.t. *acyclic $\mathcal{EL}$-TBoxes* is obtained by guessing a candidate tuple for a compact witness and then using a variant of Proposition 14 to verify that it is indeed a compact witness. The additional role name $r_@$ can be traced all the way through to Proposition 14, thus it suffices to prove the following.

Proposition 18. *Given an acyclic $\mathcal{EL}$-TBox $\mathcal{T}$, an ABox signature Σ, and sets of $\mathcal{EL}^u$-concepts Ψ_1 and Ψ_2 such that the role name $r_@$ occurs in Σ, but not in $\mathcal{T}$, Ψ_1, and Ψ_2, it is in* PSpace *to decide whether there is a tree-shaped Σ-ABox $\mathcal{A}$ with root a such that Points 1 to 3 of Proposition 14 hold.*

Proof. (sketch) We assume that $\mathcal{T}$ is in a certain normal form in which each $A \equiv C$ is of the form $A \equiv \exists r.B$ or $A \equiv B_1 \sqcap \cdots \sqcap B_n$, where $B, B_1, \ldots, B_n$ are concept names; details are given in the long version. Call a concept name A *primitive in $\mathcal{T}$* if it does not occur on the left hand side of a concept definition in $\mathcal{T}$. Let $\Psi_i = \Psi_i^l \cup \{\exists u.C \mid C \in \Psi_i^u\}$, where Ψ_i^l and Ψ_i^u are sets of $\mathcal{EL}$-concepts, for $i = 1, 2$. Let S be the set of subconcepts of $\mathcal{T} \cup \Psi_1^l \cup \Psi_1^u \cup \Psi_2^l \cup \Psi_2^u$ and w.l.o.g. $\top \in S$.

We define a recursive polynomial space procedure computing the predicate $\mathsf{ABox}(\Gamma, k)$ for $\Gamma \subseteq S$ such that $\mathsf{ABox}(\Gamma, k) = 1$ iff there is a tree-shaped Σ-ABox $\mathcal{A}$ with root a of depth $\leq k$ and such that $\mathcal{T}, \mathcal{A} \models C(a)$ iff $C \in \Gamma$ (for all $C \in S$), and $\mathcal{T}, \mathcal{A} \not\models \exists u.C(a)$ for all $C \in \Psi_2^u$. $\mathsf{ABox}(\Gamma, 0) = 1$ can be checked in polyspace. For $n > 0$, one can show that $\mathsf{ABox}(\Gamma, n) = 1$ if and only if the following conditions are satisfied:

(b1) $\top \in \Gamma$ and $\Psi_2^u \subseteq S \setminus \Gamma$;

(b2) for all $C_1, C_2 \in S$: if $C_1 \in \Gamma$ and $\mathcal{T} \models C_1 \sqsubseteq C_2$, then $C_2 \in \Gamma$;

(b3) for all $A \in \Gamma \cap \Sigma$ and $C \in \Psi_2^u$: $\mathcal{T} \not\models A \sqsubseteq \exists u.C$;

(b4) for all A primitive in $\mathcal{T}$: $A \in \Gamma$ iff there exists $B \in \Sigma \cap \Gamma$ such that $\mathcal{T} \models B \sqsubseteq A$;

(b5) for all $\exists r.C \in S$: $\exists r.C \in \Gamma$ iff there exists $B \in \Sigma \cap \Gamma$ such that $\mathcal{T} \models B \sqsubseteq \exists r.C$ or there exist $\Gamma' \subseteq S$ such that, recursively: (1) $C \in \Gamma'$, (2) for all $D \in S$: if $\mathcal{T} \models \exists r.(\bigsqcap \Gamma') \sqsubseteq D$, then $D \in \Gamma$, (3) $\mathsf{ABox}(\Gamma', n-1) = 1$.

Conditions (b1) to (b5) can be checked in polyspace. Let $d(S)$ denote the maximal number of nestings of existential restriction in concepts from S. Now one can show that an ABox with the properties from Proposition 18 exists if

1. there exists $\Gamma \subseteq S$ such that $\mathsf{ABox}(\Gamma, d(\mathcal{T}) + d(S)) = 1$, $\Psi_1^l \subseteq \Gamma$ and $\Psi_2^l \cap \Gamma = \emptyset$;
2. for all $C \in \Psi_1^u$: there exists $\Gamma \subseteq S$ such that $\mathsf{ABox}(\Gamma, d(\mathcal{T}) + d(S)) = 1$ and $C \in \Gamma$ or $\mathcal{T} \models A \sqsubseteq \exists u.C$ for some $A \in \Gamma \cap \Sigma$.

One obtains the required tree-shaped Σ-ABox by adding the roots of the tree-shaped Σ-ABoxes obtained in Point 2 for $C \in \Psi_1^u$ as $r_@$-successors to the root of the Σ-ABox obtained in Point 1. ❑

We now state the matching lower bound.

Theorem 19. *CQ-containment in $\mathcal{EL}$ with acyclic TBoxes is* PSPACE*-hard. This is true already for single TBox and full signature containment.*

We consider single TBox containment and give a polytime reduction of the validity of QBF formulas $\varphi = Q_1 p_1 \dots Q_n p_n.\vartheta$ with $\vartheta = \bigsqcap_{c \in C} c$ a propositional formula in CNF, C its clauses, and $Q_i \in \{\forall, \exists\}$, for $1 \leq i \leq n$. Recall that a *validation tree* for φ is a tree of depth n in which every level (except the leaves) corresponds to one of the quantifiers in φ. In $\forall p_i$-levels, each node has two successors, one for $p_i = \top$ and one for $p_i = \bot$. In $\exists p_i$-levels, each node has one successor, either for $p_i = \top$ or for $p_i = \bot$. Thus, every branch of a validation tree corresponds to a truth assignment to the variables $p_1, \dots, p_n$, and it is required that the propositional formula ϑ evaluates to true on every branch. We say φ is *valid* iff there exists a validation tree for φ.

Now let $\varphi = Q_1 p_1 \dots Q_n p_n.\vartheta$ be of the form above. In the reduction, the existence of a validation tree is encoded in the existence of a Σ-ABox $\mathcal{A}$ that witnesses $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$. To encode the tree structure, we use a role name r to represent the edges of the validation tree, and the concept names $L_0, \dots, L_n$ to identify the n levels. The truth values of the variables $p_1, \dots, p_n$ are represented via the concept names $P_1, \dots, P_n$ (indicating truth) and $\bar{P}_1, \dots, \bar{P}_n$ (indicating falsity). Concept names $V_1, \dots, V_n$ are used to indicate that either P_i or $\bar{P}_i$ holds. Finally concept names P_c, $c \in C$, indicate that the clause c of ϑ is true. We set $\Sigma = \{r\} \cup \{P_i, \bar{P}_i \mid i \leq n\}$ and let $\mathcal{T}$ consist of, for $i \leq n$:

$$P_i \sqsubseteq V_i \sqcap \bigsqcap_{P_i \in c, c \in C} P_c, \quad \bar{P}_i \sqsubseteq V_i \sqcap \bigsqcap_{\bar{P}_i \in C, c \in C} P_c,$$

for $i < n$ and $Q_i = \forall$:

$$L_i \equiv \exists r.(P_{i+1} \sqcap L_{i+1} \sqcap \bigsqcap_{j \leq i} V_j) \sqcap \exists r.(\bar{P}_{i+1} \sqcap L_{i+1} \sqcap \bigsqcap_{j \leq i} V_j),$$

for $i < n$ and $Q_{i+1} = \exists$: $\quad L_i \equiv \exists r.(L_{i+1} \sqcap \bigsqcap_{j \leq i+1} V_j)$,

and, finally, $L_n \equiv \bigsqcap_{c \in C} P_c$.

Observe that if $\mathcal{T}, \mathcal{A} \models L_0(a)$ for a Σ-ABox $\mathcal{A}$, then one can show by induction that there exists a tree in $\mathcal{A}$ with root a satisfying L_i at every node of level i and satisfying $P_i \vee \bar{P}_i$ at all nodes of level j with $n \geq j \geq i$ (since V_i is satisfied in all those nodes and this can only be enforced by $P_i \vee \bar{P}_i$). Moreover, in $\forall p_i$-levels we have a successor node in which P_i holds and a successor node in which $\bar{P}_i$ holds. Finally, P_c is true in all leaf nodes for all $c \in C$. We, therefore, have a validation tree in which ϑ evaluates to true in every leaf iff there exists a Σ-ABox $\mathcal{A}$ with $\mathcal{T}, \mathcal{A} \models L_0(a)$ and $\mathcal{T}, \mathcal{A} \not\models C(a)$, for every $C \in \mathcal{C}$, where

$$\begin{aligned} \mathcal{C} \quad := \quad & \{\exists r^j.(P_i \sqcap \bar{P}_i) \mid 1 \leq j \leq n, 1 \leq i \leq n\} \cup \\ & \{\exists r^j.(\bar{P}_i \sqcap \exists r.P_i) \mid 1 \leq j \leq n-1, 1 \leq i < n\} \cup \\ & \{\exists r^j.(P_i \sqcap \exists r.\bar{P}_i) \mid 1 \leq j \leq n-1, 1 \leq i \leq n\} \end{aligned}$$

It follows that φ is valid iff $L_0(x) \not\subseteq_{\mathcal{T},\Sigma} \bigsqcup_{C \in \mathcal{C}} C(x)$. It now remains to encode the disjunction on the right-hand-side of this containment problem into an extension of $L_0(x)$ to a CQ. This is done in the long version by modifying a construction from a lower bound proof for ABox updates in $\mathcal{EL}$ (Liu, Lutz, and Milicic 2008). The proof for full signature containment is similar.

6 Containment in Expressive DLs

For $\mathcal{ALC}$ and $\mathcal{ALCI}$, we establish a $\mathrm{P}^{\mathrm{NExp}}$ upper bound on IQ-containment using the same technique that was used in (Baader et al. 2010) to prove such an upper bound for query emptiness. Embarrassingly enough, we do not know whether this bound is tight, neither for emptiness nor for containment, and we do not even know whether CQ-containment is decidable in $\mathcal{ALC}$ and $\mathcal{ALCI}$. In $\mathcal{ALCF}$, both IQ-containment and CQ-containment are undecidable as a direct consequence of the corresponding results for query emptiness, shown in (Baader et al. 2010).

Fix IQs $A_1(x)$, $A_2(x)$, a consistent $\mathcal{ALCI}$-TBox $\mathcal{T}$, and an ABox signature Σ such that it is to be decided whether $A_1(x) \subseteq_{\mathcal{T},\Sigma} A_2(x)$. The central observation is that one can construct, in exponential time, a single candidate Σ-ABox $\mathcal{A}_{\mathcal{T},\Sigma}$ such that $A_1(x) \not\subseteq_{\mathcal{T},\Sigma} A_2(x)$ iff $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \models q_1[a]$ and $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models q_2[a]$ for some $a \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma})$. $\mathcal{A}_{\mathcal{T},\Sigma}$ is called the *canonical Σ-ABox for $\mathcal{T}$* and constructed as follows. The *closure* $\mathsf{cl}(\mathcal{T}, \Sigma)$ is the smallest set that contains $\Sigma \cap \mathsf{N_C}$ as well as all concepts that occur (potentially as a subconcept) in $\mathcal{T}$ and is closed under single negations. A *type* for $\mathcal{T}$ and Σ is a set $t \subseteq \mathsf{cl}(\mathcal{T}, \Sigma)$ such that for some model $\mathcal{I}$ of $\mathcal{T}$ and some $d \in \Delta^{\mathcal{I}}$, we have $t = \{C \in \mathsf{cl}(\mathcal{T}, \Sigma) \mid d \in C^{\mathcal{I}}\}$. Let $\mathfrak{T}_{\mathcal{T},\Sigma}$ denote the set of all types for $\mathcal{T}$ and Σ. Let a_t be mutually distinct individual names for $t \in \mathfrak{T}_{\mathcal{T},\Sigma}$ and define $\mathcal{A}_{\mathcal{T},\Sigma}$ as follows:

$$\begin{aligned} \mathcal{A}_{\mathcal{T},\Sigma} = \{ & A(a_t) \mid A \in t \cap \Sigma \text{ and } t \in \mathfrak{T}_{\mathcal{T},\Sigma}\} \cup \\ & \{r(a_t, a_{t'}) \mid t, t' \in \mathfrak{T}_{\mathcal{T},\Sigma} \text{ and } r \in \Sigma \text{ and} \\ & \text{for all } \exists r.C \in \mathsf{cl}(\mathcal{T}, \Sigma) : C \in t' \Rightarrow \exists r.C \in t\}. \end{aligned}$$

The set $\mathfrak{T}_{\mathcal{T},\Sigma}$ can be computed in exponential time by making use of well-known EXPTIME procedures for concept satisfiability w.r.t. TBoxes in $\mathcal{ALCI}$ (Baader et al. 2003). Thus, $\mathcal{A}_{\mathcal{T},\Sigma}$ can be computed in exponential time.

The main property of $\mathcal{A}_{\mathcal{T},\Sigma}$ is that $A_1(x) \subseteq_{\mathcal{T},\Sigma} A_2(x)$ iff $\mathsf{cert}_{\mathcal{T}}(A_1(x), \mathcal{A}_{\mathcal{T},\Sigma}) \subseteq \mathsf{cert}_{\mathcal{T}}(A_2(x), \mathcal{A}_{\mathcal{T},\Sigma})$, which is proved using homomorphisms between ABoxes. Based on $\mathcal{A}_{\mathcal{T},\Sigma}$, we can now prove the main result of this section.

Theorem 20. *IQ-containment in $\mathcal{ALCI}$ is in* $\mathrm{P}^{\mathrm{NEXP}}$.

Proof. By Theorem 3, it is sufficient to consider single TBox containment. We show that non-containment is in $\mathrm{NP}^{\mathrm{NEXP}}$ and derive from $\mathrm{NP}^{\mathrm{NEXP}} \subseteq \mathrm{P}^{\mathrm{NEXP}}$ the desired result (Hemachandra 1987). Let $\mathcal{T}$ be a consistent $\mathcal{ALCI}$-TBox, Σ be an ABox signature, and $A_1(x)$, $A_2(x)$ be IQs for which it is to be decided whether $A_1(x) \subseteq_{\mathcal{T},\Sigma} A_2(x)$ is *not* the case. The algorithm guesses $a \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma})$, and then checks (1) $a \in \mathsf{cert}_{\mathcal{T}}(A_1(x), \mathcal{A}_{\mathcal{T},\Sigma})$ and (2) $a \notin \mathsf{cert}_{\mathcal{T}}(A_2(x), \mathcal{A}_{\mathcal{T},\Sigma})$, by calling a NEXPTIME oracle that decides the following problem: given an $\mathcal{ALCI}$-TBox $\mathcal{T}'$, signature Σ', individual $a' \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T}',\Sigma'})$ and IQ $A(x)$, does $a' \notin \mathsf{cert}_{\mathcal{T}'}(A(x), \mathcal{A}_{\mathcal{T}',\Sigma'})$ hold? Such an oracle exists: it computes the canonical ABox $\mathcal{A}_{\mathcal{T}',\Sigma'}$ and guesses a map $\pi : \mathsf{Ind}(\mathcal{A}_{\mathcal{T}',\Sigma'}) \to \mathfrak{T}_{\mathcal{T}',\Sigma'}$ and accepts if (i) $A \notin \pi(a')$, (ii) $C(c) \in \mathcal{A}_{\mathcal{T}',\Sigma'}$ implies $C \in \pi(c)$, and (iii) $r(b,c) \in \mathcal{A}_{\mathcal{T}',\Sigma'}$, $C \in \pi(c)$, and $\exists r.C \in \mathsf{cl}(\mathcal{T}', \Sigma')$ implies $\exists r.C \in \pi(b)$. ❑

The best known lower bound for IQ-containment in $\mathcal{ALC}$ and $\mathcal{ALCI}$ is EXPTIME. It stems from an easy reduction of unsatisfiability in $\mathcal{ALC}$: a concept C is unsatisfiable w.r.t. $\mathcal{T}$ iff $A(x) \subseteq_{\mathcal{T},\Sigma} B(x)$ where $\Sigma = \{A\}$ and $\mathcal{T} = \{A \sqsubseteq C\}$.

We close with stating undecidability for $\mathcal{ALCF}$.

Theorem 21. *In $\mathcal{ALCF}$, IQ-containment and CQ-containment are undecidable.*

7 Query Optimization

A classical result by Chandra and Merlin states that, in relational databases, any two CQs that are equivalent and minimal w.r.t. set inclusion must be isomorphic (Chandra and Merlin 1977). Given a CQ q, one can thus find the *unique minimal CQ* that is equivalent to q by applying lazy minimization: repeatedly remove atoms that are *redundant* in the sense that dropping them preserves equivalence. The order in which atoms are dropped is irrelevant, and thus the only non-trivial part is checking equivalence (i.e., two query containment checks).

It is not hard to see that, in the presence of TBoxes, Chandra-Merlin uniqueness of minimal CQs fails. This is true even for the classical version of query containment, the full signature single TBox case, which we will generally assume throughout this section. Given a TBox $\mathcal{T}$, we call a CQ q *$\mathcal{T}$-minimal w.r.t. set inclusion* if there is no $q' \subsetneq q$ with $q' \equiv_{\mathcal{T}} q$.

Example 22. *(1) Let $\mathcal{T} = \{A \equiv B\}$ and $q(x) = A(x) \wedge B(x)$. Then $q(x) \equiv_{\mathcal{T}} A(x)$ and $q(x) \equiv_{\mathcal{T}} B(x)$, and both $A(x)$ and $B(x)$ are $\mathcal{T}$-minimal w.r.t. set inclusion, but not isomorphic.*

(2)
$$\mathcal{T} = \{A \sqsupseteq \exists r.(B_1 \sqcap \exists s.\top) \sqcap \exists r.(B_2 \sqcap \exists s.\top),\ A \sqsubseteq \exists r.(B_1 \sqcap B_2 \sqcap \exists s.\top)\ \}$$
$$q(x) = \exists y_1, y_2, y_3.B_1(y_1) \wedge B_2(y_2) \wedge r(x,y_1) \wedge r(x,y_2) \wedge s(y_1,y_3) \wedge s(y_2,x_3).$$

Then $q(x)$ is $\mathcal{T}$-minimal w.r.t. set inclusion, but equivalent to the smaller (and non-isomorphic) query $A(x)$.

Note that Example (1) above uses only CIs of the very simple form $A \sqsubseteq B$ and thus applies to any reasonable DL. The second example, where the structural differences between the original and minimized query is large, is formulated in $\mathcal{EL}$. While these examples demonstrate that naive lazy minimization does not yield the desired result in the presence of DL TBoxes, we show in the following that strong guarantees on minimized queries can be recovered when applying more careful minimization strategies.

Optimization in DL-Lite

We consider the basic member DL-Lite$_{\mathsf{core}}$ of the DL-Lite family. A first observation is that, although the result of standard lazy minimization is not unique, it still yields queries of *minimum cardinality*. In the following, we use $\#q$ to denote the number of atoms in the CQ q. Clearly, one can view a CQ q as an ABox $\mathcal{A}_q$ by treating concept atoms as concept assertions and role atoms as role assertions. We say that q is *consistent with $\mathcal{T}$* if $\mathcal{A}_q$ and $\mathcal{T}$ have a common model.

Theorem 23. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox and q_1, q_2 CQs such that $q_1 \equiv_{\mathcal{T}} q_2$ and q_1, q_2 are $\mathcal{T}$-minimal w.r.t. set inclusion and consistent with $\mathcal{T}$. Then $\#q_1 = \#q_2$.*

We now show that even stronger guarantees can be obtained, concentrating on the class of *rooted CQs*, which are connected CQs that contain at least one answer variable or individual name. For these, we show that by adopting a suitable minimization strategy, we can find an equivalent query that is $\mathcal{T}$-minimal w.r.t. set inclusion and whose relational structure (restriction to role atoms) can be found in *any* equivalent query as a subquery. The strategy also guarantees a *unique result* of query minimization, similarly to Chandra and Merlin. We slightly generalize CQs and ABoxes by also admitting concept atoms of the form $\exists r(t)$ and ABox assertions of the form $\exists r(a)$.

Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox. To break ties between equivalent concepts during minimization (see Part (1) of Example 22), we fix a strict partial order $<$ on the concepts that occur in $\mathcal{T}$ such that $\mathcal{T} \models C \equiv D$ implies $C < D$ or $D < C$ and, conversely, $C < D$ implies $\mathcal{T} \not\models C \equiv D$. Now, the minimization of a rooted CQ q that is consistent with $\mathcal{T}$ proceeds in three phases:

1. maximally extend q w.r.t. concept atoms: add $C(t)$ to q whenever $\mathcal{T}, \mathcal{A}_q \models C(t)$ and $t \in \mathsf{term}(q)$;
2. exhaustively remove redundant role atoms, in any order;
3. exhaustively remove redundant concept atoms; when two atoms $C(t)$ and $D(t)$ are both redundant, drop $D(t)$ rather than $C(t)$ whenever (i) $\mathcal{T} \models C \sqsubseteq D$, but not $\mathcal{T} \models D \sqsubseteq C$ or (ii) $\mathcal{T} \models C \equiv D$ and $C < D$.

It can be shown that applying this minimization strategy to a rooted CQ again yields a rooted CQ (or a CQ without any atoms, a special case that is easy to handle). It is easy to see that any query produced according to this strategy is $\mathcal{T}$-minimal w.r.t. set inclusion (thus Theorem 23 applies).

Example 24. *Consider $\mathcal{T} = \{A \equiv \exists r, \exists r^- \sqsubseteq B\}$, the rooted CQ $\exists y\, r(x,y) \wedge B(y)$, and $<$ with $A < \exists r$. In Step 1, we add the atoms $A(x), \exists r(x), \exists r^-(y)$. In Step 2, we drop $r(x,y)$, since this preserves equivalence under $\equiv_\mathcal{T}$. In Step 3, we remove $B(y)$ and $\exists r^-(y)$ since this again preserves equivalence under $\equiv_\mathcal{T}$. We also remove $\exists r(x)$ since $\exists r(x) \equiv_\mathcal{T} A(x)$ but $A < \exists r$. We thus obtain $A(x)$.*

We can show that with the proposed strategy, we achieve uniqueness of minimized queries as in the relational case.

Theorem 25. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox, q_1, q_2 rooted CQs such that $q_1 \equiv_\mathcal{T} q_2$ and q_1, q_2 are consistent with $\mathcal{T}$, and $\widehat{q}_1, \widehat{q}_2$ CQs obtained from q_1 and q_2 by applying the minimization strategy. Then $\widehat{q}_1$ and $\widehat{q}_2$ are isomorphic.*

To see why we need rooted CQs, consider the unrooted CQs $q_1 = \exists x\, A(x)$ and $q_2 = \exists x\, B(x)$, the TBox $\mathcal{T} = \{A \equiv \exists r, B \equiv \exists r^-\}$, and $<$ such that $A < \exists r$ and $B < \exists r^-$. Then $q_1 \equiv_\mathcal{T} q_2$, $\widehat{q}_1 = q_1$, $\widehat{q}_2 = q_2$, but q_1 and q_2 are not isomorphic.

We write q^{r} to denote the CQ obtained from q by dropping all concept atoms, i.e., only the role atoms are kept. The queries produced by the strategy are optimal regarding their relational structure in following sense.

Theorem 26. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox, q_1, q_2 rooted CQs with $q_1 \equiv_\mathcal{T} q_2$ that are consistent with $\mathcal{T}$, and let $\widehat{q}_1$ be obtained from q_1 by the minimization strategy. Then $\widehat{q}_1^{\mathsf{r}}$ is isomorphic to a subquery of q_2.*

As a consequence, the minimization strategy yields an acyclic CQ iff *any* acyclic CQ is equivalent to q w.r.t. $\mathcal{T}$. The same holds for queries of bounded treewidth or with other desirable relational properties.

Optimization in $\mathcal{EL}$

We show that, in $\mathcal{EL}$, there is a minimization strategy which is guaranteed to produce an acyclic CQ whenever the original CQ is equivalent to *any* acyclic CQ.

Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox. We introduce a slightly different notion of fork elimination than was used in Section 4. A t_1, t_2-*fork* in a CQ q consists of atoms $r(t_1, t), r(t_2, t) \in q$ or $r(t, t_1), r(t, t_2) \in q$ such that $t_1 \neq t_2$. A t_1, t_2-fork is *eliminated* by identifying t_1 and t_2. Starting with an input CQ q, our minimization strategy is to exhaustively eliminate forks such that $\mathcal{T}$-equivalence is preserved, in any order. By the following lemma and as any given CQ admits only finitely many consecutive fork eliminations, we are guaranteed to find an acyclic CQ if there is one.

Lemma 27. *Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox and q a CQ that is not acyclic, but such that $q \equiv_\mathcal{T} p$ for some acyclic CQ p. Then q contains a fork whose elimination yields q' with $q \equiv_\mathcal{T} q'$.*

In the long version, we show that, in a second step, we can further minimize the query such that it has a minimum number of variables among all equivalent queries while preserving acyclicity. This is at the expense of introducing query atoms $C(t)$ with C a subconcept of the TBox.

8 Related Work in Database Theory

Query containment is a central notion in relational database theory, due to its importance in query optimization (cf. (Abiteboul, Hull, and Vianu 1995) and references therein). Of greater relevance to the present paper is work on containment of datalog programs. Formally, a datalog program P is contained in another program P' with respect to a target predicate q if for every input database D over the (shared) input signature, the output tuples for q w.r.t. (P, D) are a subset of those obtained for (P', D). This is broadly similar to our setting, with the datalog rules playing the role of the TBox. Datalog containment was studied in the context of optimizing datalog programs by removing atoms or rules while preserving equivalence. Because equivalence of datalog programs is undecidable (Shmueli 1987), the emphasis is on sound, but incomplete approaches (Sagiv 1987). It is relevant to note that containment of *monadic* datalog programs is decidable (Cosmadakis et al. 1988), which might conceivably be used to obtain an alternative proof of decidability of some of our problems. Note, however, that any such proof would be rather involved (due to the necessity of compiling away existential restrictions), and would not yield optimal complexity results since the best-known upper bound for monadic datalog containment is 2-ExpTime (Cosmadakis et al. 1988). Finally, we note that in both the relational and deductive database settings, semantic query optimization (Chakravarthy, Grant, and Minker 1990) is based on a variant of containment that is relativized to the class of databases satisfying a given set of integrity constraints. This can be compared to our setting in which both the restricted ABox signature and the required consistency with the TBox act as constraints on the possible ABoxes.

9 Future Work

As noted earlier, using the ABox signature during optimization can lead to shorter equivalent queries. Unfortunately, the guarantees we obtained in Section 7 regarding the outcome of query minimization no longer hold in the presence of an ABox signature or two TBoxes. Thus, a relevant question for future research is the design of minimization strategies for the generalized versions of query containment.

It would be worthwhile to extend our investigation to other query languages, like unions of conjunctive queries (UCQs) or positive existential queries (PEQs). This enables further applications, such as verifying query rewritings that allow to implement CQ answering in the presence of DL TBoxes using relational database systems (Calvanese et al. 2007): given a CQ q and TBox $\mathcal{T}$ and a rewriting of q and $\mathcal{T}$ into a UCQ or PEQ q', check whether $(q, \mathcal{T}) \equiv (q', \emptyset)$, i.e., whether q' is an 'FO-rewriting' of q relative to $\mathcal{T}$.

Acknowledgements Carsten Lutz was supported by the DFG SFB/TR 8 "Spatial Cognition".

References

Abiteboul, S.; Hull, R.; and Vianu, V. 1995. *Foundations of Databases*. Addison-Wesley.

Artale, A.; Calvanese, D.; Kontchakov, R.; and Zakharyaschev, M. 2009. The DL-Lite family and relations. *J. of Artificial Intelligence Research* 36:1–69.

Baader, F.; McGuiness, D. L.; Nardi, D.; and Patel-Schneider, P., eds. 2003. *The Description Logic Handbook*. Cambridge University Press.

Baader, F.; Bienvenu, M.; Lutz, C.; and Wolter, F. 2010. Query and predicate emptiness in description logics. In *Proc. of KR*, 192–202.

Baader, F.; Brandt, S.; and Lutz, C. 2005. Pushing the $\mathcal{EL}$ envelope. In *Proc. of IJCAI*, 364–369.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2007. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. of Automated Reasoning* 39(3):385–429.

Calvanese, D.; Giacomo, G. D.; Lembo, D.; Lenzerini, M.; Poggi, A.; Rodriguez-Muro, M.; Rosati, R.; Ruzzi, M.; and Savo, D. F. 2011. The MASTRO system for ontology-based data access. *Semantic Web* 2(1):43–53.

Calvanese, D.; De Giacomo, G.; and Lenzerini, M. 1998. On the decidability of query containment under constraints. In *Proc. of PODS*, 149–158.

Calvanese, D.; Ortiz, M.; and Simkus, M. 2011. Containment of regular path queries under description logic constraints. In *Proc. of IJCAI*, 805–812.

Chakravarthy, U. S.; Grant, J.; and Minker, J. 1990. Logic-based approach to semantic query optimization. *ACM Transactions on Database Systems* 15(2):162–207.

Chandra, A. K., and Merlin, P. M. 1977. Optimal implementation of conjunctive queries in relational data bases. In *Proc. of STOC*, 77–90.

Cosmadakis, S. S.; Gaifman, H.; Kanellakis, P. C.; and Vardi, M. Y. 1988. Decidable optimization problems for database logic programs (preliminary report). In *Proc.of STOC*, 477–490.

Glimm, B.; Lutz, C.; Horrocks, I.; and Sattler, U. 2008. Conjunctive query answering for the description logic $\mathcal{SHIQ}$. *J. of Artificial Intelligence Research* 31:157–204.

Gonçalves, R. S.; Parsia, B.; and Sattler, U. 2011. Analysing the evolution of the NCI thesaurus. In *Proc. of CBMS*, 1–6.

Hemachandra, L. A. 1987. The strong exponential hierarchy collapses. In *Proc. of STOC*, 110–122.

Hustadt, U.; Motik, B.; and Sattler, U. 2005. Data complexity of reasoning in very expressive description logics. In *Proc. of IJCAI*, 466–471.

IHTSDO. 2008. *SNOMED Clinical Terms User Guide*. The International Health Terminology Standards Development Organisation (IHTSDO). Available from `http://www.ihtsdo.org/publications/snomed-docs/`.

Jimnez-Ruiz, E.; Grau, B. C.; Horrocks, I.; and Llavori, R. B. 2011. Supporting concurrent ontology development: Framework, algorithms and tool. *Data & Knowledge Engineering* 70(1):146–164.

Klein, M. C. A.; Fensel, D.; Kiryakov, A.; and Ognyanov, D. 2002. Ontology versioning and change detection on the web. In *Proc. of EKAW*, 247–259.

Konev, B.; Ludwig, M.; Walther, D.; and Wolter, F. 2011. The logical diff for the lightweight description logic $\mathcal{EL}$. Technical report, University of Liverpool, http://www.liv.ac.uk/~frank/publ/.

Konev, B.; Walther, D.; and Wolter, F. 2008. The logical difference problem for description logic terminologies. In *Proc. of IJCAR*, 259–274.

Kontchakov, R.; Lutz, C.; Toman, D.; Wolter, F.; and Zakharyaschev, M. 2010. The combined approach to query answering in DL-Lite. In *Proc. of KR*, 247–257.

Liu, H.; Lutz, C.; and Milicic, M. 2008. The projection problem for $\mathcal{EL}$ actions. In *Proc. of DL*.

Lutz, C.; Toman, D.; and Wolter, F. 2009. Conjunctive query answering in the description logic $\mathcal{EL}$ using a relational database system. In *Proc. of IJCAI*, 2070–2075.

Lutz, C. 2008. The complexity of conjunctive query answering in expressive description logics. In *Proc. of IJCAR*, 179–193.

Noy, N. F., and Musen, M. A. 2002. PromptDiff: A fixed-point algorithm for comparing ontology versions. In *Proc. of AAAI*, 744–750.

Ortiz, M.; Rudolph, S.; and Simkus, M. 2011. Query answering in the Horn fragments of the description logics $\mathcal{SHOIQ}$ and $\mathcal{SROIQ}$. In *Proc. of IJCAI*, 1039–1044.

Pérez-Urbina, H.; Horrocks, I.; and Motik, B. 2009. Efficient query answering for OWL 2. In *Proc. of ISWC*, 489–504.

Sagiv, Y., and Yannakakis, M. 1980. Equivalences among relational expressions with the union and difference operators. *J. of the ACM* 27(4):633–655.

Sagiv, Y. 1987. Optimizing datalog programs. In *Proc. of PODS*, 349–362.

Shmueli, O. 1987. Decidability and expressiveness of logic queries. In *Proc. of PODS*, 237–249.

Sioutos, N.; de Coronado, S.; Haber, M.; Hartel, F.; Shaiu, W.; and Wright, L. 2006. NCI thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. of Biomedical Informatics* 40(1):30–43.

A Proofs for Section 2

Theorem 3. Let $\mathcal{L} \in \{\mathcal{EL}, \mathcal{EL}_\bot, \mathcal{ALC}, \mathcal{ALCI}, \mathcal{ALCF}\}$ and $Q \in \{\text{CQ,IQ}\}$. Then there is a polynomial reduction of Q-containment in $\mathcal{L}$ to single TBox Q-containment in $\mathcal{L}$.

We start with the case of $\mathcal{EL}_\bot$ and CQs also considered in the proof sketch of the main paper, but without any restriction of the ABox. This also settles the case of $\mathcal{EL}$ TBoxes and CQs. Let $\mathcal{T}_1, \mathcal{T}_2$ be $\mathcal{EL}_\bot$-TBoxes, q_1, q_2 CQs of the same arity, and Σ an ABox signature. Introduce fresh concept names $s_1(A)$ and $s_2(A)$ for every concept name A and fresh role names $s_1(r)$, $s_2(r)$ for every role name $r \notin \Sigma$. We extend s_1, s_2 to concepts and CQs in the obvious way; for example, $s_1(C)$ is C with every $X \in \mathsf{N_R} \cup \mathsf{N_C}$ except role names from Σ replaced with $s_1(X)$.

The relativization C_B of an $\mathcal{EL}_\bot$-concept C to a concept name B is defined inductively as follows:

$$A_B = A \sqcap B, \quad \top_B = B, \quad \bot_B = \bot$$
$$(C \sqcap D)_B = C_B \sqcap D_B, \quad (\exists r.C)_B = B \sqcap \exists r.C_B$$

The relativization q_B of a CQ q to B adds $B(t)$ to q for every term t in q. Let E, E_1, E_2 be fresh concept names, s a fresh role name, and set $\Sigma' = \Sigma \cup \{E\}$ and

$$\begin{array}{rcl} \mathcal{T} & = & \mathcal{T}_1' \cup \mathcal{T}_2' \cup \{E \sqsubseteq E_1 \sqcap E_2, E_1 \sqsubseteq \exists s.E_2\} \\ & & \{A \sqsubseteq s_1(A) \sqcap s_2(A) \sqcap E_1 \sqcap E_2\} \end{array}$$

where $\mathcal{T}_i' = \{s_i(C)_{E_i} \sqsubseteq s_i(D)_{E_i} \mid C \sqsubseteq D \in \mathcal{T}_i\}$, for $i = 1, 2$. Intuitively, E_i represents the 'active domain' of $\mathcal{T}_i$ and E is introduced to deal with individual names that occur in the ABox, but are not used in any concept assertion. The purpose of the inclusion $E_1 \sqsubseteq \exists s.E_2$ is to ensure that E_2 is non-empty whenever E_1 is non-empty.

Lemma 28. *$(\mathcal{T}_1, q_1) \subseteq_\Sigma (\mathcal{T}_2, q_2)$ iff $s_1(q_1)_{E_1} \subseteq_{\mathcal{T}, \Sigma'} s_2(q_2)_{E_2}$.*

Proof. (1) $\Rightarrow$ (2) Assume $s_1(q_1)_{E_1} \not\subseteq_{\mathcal{T}, \Sigma'} s_2(q_2)_{E_2}$. Consider a Σ'-ABox $\mathcal{A}$ such that $\vec{a} \in \mathsf{cert}_{\mathcal{T}}(s_1(q_1)_{E_1}, \mathcal{A})$ and $\vec{a} \notin \mathsf{cert}_{\mathcal{T}}(s_2(q_2)_{E_2}, \mathcal{A})$.

Let I be the set of individuals a in $\mathcal{A}$ such that $A(a) \in \mathcal{A}$ for some $A \in \Sigma \cup \{E\}$ and let

$$\begin{array}{rcl} \mathcal{A}_0 & = & \{A(a) \mid a \in I, A \in \Sigma, A(a) \in \mathcal{A}\} \cup \\ & & \{r(a,b) \mid a, b \in I, r(a,b) \in \mathcal{A}\}. \end{array}$$

We show that (i) $\mathcal{A}_0$ is consistent w.r.t. $\mathcal{T}_1$ and $\mathcal{T}_2$, and (ii) $\vec{a}$ is a certain answer to q_1 w.r.t. $\mathcal{A}_0$ and $\mathcal{T}_1$ and (iii) $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{A}_0$ and $\mathcal{T}_2$.

For (i), note that there is a model $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$ (since $\vec{a} \notin \mathsf{cert}_{\mathcal{T}}(s_2(q_2)_{E_2}, \mathcal{A})$. Moreover $E_1^{\mathcal{I}} \neq \emptyset$, since $\vec{a} \in \mathsf{cert}_{\mathcal{T}}(s_1(q_1)_{E_1}, \mathcal{A})$. But then $E_2^{\mathcal{I}} \neq \emptyset$ since $E_1 \sqsubseteq \exists s.E_2 \in \mathcal{T}$. It follows that the restrictions of $\mathcal{I}$ to $E_1^{\mathcal{I}}$ and $E_2^{\mathcal{I}}$ are models of $\mathcal{T}_1$ and $\mathcal{A}_0$ and $\mathcal{T}_2$ and $\mathcal{A}_0$, respectively.

For (ii), let $\mathcal{I}$ be a model of $\mathcal{T}_1$ and $\mathcal{A}_0$. Assume $\mathcal{I} \not\models q_1[\vec{a}]$. We may assume that for all $a, b \in \mathsf{Ind}(\mathcal{A}_0)$ and role names r: $(a,b) \in r^{\mathcal{I}}$ iff $r(a,b) \in \mathcal{A}_0$. We construct a model $\mathcal{I}'$ of $\mathcal{T}$ and $\mathcal{A}$ with $\mathcal{I}' \not\models s_1(q_1)_{E_1}[\vec{a}]$ and thus derive a contradiction. To construct $\mathcal{I}'$ let $\mathcal{J}$ be some model of $\mathcal{T}_2$ and $\mathcal{A}_0$. Again, we may assume that for all $a, b \in \mathsf{Ind}(\mathcal{A}_0)$ and role names r: $(a,b) \in r^{\mathcal{J}}$ iff $r(a,b) \in \mathcal{A}_0$. Assume that $\Delta^{\mathcal{I}} \cap \Delta^{\mathcal{J}} = \mathsf{Ind}(\mathcal{A}_0)$. Let $\Delta^{\mathcal{I}'} = \Delta^{\mathcal{I}} \cup \Delta^{\mathcal{J}} \cup (\mathsf{Ind}(\mathcal{A}) \setminus \mathsf{Ind}(\mathcal{A}_0))$ and let

- $E^{\mathcal{I}'} = \{a \in \mathsf{Ind}(\mathcal{A}) \mid E(a) \in \mathcal{A}\}$;
- $E_1^{\mathcal{I}'} = \Delta^{\mathcal{I}}$, $E_2^{\mathcal{I}'} = \Delta^{\mathcal{J}}$;
- $r^{\mathcal{I}'} = r^{\mathcal{I}} \cup r^{\mathcal{J}}$ for all $r \in \Sigma$;
- $A^{\mathcal{I}'} = \{a \in \mathsf{Ind}(\mathcal{A}) \mid A(a) \in \mathcal{A}\}$ for all $A \in \Sigma$;
- $s_1(X)^{\mathcal{I}'} = X^{\mathcal{I}}$, for all X from $\mathcal{T}_1, q_1, \mathcal{T}_2, q_2$;
- $s_2(X)^{\mathcal{I}'} = X^{\mathcal{J}}$, for all X from $\mathcal{T}_1, q_1, \mathcal{T}_2, q_2$.
- $a^{\mathcal{I}'} = a$, for all $a \in \mathsf{Ind}(\mathcal{A})$.

It is readily checked that $\mathcal{I}$ is a model of $\mathcal{T}$ and $\mathcal{A}$. Moreover, $\mathcal{I}' \not\models s_1(q_1)_{E_1}[\vec{a}]$, as required.

For (iii), take a model $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$ such that $\mathcal{I} \not\models s_2(q_2)_{E_2}[\vec{a}]$. The argument used for (i) shows that $E_2^{\mathcal{I}} \neq \emptyset$. We may assume that $a^{\mathcal{I}} \in E_2^{\mathcal{I}}$ iff $a \in \mathsf{Ind}(\mathcal{A}_0)$. Let $\mathcal{I}'$ be the restriction of $\mathcal{I}$ to $E_2^{\mathcal{I}}$. Then $\mathcal{I}'$ is a model of $\mathcal{T}_2$ and $\mathcal{A}_0$ and $\mathcal{I}' \not\models q_2[\vec{a}]$. Thus $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{A}_0$ and $\mathcal{T}_2$, as required.

(2) $\Rightarrow$ (1) Assume $(\mathcal{T}_1, q_1) \not\subseteq_\Sigma (\mathcal{T}_2, q_2)$. Take a Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $\vec{a}$ is a certain answer to q_1 w.r.t. $\mathcal{A}$ and $\mathcal{T}_1$ and not a certain answer to q_2 w.r.t. $\mathcal{A}$ and $\mathcal{T}_2$.

Let $\mathcal{A}_0 = \mathcal{A} \cup \{E(a) \mid a \in \mathsf{Ind}(\mathcal{A})\}$. (If $\mathcal{A} = \emptyset$, we take an individual a and set $\mathcal{A}_0 = \{E(a)\}$. As this case is straightforward we consider the case $\mathcal{A} \neq \emptyset$ only). We show (i) $\vec{a}$ is a certain answer to $s_1(q_1)_{E_1}$ w.r.t. $\mathcal{A}_0$ and $\mathcal{T}$ and (ii) $\vec{a}$ is not a certain answer to $s_2(q_2)_{E_2}$ w.r.t. $\mathcal{A}_0$ and $\mathcal{T}$. For (i), the proof is by contradiction. Let $\mathcal{I}$ be a model of $\mathcal{A}_0$ and $\mathcal{T}$ such that $\mathcal{I} \not\models s_1(q_1)_{E_1}[\vec{a}]$. Let $\mathcal{I}'$ be the restriction of $\mathcal{I}$ to $E_1^{\mathcal{I}}$. Then $\mathcal{I}'$ is a model of $\mathcal{T}_1$ and $\mathcal{A}$ and $\mathcal{I}' \not\models q_2[\vec{a}]$. Thus $\vec{a}$ is not a certain answer to q_1 w.r.t. $\mathcal{A}$ and $\mathcal{T}_1$, and we have derived a contradiction.

For (ii) take a model $\mathcal{I}$ of $\mathcal{T}_2$ and $\mathcal{A}$ such that $\mathcal{I} \not\models q_2[\vec{a}]$. We may assume that for all $a, b \in \mathsf{Ind}(\mathcal{A})$ and role names r: $(a,b) \in r^{\mathcal{I}}$ iff $r(a,b) \in \mathcal{A}$. We construct a model $\mathcal{I}'$ of $\mathcal{T}$ and $\mathcal{A}_0$ with $\mathcal{I}' \not\models s_2(q_2)_{E_2}[\vec{a}]$. To construct $\mathcal{I}'$, let $\mathcal{J}$ be some model of $\mathcal{T}_1$ and $\mathcal{A}$. Again, we may assume that for all $a, b \in \mathsf{Ind}(\mathcal{A})$ and role names r: $(a,b) \in r^{\mathcal{J}}$ iff $r(a,b) \in \mathcal{A}$. Assume that $\Delta^{\mathcal{I}} \cap \Delta^{\mathcal{J}} = \mathsf{Ind}(\mathcal{A})$. Let $\Delta^{\mathcal{I}'} = \Delta^{\mathcal{I}} \cup \Delta^{\mathcal{J}}$ and let

- $E^{\mathcal{I}'} = \mathsf{Ind}(\mathcal{A})$;
- $E_1^{\mathcal{I}'} = \Delta^{\mathcal{J}}$, $E_2^{\mathcal{I}'} = \Delta^{\mathcal{I}}$;
- $r^{\mathcal{I}'} = r^{\mathcal{I}} \cup r^{\mathcal{J}}$ for all $r \in \Sigma$;
- $A^{\mathcal{I}'} = \{a \in \mathsf{Ind}(\mathcal{A}) \mid A(a) \in \mathcal{A}\}$ for all $A \in \Sigma$;
- $s_1(X)^{\mathcal{I}'} = X^{\mathcal{J}}$, for all X from $\mathcal{T}_1, q_1, \mathcal{T}_2, q_2$;
- $s_2(X)^{\mathcal{I}'} = X^{\mathcal{I}}$, for all X from $\mathcal{T}_1, q_1, \mathcal{T}_2, q_2$.
- $a^{\mathcal{I}'} = a$, for all $a \in \mathsf{Ind}(\mathcal{A})$.

It is readily checked that $\mathcal{I}$ is a model of $\mathcal{T}$ and $\mathcal{A}_0$. Moreover, $\mathcal{I}' \not\models s_2(q_2)_{E_2}[\vec{a}]$, as required. ❏

We now consider $\mathcal{EL}$ for IQs. Observe that the relativization of an IQ is not an IQ again, but a query of the form $A_1(x) \wedge A_2(x)$. Now, by adding to the TBox inclusions of the form $A \equiv A_1 \sqcap A_2$ one can readily show that single TBox containment in $\mathcal{EL}$ for CQs of the form $A_1(x) \wedge A_2(x)$ is polynomially equivalent to single TBox IQ-containment in $\mathcal{EL}$. The same holds for $\mathcal{EL}_\bot$ and $\mathcal{ALC}$, $\mathcal{ALCI}$, and $\mathcal{ALCF}$.

For $\mathcal{ALC}$, $\mathcal{ALCI}$, and $\mathcal{ALCF}$, the proof is the same except that we must extend the definition of relativized concepts C_B to $\mathcal{ALCI}$-concepts. We do this by setting

$$(\neg C)_B = (B \sqcap \neg C_B), \quad (\forall r.C)_B = B \sqcap \forall.(\neg B \sqcup C_B)$$

Finally we note that for acyclic $\mathcal{EL}$, we can define the TBoxes $\mathcal{T}_i'$ as follows:

$$\mathcal{T}_i' = \{s_i(A) \equiv s_i(D)_{E_i} \mid A \equiv D \in \mathcal{T}_i\}$$

which preserves acyclicity.

Theorem 4 Let $\mathcal{L} \in \{\mathcal{EL}_\bot, \text{DL-Lite}_{\mathsf{core}}, \text{DL-Lite}_{\mathsf{horn}}, \mathcal{ALC}, \mathcal{ALCI}, \mathcal{ALCF}\}$ and $Q \in \{\text{CQ},\text{IQ}\}$. Then Q-containment in $\mathcal{L}$ can be polynomially reduced to full signature Q-containment in $\mathcal{L}$.

Proof. Assume TBoxes $\mathcal{T}_1, \mathcal{T}_2$, queries q_1, q_2 of the same arity, and a signature Σ are given. We can assume $\mathsf{sig}(\mathcal{T}_1, q_1) \cap \mathsf{sig}(\mathcal{T}_2, q_2) \subseteq \Sigma$. We define

$$\begin{aligned}\mathcal{T}_2' \;=\; & \mathcal{T}_2 \cup \{A \sqsubseteq \bot \mid A \in \mathsf{sig}(\mathcal{T}_1, q_1) \setminus \Sigma\} \cup \\ & \{\exists r.\top \sqsubseteq \bot \mid r \in \mathsf{sig}(\mathcal{T}_1, q_1) \setminus \Sigma\}\end{aligned}$$

If $\mathcal{T}$ is in any of the languages $\mathcal{EL}_\bot$, DL-Lite$_{\mathsf{core}}$, DL-Lite$_{\mathsf{horn}}$, $\mathcal{ALC}$, $\mathcal{ALCF}$, then $\mathcal{T}_2'$ is in the same language.

Claim. $(\mathcal{T}_1, q_1) \subseteq_\Sigma (\mathcal{T}_2, q_2)$ iff $(\mathcal{T}_1, q_1) \subseteq (\mathcal{T}_2', q_2)$

($\Rightarrow$). Assume there exists an ABox $\mathcal{A}$ such that $\vec{a}$ is a certain answer to q_1 w.r.t. $\mathcal{T}_1$ and $\mathcal{A}$ but $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{T}_2'$ and $\mathcal{A}$. Then $\mathcal{A}$ is consistent w.r.t. $\mathcal{T}_2'$. Hence the signature of $\mathcal{A}$ does not contain any non-Σ symbols from $\mathcal{T}_1, q_1$. Denote by $\mathcal{A}'$ the ABox obtained from $\mathcal{A}$ by removing all assertions containing non-Σ-symbols. Clearly, $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{T}_2$ and $\mathcal{A}'$. Moreover, $\vec{a}$ is still a certain answer to q_1 w.r.t. $\mathcal{T}_1$ and $\mathcal{A}'$, as required.

($\Leftarrow$) Assume there exists an Σ-ABox $\mathcal{A}$ such that $\vec{a}$ is a certain answer to q_1 w.r.t. $\mathcal{T}_1$ and $\mathcal{A}$ but $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{T}_2$ and $\mathcal{A}$. Then $\vec{a}$ is not a certain answer to q_2 w.r.t. $\mathcal{T}_2'$ and $\mathcal{A}$, and we are done. ❑

Lemma 6. Let q be a CQ, $\mathcal{T}$ an $\mathcal{ALCFI}$-TBox, Σ an ABox signature, A a concept name A that does not occur in q, Σ, nor $\mathcal{T}$, and q_A any query with the same arity as q that uses A. Then there exists a Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$ with $\mathsf{cert}_\mathcal{T}(q, \mathcal{A}) \neq \emptyset$ iff $q \not\subseteq_{\mathcal{T},\Sigma} q_A$.

Proof. For the first direction, suppose that $\mathcal{A}$ is a Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$ and such that $\mathsf{cert}_\mathcal{T}(q, \mathcal{A}) \neq \emptyset$. Let $\vec{a}$ be such that $\mathcal{T}, \mathcal{A} \models q(\vec{a})$. Take any model $\mathcal{I}$ of $\mathcal{A}$ and $\mathcal{T}$. Then the interpretation $\mathcal{I}'$ obtained from $\mathcal{I}$ by interpreting A as $\emptyset$ is still a model of $\mathcal{A}$ and $\mathcal{T}$. Moreover, $\mathcal{I}' \not\models q_A(\vec{a})$. Hence, $\mathcal{A}$ witnesses $q \not\subseteq_{\mathcal{T},\Sigma} q_A$.

For the second direction, suppose $q \not\subseteq_{\mathcal{T},\Sigma} q_A$. It follows that $\mathsf{cert}_\mathcal{T}(q, \mathcal{A}) \neq \emptyset$, since otherwise, $q \subseteq_{\mathcal{T},\Sigma} q_A$ would hold trivially. ❑

B Proofs for Section 3

Throughout the section, we will use R (possibly with subscripts) to stand for either a role or an inverse role, and we will use R^- to denote s^- if $R = s$ and s if $R = s^-$ ($s \in \mathsf{N_R}$). Likewise, we use $R(a, b)$ to refer to the assertion $s(a, b)$ if $R = s$ and $s(b, a)$ if $R = s^-$.

We begin by recalling the definition of canonical models for DL-Lite, as it will prove useful in some of the proofs. To construct the *canonical model*, $\mathcal{I}_{\mathcal{T},\mathcal{A}}$, for a DL-Lite$_{\mathsf{horn}}$ TBox $\mathcal{T}$ and ABox $\mathcal{A}$, we start with $\mathcal{A}$ and then exhaustively apply the CIs from $\mathcal{T}$, always introducing *new* elements in role domain and ranges if necessary. Formally, the domain of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ consists of *paths* of the form $ac_{R_1} \cdots c_{R_n}$, $n \geq 0$, such that $a \in \mathsf{Ind}(\mathcal{A})$, each $R_i \in \mathsf{N_R} \cup \{r^- \mid r \in \mathsf{N_R}\}$, and the following conditions hold:

(agen) $\mathcal{T}, \mathcal{A} \models \exists R_1(a)$ but $R_1(a, b) \notin \mathcal{A}$ for all $b \in \mathsf{Ind}(\mathcal{A})$, in which case we write $a \rightsquigarrow c_{R_1}$;

(rgen) for $i < n$, $\mathcal{T} \models \exists R_i^- \sqsubseteq \exists R_{i+1}$ and $R_i^- \neq R_{i+1}$, in which case we write $c_{R_i} \rightsquigarrow c_{R_{i+1}}$.

We denote the last element in a path σ by $\mathsf{tail}(\sigma)$, and define $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ by taking:

$$\begin{aligned}\Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} &= \{ac_{R_1} \cdots c_{R_n} \mid a \in \mathsf{Ind}(\mathcal{A}), a \rightsquigarrow c_{R_1} \rightsquigarrow \cdots \rightsquigarrow c_{R_n}\}, \\ a^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} &= a, \text{ for } a \in \mathsf{Ind}(\mathcal{A}), \\ A^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} &= \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{A}, \mathcal{T} \models A(a)\} \cup \\ &\quad \{\sigma c_R \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \mid \mathcal{T} \models \exists R^- \sqsubseteq A\}, \\ r^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} &= \{(a, b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mid r(a, b) \in \mathcal{A}\} \cup \\ &\quad \{(\sigma, \sigma c_r) \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \times \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \mid \mathsf{tail}(\sigma) \rightsquigarrow c_r\} \cup \\ &\quad \{(\sigma c_{r^-}, \sigma) \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \times \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \mid \mathsf{tail}(\sigma) \rightsquigarrow c_{r^-}\}.\end{aligned}$$

It is standard to show the following:

Lemma 29. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{horn}}$ TBox, and let $\mathcal{A}$ be an ABox consistent with $\mathcal{T}$. For any k-ary conjunctive query q and k-tuple $\vec{a}$ of individuals from $\mathsf{Ind}(\mathcal{A})$, $\mathcal{I}_{\mathcal{T},\mathcal{A}} \models q[\vec{a}]$ iff $\vec{a} \in \mathsf{cert}_\mathcal{T}(q, \mathcal{A})$.*

We now complete the proofs which were only sketched in Section 3.

Theorem 7. IQ-containment in DL-Lite$_{\mathsf{core}}$ is in PTIME.

Proof. We prove the following:

Claim. $(\mathcal{T}_1, A(x)) \subseteq_\Sigma (\mathcal{T}_2, B(x))$ if and only if for every $C \in (\mathsf{N_C} \cap \Sigma) \cup \{\exists r, \exists r^- \mid r \in \mathsf{N_R} \cap \Sigma\}$ we have that $\mathcal{T}_1 \models C \sqsubseteq A$ implies $\mathcal{T}_2 \models C \sqsubseteq B$.

($\Rightarrow$). Suppose $(\mathcal{T}_1, A(x)) \subseteq_\Sigma (\mathcal{T}_2, B(x))$. Let $C \in (\mathsf{N_C} \cap \Sigma) \cup \{\exists r, \exists r^- \mid r \in \mathsf{N_R} \cap \Sigma\}$ be such that $\mathcal{T}_1 \models C \sqsubseteq A$. We define a Σ-ABox $\mathcal{A}$ which realizes $C(a)$ as follows. If $C \in \mathsf{N_C}$ then $\mathcal{A} = \{C(a)\}$. If $C = \exists R$, then $\mathcal{A} = \{R(a, b)\}$. Then $\mathcal{T}_1, \mathcal{A} \models A(a)$, and since $(\mathcal{T}_1, A(x)) \subseteq_\Sigma (\mathcal{T}_2, B(x))$, we must also have $\mathcal{T}_2, \mathcal{A} \models B(a)$. Since $\mathcal{A}$ only makes C true at a, it follows that $\mathcal{T}_2 \models C \sqsubseteq B$.

($\Leftarrow$). Suppose that for every $C \in (\mathsf{N_C} \cap \Sigma) \cup \{\exists r, \exists r^- \mid r \in \mathsf{N_R} \cap \Sigma\}$, $\mathcal{T}_1 \models C \sqsubseteq A$ implies $\mathcal{T}_2 \models C \sqsubseteq B$. Consider a Σ-ABox $\mathcal{A}$ consistent with $\mathcal{T}_1$ and $\mathcal{T}_2$ such that $\mathcal{T}_1, \mathcal{A} \models A(a)$.

It is a well-known property of DL-Lite$_{\mathsf{core}}$ that there must exist a single assertion $\alpha \in \mathcal{A}$ such that $\{\alpha\} \models C(a)$ for some concept C such that $\mathcal{T}_1 \models C \sqsubseteq A$. Since C must involve a concept or role name from Σ, our earlier assumption yields $\mathcal{T}_2 \models C \sqsubseteq B$. Hence, $\mathcal{T}_2, \mathcal{A} \models B(a)$.

Finally, to complete the proof, we recall that subsumption in DL-Lite$_{\mathsf{core}}$ is tractable, and so the property described in the claim can be verified in polynomial time. ❑

Theorem 8. In DL-Lite$_{\mathsf{horn}}$, IQ-containment is coNP-complete and CQ-containment is in Π_2^p.

Proof. We detail the proof of the upper bound for CQ-containment. A *witness* for $(\mathcal{T}_1, q_1) \not\subseteq_\Sigma (\mathcal{T}_2, q_2)$ consists of a Σ-ABox $\mathcal{A}$, a tuple of individuals $\vec{a}$ from $\mathsf{Ind}(\mathcal{A})$, and a mapping $\pi : \mathsf{vars}(q_1) \to \Delta^{\mathcal{I}_{\mathcal{T}_1,\mathcal{A}}}$ such that (i) $\mathcal{A}$ is consistent with $\mathcal{T}_1$ and such that $\mathcal{T}_2, \mathcal{A} \not\models q[\vec{a}]$, and (ii) π defines a $\vec{a}$-match of q_1 in the canonical model for $\mathcal{A}$ and $\mathcal{T}_1$. By the following claim, it suffices to consider witnesses with $|\mathsf{Ind}(\mathcal{A})| \leq |\mathsf{term}(q)| \times |\Sigma|$.

Claim 1. Let $\mathcal{T}, \mathcal{A} \models q[\vec{a}]$ for a Σ-ABox $\mathcal{A}$ such that $\mathcal{T}, \mathcal{A}$ is consistent. Then there exists $\mathcal{A}' \subseteq \mathcal{A}$ such that $|\mathsf{Ind}(\mathcal{A}')| \leq |\mathsf{term}(q)| \times |\Sigma|$ and $\mathcal{T}, \mathcal{A}' \models q[\vec{a}]$.

To prove Claim 1, suppose $\mathcal{T}, \mathcal{A} \models q[\vec{a}]$ for a Σ-ABox $\mathcal{A}$ such that $\mathcal{T}, \mathcal{A}$ is consistent. Then, by Lemma 29, there exists a match π of q in $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ such that $\pi(v_i) = a_i$. Let $\mathcal{F}$ be the set of all a such that there exists $a \cdot w$ in the range of π (where w can be the empty word). Now let $\mathcal{A}'$ be a subset of $\mathcal{A}$ consisting of:

- $B(a) \in \mathcal{A}$ such that $a \in \mathcal{F}$;
- $r(a, b) \in \mathcal{A}$ such that $a, b \in \mathcal{F}$;
- some assertion $r(a, b) \in \mathcal{A}$, if $a \in \mathcal{F}$ and $\mathcal{A}$ contains an assertion of such a form;
- some assertion $r(a, b) \in \mathcal{A}$, if $b \in \mathcal{F}$ and $\mathcal{A}$ contains an assertion of such a form.

By construction, $\mathcal{A}'$ contains at most $|\mathsf{term}(q)| \times |\Sigma|$ assertions. An inspection of the canonical model construction shows that $\mathcal{I}_{\mathcal{T},\mathcal{A}'}$ contains all $\pi(t)$, $t \in \mathsf{term}(q)$, and that $\pi(t) \in A^{\mathcal{I}_{\mathcal{T},\mathcal{A}'}}$ iff $\pi(t) \in A^{\mathcal{I}_{\mathcal{T},\mathcal{A}'}}$ and $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{\mathcal{T},\mathcal{A}'}}$ iff $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ for all $t, t_1, t_2 \in \mathsf{term}(q)$, all concept names A, and all roles r. Thus, $\mathcal{I}_{\mathcal{T},\mathcal{A}'} \models q[\vec{a}]$. By Lemma 29, $(\mathcal{T}, \mathcal{A}') \models q[\vec{a}]$, as required.

We now establish a further claim in order to ensure that we can find always find an $\vec{a}$-match of q_1 which involves only a polynomial-size portion of the canonical model.

Claim 2. Let $\mathcal{T}, \mathcal{A} \models q[\vec{a}]$ for a Σ-ABox $\mathcal{A}$ such that $\mathcal{T}, \mathcal{A}$ is consistent. Then there exists an $\vec{a}$-match π of q in $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ such that every path in the range of π has length at most $|\mathcal{T}| + |\mathsf{term}(q)| + 1$.

To show Claim 2, consider an arbitrary $\vec{a}$-match π for q in $\mathcal{I}_{\mathcal{T},\mathcal{A}}$. Let P_π be the set of paths in the range of π, and let P'_π be the set of paths in P which have no proper prefix in P. By definition, for every path σ in P_π, there is a path $\sigma' \in P'_\pi$ which is a prefix of σ. Moreover, all paths σ'' which have σ' as prefix and are prefixes of σ must belong to P_π. Hence the maximal length of any path in P_π cannot exceed by more than $|\mathsf{term}(q)|$ the maximal length of a path in P'_π . It thus suffices to show that we can find a match π such that the paths in P'_π have length at most $|\mathcal{T}| + 1$. For this, we use the fact that $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ is regular in the following sense: if σ and σ' are paths such that $\mathsf{tail}(\sigma) = \mathsf{tail}(\sigma')$, then the submodel of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ generated by σ is precisely the the submodel of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ generated by σ'. Now consider some $\pi(t) = ac_{R_1} \cdots c_{R_n} \in P'_\pi$. If $n > |\mathcal{T}|$, then there exists some $1 \leq i < j \leq n$ such that $R_i = R_j$, and so, by the above property, the submodels of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ generated by $ac_{R_1} \dots c_{R_i}$ and by $ac_{R_1} \dots c_{R_j}$ are identical. Hence one can alter π by replacing every path $ac_{R_1} \cdots c_{R_n} \cdot w$ with the (shorter) path $ac_{R_1} \dots c_{R_i} c_{R_{j+1}} \dots c_{R_n} \cdot w$, without losing the property that π is an $\vec{a}$-match of q. By iterating this operation, we ensure that every path in P'_π has length at most $|\mathcal{T}| + 1$. This completes the proof of Claim 2.

Now to decide CQ non-containment, we first guess a potential witness $(\mathcal{A}, \vec{a}, \pi)$ with $|\mathsf{Ind}(\mathcal{A})| \leq |\mathsf{term}(q)| \times |\Sigma|$ and such that all paths in the range of π have length at most $|\mathcal{T}| + |\mathsf{term}(q_1)| + 1$. To verify that this potential witness satisfies the required conditions, we first check in polynomial time that $\mathcal{A}_1$ is consistent with $\mathcal{T}$ and that the mapping π defines a $\vec{a}$-match for q_1 in the canonical model of $\mathcal{T}_1, \mathcal{A}$. Then we can verify in coNP that $\mathcal{T}_2, \mathcal{A} \not\models q_2(\vec{a})$. If the guessed witness is valid, this proves $(\mathcal{T}_1, q_1) \not\subseteq_\Sigma (\mathcal{T}_2, q_2)$, and if no valid witness of the required size exists, then by the above claims, there is no counterexample to $(\mathcal{T}_1, q_1) \subseteq_\Sigma (\mathcal{T}_2, q_2)$, so containment holds. ❑

To prove Theorem 9, we show the following:

Lemma 30. *$\forall\vec{x}\exists\vec{y}\,\varphi(\vec{x}, \vec{y})$ is valid iff $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$.*

Proof. First assume that $\forall\vec{x}\exists\vec{y}\,\varphi(\vec{x}, \vec{y})$ is valid. Take a Σ-ABox $\mathcal{A}$ such that $\mathcal{T}, \mathcal{A} \models q_1$. Take the following model $\mathcal{I}$ of $\mathcal{A}$ and $\mathcal{T}$:

$$
\begin{array}{lcll}
\Delta^{\mathcal{I}} & = & \mathsf{Ind}(\mathcal{A}) & \\
X^{\mathcal{I}} & = & \{a \mid X(a) \in \mathcal{A}\} & \text{for } X \in \{T, F\} \\
V^{\mathcal{I}} & = & T^{\mathcal{I}} \cup F^{\mathcal{I}} & \\
r_v^{\mathcal{I}} & = & \{(a, b) \mid r_v(a, b) \in \mathcal{A}\} & \text{for all } v \in \vec{x} \cup \vec{y} \\
a^{\mathcal{I}} & = & a & \text{for all } a \in \mathsf{Ind}(\mathcal{A})
\end{array}
$$

Let π be a match of q_1 in $\mathcal{I}$. Define an assignment f of a truth value to each universal variable x as follows. Since q_1 contains $V(u_x)$, we must have $\pi(u_x) \in V^{\mathcal{I}}$. By definition of $\mathcal{I}$, this implies that $T(\pi(u_x)) \in \mathcal{A}$ or $F(\pi(u_x)) \in \mathcal{A}$. We set $f(x)$ to true in the former case and to false in the latter case. Ties are broken arbitrarily. Since $\forall\vec{x}\exists\vec{y}\,\varphi(\vec{x}, \vec{y})$ is valid, we can extend f to the existential variables such that φ is satisfied. We define a match of q_2 in $\mathcal{I}$ as follows.

- for each clause c in φ, there is a literal ℓ in c that evaluates to true under f. Map
  - v_c to $\pi(u_{c\ell})^{\mathcal{I}}$;
  - if $\ell = x$ or $\ell = \neg x$ with x an universal variable, then map v_{cx} to $\pi(u_x)^{\mathcal{I}}$;

- if x is an universal variable that occurs positively in c and is distinct from the variable in ℓ, then map v_{cx} to $\pi(u_T^\forall)^\mathcal{I}$;
- if x is an universal variable that occurs negatively in c and is distinct from the variable in ℓ, then map v_{cx} to $\pi(u_F^\forall)^\mathcal{I}$;

- each existential variable y with $f(y) = \mathsf{true}$ is mapped to $\pi(u_T^\exists)^\mathcal{I}$;
- each existential variable y with $f(y) = \mathsf{false}$ is mapped to $\pi(u_F^\exists)^\mathcal{I}$.

By going through the atoms in q_2, it can be checked that the above indeed defines a match of q_2 in $\mathcal{I}$. By construction of $\mathcal{I}$, this model can be homomorphically embedded into any model of $\mathcal{A}$ and $\mathcal{T}$. It follows that q_2 has a match in each such model, so $\mathcal{T}, \mathcal{A} \models q_2$.

For the converse direction, assume that $q_1 \subseteq_{\mathcal{T},\Sigma} q_2$ and let f be a truth assignment for the universal variables. Define an ABox $\mathcal{A}$ that consists of the following assertions:

- $T(a_x)$ for each universal variable x with $f(x) = \mathsf{true}$;
- $F(a_x)$ for each universal variable x with $f(x) = \mathsf{false}$;
- $F(a_F^\forall), T(a_T^\forall)$;
- for each clause c in φ and each literal ℓ in c, the following assertions:
  - $A_c(a_{c\ell})$;
  - $r_x(a_{c\ell}, a_x)$ if $\ell = x$ or $\ell = \neg x$, with x an universal variable;
  - $r_y(a_{c\ell}, a_T^\exists)$ if $\ell = y$ is a existential variable;
  - $r_y(a_{c\ell}, a_F^\exists)$ if $\ell = \neg y$, with y a existential variable;
  - for each universal variable x different from the variable in ℓ, the assertions $r_x(a_{c\ell}, a_F^\forall)$ and $r_x(a_{c\ell}, a_T^\forall)$;
  - for each existential variable y different from the variable in ℓ, the assertions $r_y(a_{c\ell}, a_F^\exists)$ and $r_y(a_{c\ell}, a_T^\exists)$.

It can be verified that $\mathcal{T}, \mathcal{A} \models q_1$, by mapping each variable in q_1 to the corresponding individual name in $\mathcal{A}$. Thus, $\mathcal{T}, \mathcal{A} \models q_2$. Let $\mathcal{I}$ be the model of $\mathcal{T}$ and $\mathcal{A}$ that is defined as in the other direction of this proof. We have $\mathcal{I} \models q_2$, thus there is a match π of q_2 in $\mathcal{I}$. Extend f to the existential variables by setting $f(y) = \mathsf{true}$ if $\pi(v_y) = a_T^\exists$ and $f(y) = \mathsf{false}$ if $\pi(v_y) = a_F^\exists$ (due to the atom $r_y(v_c, v_y)$ in q_2 and the construction of $\mathcal{A}$, there are no other choices for $\pi(v_y)$). It remains to show that f satisfies each clause c in φ. Fix a c. Due to the use of the concept name A_c in $\mathcal{A}$ and q_2, we have $\pi(v_c) = a_{c\ell}$ for some literal ℓ in c. It suffices to show that f satisfies ℓ. Distinguish the following cases:

- $\ell = x$ an universal variable.

  Then $r_x(v_c, v_{cx}) \in q_2$ and $T(v_{cx}) \in q_2$. The former and the definition of $\mathcal{A}$ yields $\pi(v_{cx}) = a_x$. The latter yields $T(a_x) \in \mathcal{A}$, thus $f(x) = \mathsf{true}$ and f satisfies ℓ.
- $\ell = \neg x$ with x an universal variable.

  Then $r_x(v_c, v_{cx}) \in q_2$ and $F(v_{cx}) \in q_2$. The former and the definition of $\mathcal{A}$ yields $\pi(v_{cx}) = a_x$. The latter yields $F(a_x) \in \mathcal{A}$, thus $f(x) = \mathsf{false}$ and f satisfies ℓ.
- $\ell = y$ a existential variable.

  By construction of $\mathcal{A}$, since $\pi(v_c) = a_{c\ell}$ and due to the role atom $r_y(v_c, v_y)$ in q_2, we must have $\pi(v_y) = u_T^\exists$. By definition of f, $f(y) = \mathsf{true}$ and thus f satisfies ℓ.
- $\ell = \neg y$, with y a existential variable.

  By construction of $\mathcal{A}$, since $\pi(v_c) = a_{c\ell}$ and due to the role atom $r_y(v_c, v_y)$ in q_2, we must have $\pi(v_y) = u_F^\exists$. By definition of f, $f(y) = \mathsf{false}$ and thus f satisfies ℓ.

❑

C Proofs for Section 4

The notion of *canonical models* will play an important role in the proofs in this section, so we recall the definition here. Let $\mathcal{T}$ be an $\mathcal{EL}_\bot$-TBox and $\mathcal{A}$ a (possibly infinite) ABox that is consistent w.r.t. $\mathcal{T}$. For $a \in \mathsf{Ind}(\mathcal{A})$, a *path* for $\mathcal{A}$ and $\mathcal{T}$ is a finite sequence $a\, r_1\, C_1\, r_2 C_2 \cdots r_n\, C_n$, $n \geq 0$, where the C_i are concepts that occur in $\mathcal{T}$ (potentially as a subconcept) and the r_i are roles such that the following conditions are satisfied:

- $a \in \mathsf{Ind}(\mathcal{A})$,
- $\mathcal{T}, \mathcal{A} \models \exists r_1.C_1(a)$ if $n \geq 1$,
- $\mathcal{T} \models C_i \sqsubseteq \exists r_{i+1}.C_{i+1}$ for $1 \leq i < n$.

The domain $\Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ of the *canonical model* $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ for $\mathcal{T}$ and $\mathcal{A}$ is the set of all paths for $\mathcal{A}$ and $\mathcal{T}$. If $p \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \setminus \mathsf{Ind}(\mathcal{A})$, then $\mathsf{tail}(p)$ denotes the last concept C_n in p. Set

$$
\begin{array}{lcl}
A^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} & := & \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{T}, \mathcal{A} \models A(a)\} \cup \\
 & & \{p \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \setminus \mathsf{Ind}(\mathcal{A}) \mid \mathcal{T} \models \mathsf{tail}(p) \sqsubseteq A\} \\
r^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} & := & \{(a,b) \mid r(a,b) \in \mathcal{A}\} \cup \\
 & & \{(p,q) \in \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \times \Delta^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \mid \\
 & & \qquad q = p \cdot r\, C \text{ for some concept } C\} \\
a^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} & := & a \text{ for all } a \in \mathsf{Ind}(\mathcal{A})
\end{array}
$$

It is standard to prove the following.

Lemma 31. *$\mathcal{I}_{\mathcal{T},\mathcal{A}}$ is a model of $\mathcal{T}$ and $\mathcal{A}$ such that:*

1. *for any $a \in \mathsf{Ind}(\mathcal{A})$ and $\mathcal{EL}_\bot$-concept C, $a^{\mathcal{I}_{\mathcal{T},\mathcal{A}}} \in C^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$ iff $\mathcal{T}, \mathcal{A} \models C(a)$;*
2. *for any k-ary conjunctive query q and $(a_1, \ldots, a_k) \in \mathsf{N}_\mathsf{I}^k$, $\mathcal{I}_{\mathcal{T},\mathcal{A}} \models q[a_1, \ldots, a_k]$ iff $(a_1, \ldots, a_k) \in \mathsf{cert}_{\mathcal{T},\mathcal{A}}(q)$.*

Theorem 10. IQ-containment in $\mathcal{EL}$ is EXPTIME-hard.

Proof. The proof is by reduction of instance query emptiness in $\mathcal{EL}_\bot$, proved EXPTIME-hard in (Baader et al. 2010).

Claim. $A(x)$ is Σ-empty w.r.t. $\mathcal{T}$ iff $A(x) \subseteq_{\mathcal{T}',\Sigma} B(x)$ where B is a fresh concept name and $\mathcal{T}'$ is obtained from $\mathcal{T}$ by

1. replacing every assertion $C \sqsubseteq \bot$ in $\mathcal{T}$ with $C \sqsubseteq B$ and
2. adding $\exists r.B \sqsubseteq B$ for every role r in $\mathcal{T}$ and Σ.

"if". Assume that $A(x)$ is not Σ-empty w.r.t. $\mathcal{T}$, i.e., there is a Σ-ABox $\mathcal{A}$ and an $a \in \mathsf{Ind}(\mathcal{A})$ such that $\mathcal{A}$ is consistent w.r.t. $\mathcal{T}$ and $\mathcal{T}, \mathcal{A} \models A(a)$. We can assume w.l.o.g. that

($*$) every $b \in \mathsf{Ind}(\mathcal{A})$ is reachable from a in $\mathcal{A}$, i.e, there are $r_1(a_0, a_1), \ldots, r_n(a_{n-1}, a_n) \in \mathcal{A}$ such that $a = a_0$ and $a_n = b$.

Indeed, all unreachable ABox individuals can simply be dropped from $\mathcal{A}$. Every model $\mathcal{I}$ of $\mathcal{A}$ and $\mathcal{T}$ can be converted into a model of $\mathcal{A}$ and $\mathcal{T}'$ by setting $B^{\mathcal{I}} = \emptyset$. Thus, $\mathcal{T}', \mathcal{A} \not\models B(a)$. To prove that $A(x) \not\subseteq_{\mathcal{T}',\Sigma} B(x)$ as required, it thus remains to show that $\mathcal{T}', \mathcal{A} \models A(a)$. Suppose that the contrary is true. Then $a^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}} \notin A^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}}$, where $\mathcal{I}_{\mathcal{T}',\mathcal{A}}$ is the canonical model of $\mathcal{T}'$ and $\mathcal{A}$. Since $\mathcal{T}', \mathcal{A} \not\models B(a)$, we also have $a^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}} \notin B^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}}$. By $(*)$ and construction of $\mathcal{T}'$ and $\mathcal{I}_{\mathcal{T}',\mathcal{A}}$, this means that $C^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}} = \emptyset$ whenever $C \sqsubseteq \bot \in \mathcal{T}$. It follows that $\mathcal{I}^{\mathcal{T}',\mathcal{A}}$ is a model of $\mathcal{T}$, in contradiction to $\mathcal{T}, \mathcal{A} \models A(a)$.

"only if". Assume that $A(x) \not\subseteq_{\mathcal{T}',\Sigma} B(x)$ and let $\mathcal{A}$ be a Σ-ABox and $a \in \mathsf{Ind}(\mathcal{A})$ such that $\mathcal{T}', \mathcal{A} \models A(a)$ and $\mathcal{T}', \mathcal{A} \not\models B(a)$ (consistency trivially holds since $\mathcal{T}'$ does not contain $\bot$). As above, we can w.l.o.g. assume $(*)$. Since $\mathcal{T}', \mathcal{A} \not\models B(a)$, we must have $a^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}} \notin B^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}}$. By $(*)$ and construction of $\mathcal{T}'$ and $\mathcal{I}_{\mathcal{T}',\mathcal{A}}$, this means that $C^{\mathcal{I}_{\mathcal{T}',\mathcal{A}}} = \emptyset$ whenever $C \sqsubseteq \bot \in \mathcal{T}$. It follows that $\mathcal{I}^{\mathcal{T}',\mathcal{A}}$ is a model of $\mathcal{T}$. Thus, $\mathcal{A}$ is consistent w.r.t. $\mathcal{T}$ and $\mathcal{T}, \mathcal{A} \not\models B(a)$. It remains to show $\mathcal{T}, \mathcal{A} \models A(a)$, which follows from the fact that every model $\mathcal{I}$ of $\mathcal{A}$ and $\mathcal{T}$ can be converted into a model of $\mathcal{A}$ and $\mathcal{T}'$ by setting $B^{\mathcal{I}} = \emptyset$. ❑

Theorem 11. In $\mathcal{EL}$, IQ-containment can be polynomially reduced to full signature IQ-containment.

Proof. We aim to decide $(\mathcal{T}_1, A(x)) \subseteq_\Sigma (\mathcal{T}_2, B(x))$. We can assume without loss of generality that $\mathsf{sig}(\mathcal{T}_1, q_1) \cap \mathsf{sig}(\mathcal{T}_2, q_2) \subseteq \Sigma$. We define

$$\begin{aligned}\mathcal{T}_2' \quad = \quad & \mathcal{T}_2 \cup \{A \sqsubseteq X \mid A \in \mathsf{sig}(\mathcal{T}_1, q_1) \setminus \Sigma\} \cup \\ & \{\exists r.\top \sqsubseteq X \mid r \in \mathsf{sig}(\mathcal{T}_1, q_1) \setminus \Sigma\} \cup \\ & \{\exists r.X \sqsubseteq X \mid r \in \mathsf{sig}(\mathcal{T}_1, q_1)\} \cup \{X \sqsubseteq B\}\end{aligned}$$

where X is a fresh concept name.

Claim. $(\mathcal{T}_1, q_1) \subseteq_\Sigma (\mathcal{T}_2, q_2)$ iff $(\mathcal{T}_1, q_1) \subseteq (\mathcal{T}_2', q_2)$

$(\Rightarrow)$. Assume there exists an ABox $\mathcal{A}$ and individual a such that $\mathcal{T}_1, \mathcal{A} \models A(a)$ and $\mathcal{T}_2', \mathcal{A} \not\models B(a)$. We can assume w.l.o.g. that all individuals from $\mathcal{A}$ are reachable from a (cf. proof of Theorem 10 for a formal definition of reachability). It follows from the fact that $\mathcal{T}_2', \mathcal{A} \not\models B(a)$ and the structure of $\mathcal{T}_2'$ that $\mathcal{A}$ does not contain any non-Σ symbols from $\mathsf{sig}(\mathcal{T}_1, q_1)$. Denote by $\mathcal{A}'$ the Σ-ABox obtained from $\mathcal{A}$ by removing all assertions containing non-Σ-symbols. As $\mathcal{A}' \subseteq \mathcal{A}$ and $\mathcal{T}_2 \subseteq \mathcal{T}_2'$, we must have $\mathcal{T}_2, \mathcal{A}' \not\models B(a)$. Moreover, we also have $\mathcal{T}_1, \mathcal{A}' \models A(a)$, since $\mathcal{A}$ and $\mathcal{A}'$ agree on the signature of $(\mathcal{T}_1, q_1)$.

$(\Leftarrow)$ Assume there exists an Σ-ABox $\mathcal{A}$ and individual a such that $\mathcal{T}_1, \mathcal{A} \models A(a)$ and $\mathcal{T}_2, \mathcal{A} \not\models B(a)$. As $\mathcal{A}$ contains no non-Σ symbols, the canonical models of $(\mathcal{T}_2, \mathcal{A})$ and $(\mathcal{T}_2', \mathcal{A})$ will be identical, and so $\mathcal{T}_2', \mathcal{A} \not\models B(a)$. Thus, the ABox $\mathcal{A}$ witnesses $(\mathcal{T}_1, A(x)) \not\subseteq (\mathcal{T}_2', B(x))$. ❑

Theorem 13. If $\mathcal{T}$ is normalized, $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$ iff there is a compact witness for $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$.

We break down the proof of Theorem 13 into several lemmas. The first lemma relates query entailment to the existence of a match candidate. It utilizes the closure of an ABox w.r.t. a TBox, defined as follows: $\mathsf{close}(\mathcal{A}, \mathcal{T}) = \mathcal{A} \cup \{A(a) \mid \mathcal{T}, \mathcal{A} \models A(a), a \in \mathsf{Ind}(\mathcal{A}), A \in \mathsf{N_C}\}$.

Lemma 12. Suppose $\mathcal{A}$ is consistent with $\mathcal{T}$. Then $\mathcal{T}, \mathcal{A} \models q(\vec{a})$ if and only if there exists a fork rewriting q' of q and an $\vec{a}$-match candidate $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ for q' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$ such that $\mathcal{T}, \mathcal{A} \models C(a)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = a\}$.

Proof. Suppose $\mathcal{A}$ is consistent with $\mathcal{T}$. For the first direction, let q' be a fork rewriting of q and $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ be an $\vec{a}$-match candidate for q' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$ such that $\mathcal{T}, \mathcal{A} \models C(a)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = a\}$. Because of the latter property, we can find for each $1 \leq i \leq n$, a match π_i for the query p_i in the canonical model $\mathcal{I}_{\mathcal{A},\mathcal{T}}$ of $\mathcal{T}, \mathcal{A}$ such that $\pi_i(t_i) = f(p_i)$ (where we suppose $p_i = r_i(t_i, y_i) \wedge p_i'$ for $1 \leq i \leq n$, as in the paper). Likewise, for each $1 \leq i \leq m$, we can find a match $\hat{\pi}_i$ for $\hat{p}_i$ in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$ such that $\hat{\pi}_i(x)$ belongs to the tree rooted at $f(\hat{p}_i)$ for every $x \in \mathsf{term}(\hat{p}_i)$. Now we use the matches π_i and $\hat{\pi}_i$ to construct a $\vec{a}$-match π for q' in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. First, for each $x \in \mathsf{term}(p_0) \cup \{t_1, \ldots, t_n\}$, we set $\pi(x) = f(x)$. Then for each $x \in \mathsf{term}(p_i)$ (resp. $x \in \mathsf{term}(\hat{p}_i)$), we set $\pi(x) = \pi_i(x)$ (resp. $\pi(x) = \hat{\pi}_i(x)$). Notice that our definition is well-defined since the queries $p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m$ can only intersect on the terms in $\{t_1, \ldots, t_n\}$, and for each such term x, we use $\pi(x) = f(x)$. We now prove that π does indeed define a $\vec{a}$-match π for q' in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. First we remark that because of conditions 2 and 3, $\pi(x_i) = a_i$ for every answer variable x_i, and $\pi(a) = a$ for every individual name a. Now consider some atom $A(t) \in q'$. If $A(t)$ belongs to some p_i or $\hat{p}_i$, we simply use the properties of the matches π_i and $\hat{\pi}_i$. Otherwise, we must have $A(t) \in p_0$, and we can use condition 4 of match candidates to derive $A(f(t)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$, and hence $f(t) \in A^{\mathcal{I}_{\mathcal{T},\mathcal{A}}}$. Next consider a role atom $r(t, x) \in q'$. If $r(t, x)$ belongs to some p_i or $\hat{p}_i$, we can again use the properties of the matches π_i and $\hat{\pi}_i$. Otherwise, $r(t, x) \in p_0$, so condition 5 gives us $r(f(t), f(x)) \in \mathcal{A}$, and thus $(f(t), f(x)) \in r^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$.

For the other direction, suppose $\mathcal{T}, \mathcal{A} \models q(\vec{a})$, and let π be a $\vec{a}$-match for q in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. Set q' equal to the maximal fork rewriting of q such that π is a $\vec{a}$-match for q' in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. We aim to construct a $\vec{a}$-match candidate $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ for q' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$ such that $\mathcal{T}, \mathcal{A} \models C(a)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = a\}$. For p_0, we take all atoms α of q' such that π maps all terms in α to some element of $\mathsf{Ind}(\mathcal{A})$. For each $t \in \mathsf{term}(p_0)$, we set $f(t) = \pi(t)$. Then for each atom $r(t, x) \in \alpha$ such that $\pi(t) \in \mathsf{Ind}(\mathcal{A})$ but $\pi(x) \notin \mathsf{Ind}(\mathcal{A})$, we create a query $p_i = r(t, x) \wedge q'|_{\mathsf{Reach}'_{q'}(x)}$, and we set $f(p_i) = \pi(t)$. Finally, for each atom $r(t, x)$ such that t has no predecessor in q' and $f(t) \notin \mathsf{Ind}(\mathcal{A})$, we create a query $\hat{p}_i = q'|_{\mathsf{Reach}'_{q'}(t)}$, and we set $f(\hat{p}_i) = a$ where a is the unique individual such that $\pi(t)$ lies in the tree below a. It is easy to verify that $p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m$ partitions

the atoms of q'. Condition 1 follows from the fact that q' is a fork rewriting of q, and each p_i ($\hat{p}_i$) is mapped by π into a tree structure in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. Conditions 2 and 3 follow from the fact that $\pi(x_i) = a_i$ for every answer variables x_i, and $\pi(a) = a$ for every individual in $\mathsf{term}(q) \cap \mathsf{N_I}$. For condition 4, we note that if $A(t) \in p_0$, then $f(t) = \pi(t) \in \mathsf{Ind}(\mathcal{A})$, and so we must have $f(t) \in A^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$, and hence $\mathcal{T}, \mathcal{A} \models A(f(t))$ and $A(f(t)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$. For condition 5, suppose $r(t, x) \in p_0$. Then $(\pi(t), \pi(x)) \in r^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$, and since $f(t) = \pi(t) = \mathsf{Ind}(\mathcal{A})$ and $f(x) = \pi(x) = \mathsf{Ind}(\mathcal{A})$, also $(f(t), f(x)) \in r^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$. Using the definition of canonical models, we obtain $r(f(t), f(x)) \in \mathcal{A}$, hence $r(f(t), f(x)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$. Condition 6 follows from the fact that all answer variables and constants belong to either p_0 or to the first atom in some p_i, but never to any $p_{i'}$ or $\hat{p}_i$. Condition 7 follows directly from our definition of the p_i and $\hat{p}_i$. Finally, for condition 8, we use the fact that the p_i and $\hat{p}_i$ are satisfied in distinct subtrees of $\mathcal{I}_{\mathcal{A},\mathcal{T}}$. ❑

We are now ready to prove the "if" direction of Theorem 13.

Lemma 32. *If $\mathcal{T}$ is normalized, then $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$ if there is a compact witness for $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$.*

Proof. Suppose $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ is a compact witness for $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$. Then for each $a \in \mathsf{Ind}(\mathcal{A}_w)$, we can find a Σ-ABox $\mathcal{A}_a$ which satisfies condition 4 of the definition of compact witnesses. We can assume without loss of generality that each $\mathcal{A}_a$ is such that $\mathsf{Ind}(\mathcal{A}_a) \cap \mathsf{term}(q_2) \subseteq \{a\}$, i.e. the only individual which can appear both in $\mathcal{A}_a$ and the query q_2 is a. We let $\mathcal{A} = \mathcal{A}_w^\Sigma \cup \bigcup_a \mathcal{A}_a$, where $\mathcal{A}_w^\Sigma$ denotes the projection of $\mathcal{A}_w$ onto Σ. We claim that $\mathcal{A}$ is a Σ-Abox consistent with $\mathcal{T}$ such that $\mathcal{T}, \mathcal{A} \models q_1[\vec{a}_w]$ and $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{a}_w]$.

We start by proving that $\mathcal{A}$ is consistent with $\mathcal{T}$. For each $a \in \mathsf{Ind}(\mathcal{A}_w)$, let $\mathcal{I}_a = (\Delta^{\mathcal{I}_a}, \cdot^{\mathcal{I}_a})$ be the canonical model of $\mathcal{T}, \mathcal{A}_a$. Note that by definition the $\Delta^{\mathcal{I}_a}$ are pairwise disjoint. We will use $\mathcal{I}_w$ to denote the canonical model of $\mathcal{A}_w, \mathcal{T}$. Now we construct an interpretation $\mathcal{J} = (\Delta^{\mathcal{J}}, \cdot^{\mathcal{J}})$ with $\Delta^{\mathcal{J}} = \cup_a \Delta^{\mathcal{I}_a}$ and $\cdot^{\mathcal{J}}$ defined as follows:

- $a^{\mathcal{J}} = a$
- $A^{\mathcal{J}} = \cup_a A^{\mathcal{I}_a}$
- $r^{\mathcal{J}} = \cup_a r^{\mathcal{I}_a}$, if $r \in \mathsf{N_R} \setminus \Sigma$
- $r^{\mathcal{J}} = \cup_a r^{\mathcal{I}_a} \cup \{(b, c) \mid r(b, c) \in \mathcal{A}_w\}$, if $r \in \mathsf{N_R} \cap \Sigma$

It is clear from the definition that $\mathcal{J}$ is a model of the assertions in $\bigcup_a \mathcal{A}_a$ and the role assertions in $\mathcal{A}_w^\Sigma$. Consider some concept assertion $A(a) \in \mathcal{A}_w^\Sigma$. From items (a) and (b) of condition 4, we obtain $\mathcal{T}, \mathcal{A}_a \models A(a)$, and hence $a \in A^{\mathcal{I}_a} \subseteq A^{\mathcal{J}}$. To prove that $\mathcal{J}$ is also a model of $\mathcal{T}$, we show that every axiom of $\mathcal{T}$ is satisfied at every element of $\Delta^{\mathcal{J}}$. We restrict our attention to elements from $\mathsf{Ind}(\mathcal{A}_w)$ (the roots of the interpretations $\mathcal{I}_a$), since all other elements of $\Delta^{\mathcal{J}}$ conserve their types from the interpretation $\mathcal{I}_a$ from which they originated, and hence must satisfy all axioms. First consider an axiom $A \sqsubseteq B \in \mathcal{T}$ and suppose $a \in A^{\mathcal{J}}$. Then $a \in A^{\mathcal{I}_a}$, hence $a \in B^{\mathcal{I}_a} \subseteq B^{\mathcal{J}}$. Next take an axiom $A \sqsubseteq \exists r.B \in \mathcal{T}$ and $a \in A^{\mathcal{J}}$. Then $a \in A^{\mathcal{I}_a}$, hence $a \in (\exists r.B)^{\mathcal{I}_a}$. As $\mathcal{J}$ contains $\mathcal{I}_a$, we also have $a \in (\exists r.B)^{\mathcal{J}}$. Conjunctive axioms of the form $A_1 \sqcap A_2 \sqsubseteq B$ are handled similarly. Finally consider $\exists r.A \sqsubseteq B \in \mathcal{T}$ and $a \in (\exists r.A)^{\mathcal{J}}$. Then either $a \in (\exists r.A)^{\mathcal{I}_a}$ (in which case we proceed as before), or there is some $b \in \mathsf{Ind}(\mathcal{A}_w)$ such that $(a, b) \in r^{\mathcal{J}}$ and $b \in A^{\mathcal{J}}$. In the latter case, the properties of canonical models imply that $r(a, b) \in \mathcal{A}_w$. From $b \in A^{\mathcal{J}}$, we obtain $b \in A^{\mathcal{I}_b}$, and then by Lemma 31, $\mathcal{A}_b, \mathcal{T} \models A(b)$. We can then apply condition 4 of the compact witness definition to get $A(b) \in \mathcal{A}_w$, which together with $r(a, b) \in \mathcal{A}_w$ yields $\mathcal{A}_w, \mathcal{T} \models B(a)$. Again using condition 4, we can derive $a \in B^{\mathcal{I}_a} \subseteq B^{\mathcal{J}}$. This completes our proof that $\mathcal{J} \models \mathcal{T}$.

To show $\mathcal{T}, \mathcal{A} \models q_1[\vec{a}_w]$, we first show that $\Pi_w = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ is a $\vec{a}_w$-match candidate for q_1' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$. Conditions 1-3 and 6-8 are satisfied since we know Π_w to be a $\vec{a}_w$-match candidate for q_1' and $\mathcal{A}_w$. For condition 4, consider some $A(t) \in p_0$. We know that $A(f(t)) \in \mathcal{A}_w$, and hence $\mathcal{A}_{f(t)}, \mathcal{T} \models A(f(t))$. It follows that $A(f(t)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$. For condition 5, if $r(t, t') \in p_0$, then $r(f(t), f(t')) \in \mathcal{A}_w$, hence $r(f(t), f(t')) \in \mathcal{A}_w^\Sigma \subseteq \mathcal{A} \subseteq \mathsf{close}(\mathcal{A}, \mathcal{T})$. Now we can apply condition 4 of compact witnesses to infer that $\mathcal{A}_a, \mathcal{T} \models C(a)$ (hence $\mathcal{T}, \mathcal{A} \models C(a)$) for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = a\}$. By applying Lemma 12, we obtain $\mathcal{T}, \mathcal{A} \models q_1[\vec{a}_w]$.

To show $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{a}_w]$, we suppose for a contradiction that $\mathcal{T}, \mathcal{A} \models q_2[\vec{a}_w]$. It follows by Lemma 12 that there is a fork rewriting q_2' of q_2 and a $\vec{a}_w$-match candidate $\Pi = \langle p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m, f\rangle$ for q_2' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$ such that $\mathcal{T}, \mathcal{A} \models C(a)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = a\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = a\}$. We assume without loss of generality that q_2' is the finest fork rewriting for which such a match candidate exists. Now we wish to use Π to construct a $\vec{a}_w$-match candidate $\Psi = (s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l, g)$ for q_2' and $\mathcal{A}_w$, which will later serve to show that the compact witness $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ does not satisfy condition 4(d), yielding the desired contradiction. We define s_0 as the atoms in p_0 whose terms are all mapped to $\mathsf{Ind}(\mathcal{A}_w)$ by f. The queries $s_1, \ldots, s_l$ include all queries p_i such that $f(p_i) \in \mathsf{Ind}(\mathcal{A}_w)$ as well as all queries of the form $r(t, x) \wedge q_2'|_{\mathsf{Reach}_{q_2'}(x)}$ such that $f(t) \in \mathsf{Ind}(\mathcal{A}_w)$ and $f(x) \notin \mathsf{Ind}(\mathcal{A}_w)$. For the queries $\hat{s}_1, \ldots, \hat{s}_l$, we take all queries $\hat{p}_i$ as well as all queries of the form $q_2'|_{\mathsf{Reach}_{q_2'}(x)}$ such that $f(x) \notin \mathsf{Ind}(\mathcal{A}_w)$ and x is not reachable from any other term in $\mathsf{term}(q_2')$. The function $g : \mathsf{term}(q_2') \to \mathsf{Ind}(\mathcal{A}_w)$ is defined as follows:

- for $t \in \mathsf{term}(q_2')$ such that $f(t) \in \mathsf{Ind}(\mathcal{A}_w)$: $g(t) = f(t)$
- for $t \in \mathsf{term}(q_2')$ such that $f(t) \in \mathsf{Ind}(\mathcal{A}) \setminus \mathsf{Ind}(\mathcal{A}_w)$: $g(t) = a$ where a is the unique individual in $\mathcal{A}_w$ such that $f(t) \in \mathsf{Ind}(\mathcal{A}_a)$

We must verify that Ψ satisfies all the requirements of a $\vec{a}_w$-match candidate for q_2' and $\mathcal{A}_w$. We first remark that every query s_i ($1 \leq i \leq n$) is either equal to some p_i, or it consists of a first part which is mapped by f into a sub-tree $\mathcal{B}$ of some ABox $\mathcal{A}_a$, together with the (tree-shaped) queries p_i which are mapped by f to the individuals in $\mathcal{B}$. It follows that s_i contains no cycles, and if $r(x, y), r(z, y) \in s_i$, then

x and z must always be mapped to the same element. This means x and z can be identified by fork elimination, and since q_2' is assumed to be a finest fork rewriting, we must have $x = z$. Thus, the queries $s_1, \ldots, s_l$ are all tree-shaped. The queries $\hat{s}_j$ either correspond to some $\hat{p}_i$ or to queries which are partially satisfied inside some $\mathcal{A}_a$ and completed by queries of the form p_i. Thus, using similar reasoning, we can conclude that the queries $\hat{s}_j$ are also tree-shaped. We next show that the queries $s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l$ include all the atoms of q_2'. Each atom $\alpha \in p_0$ either belongs to s_0 or belongs to some s_i. Every query $\hat{p}_i$ is equal to some $\hat{s}_j$. Finally, each query p_i ($i \geq 1$) either is equal to some s_j, or it is a subquery of some s_j (if there is a path to some element of $\mathsf{Ind}(\mathcal{A}_w)$), or is a subquery of some $\hat{s}_j$ (if there is no path back to $\mathsf{Ind}(\mathcal{A}_w)$). It is not hard to see that these queries are pairwise disjoint, so $s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l$ define a partitioning of the atoms of q_2'. We now show that conditions 1 to 8 are satisfied.

Condition 1: to show that $s_1, \ldots, s_k \in \mathsf{Trees}^+(q_2')$ and $\hat{s}_1, \ldots, \hat{s}_l \in \mathsf{Trees}(q_2')$, we use the definition of the queries $s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l$ and the fact that they are all tree-shaped.

Condition 2: trivially holds since for each $a_i \in \vec{a_w}$, we have $a_i \in \mathsf{Ind}(\mathcal{A}_w)$, and hence $g(x_i) = f(x_i) = a_i$.

Condition 3: we know that $f(c) = c$ for all $c \in \mathsf{term}(q_2') \cap \mathsf{N_I}$. As we have assumed that $\mathsf{Ind}(\mathcal{A}_a) \cap \mathsf{term}(q_2) \subseteq \{a\}$ for each $a \in \mathsf{Ind}(\mathcal{A}_w)$, it follows that $\mathsf{Ind}(\mathcal{A}) \cap \mathsf{term}(q_2) \subseteq \mathsf{Ind}(\mathcal{A}_w)$, and so $g(c) = f(c) = c \in \mathsf{Ind}(\mathcal{A}_w)$ for every $c \in \mathsf{term}(q_2') \cap \mathsf{N_I}$.

Condition 4: let $A(t) \in s_0$, which means $A(t) \in p_0$ and $f(t) \in \mathsf{Ind}(\mathcal{A}_w)$. We know that $A(f(t)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$ from condition 4 for Π, hence $\mathcal{A}_w^\Sigma \cup \cup_a \mathcal{T}, \mathcal{A}_a \models A(f(t))$ (since $\mathcal{A} = \mathcal{A}_w^\Sigma \cup \bigcup_a \mathcal{A}_a$). We aim to show $A(g(t)) \in \mathcal{A}_w$, or equivalently, $A(f(t)) \in \mathcal{A}_w$ (since $f = g$ for terms in s_0). We start by showing how the canonical model of $\mathcal{T}, \mathcal{A}$ can be obtained by combining the canonical models of the ABoxes $\mathcal{A}_w^\Sigma$ and $\mathcal{A}_a$. Let $\mathcal{I}$ be the canonical model of $\mathcal{T}, \mathcal{A}$, let $\mathcal{J}_w^\Sigma$ be the canonical model of $\mathcal{A}_w^\Sigma, \mathcal{T}$, and for each $a \in \mathsf{Ind}(\mathcal{A}_w)$, let $\mathcal{J}_a$ be the canonical model of $\mathcal{T}, \mathcal{A}_a$. Define a new interpretation $\mathcal{K}$ by taking the union of $\mathcal{J}_w^\Sigma$ and the $\mathcal{J}_a$. More precisely:

- $\Delta^\mathcal{K} = \cup_a \Delta^{\mathcal{J}_a} \cup \Delta^{\mathcal{J}_w^\Sigma}$
- $A^\mathcal{K} = \cup_a A^{\mathcal{J}_a} \cup A^{\mathcal{J}_w^\Sigma}$
- $r^\mathcal{K} = \cup_a r^{\mathcal{J}_a} \cup r^{\mathcal{J}_w^\Sigma}$

We claim that $\mathcal{K}$ is in fact equal to $\mathcal{I}$. First note that every element in $\Delta^\mathcal{K}$ must belong to the universe of $\mathcal{I}$, because $\mathcal{A}_w^\Sigma \cup \bigcup_a \mathcal{A}_a \subseteq \mathcal{A}$. We have $A^\mathcal{K} \subseteq A^\mathcal{I}$ and $r^\mathcal{K} \subseteq r^\mathcal{I}$ for every $A \in \mathsf{N_C}$ and $R \in \mathsf{N_R}$ for the same reason. This shows that everything in $\mathcal{K}$ belongs to the canonical model $\mathcal{I}$. What remains to be shown is that nothing is missing, i.e. $\Delta^\mathcal{I} \subseteq \Delta^\mathcal{K}$, and $A^\mathcal{I} \subseteq A^\mathcal{K}$ and $r^\mathcal{I} \subseteq r^\mathcal{K}$ for every $A \in \mathsf{N_C}$ and $r \in \mathsf{N_R}$. For this, we use the fact that if such were not the case, there would be some axiom of $\mathcal{T}$ which is not satisfied by $\mathcal{K}$. Thus, we only need to show that $\mathcal{K}$ is a model of $\mathcal{T}$. Consider suppose $A \sqsubseteq B \in \mathcal{T}$ and $c \in A^\mathcal{K}$. Then we must have $c \in \cup_a A^{\mathcal{J}_a} \cup A^{\mathcal{J}_w^\Sigma}$, and since these hence interpretations are known to be models of $\mathcal{T}$, $c \in \cup_a B^{\mathcal{J}_a} \cup B^{\mathcal{J}_w^\Sigma} = B^\mathcal{K}$. The case where $A \sqsubseteq \exists r.B \in \mathcal{T}$ and $c \in A^\mathcal{K}$ is similar. Next suppose $A_1 \sqcap A_2 \sqsubseteq B \in \mathcal{T}$ and $c \in (A_1 \sqcap A_2)^\mathcal{K}$. Then $c \in \cup_a A_1^{\mathcal{J}_a} \cup A_1^{\mathcal{J}_w^\Sigma}$ and $c \in \cup_a A_2^{\mathcal{J}_a} \cup A_2^{\mathcal{J}_w^\Sigma}$. If $A_1 \sqcap A_2(c)$ holds in $\mathcal{J}_w^\Sigma$ or in one of the $\mathcal{J}_a$, then $B(c)$ will hold in that same interpretation, and hence also in $\mathcal{K}$. The only interesting case is thus when $A_1(c)$ and $A_2(c)$ hold in different interpretations. Because the universes of the $\mathcal{J}_a$ are pairwise disjoint, the only possibility is that $c \in \mathsf{Ind}(\mathcal{A}_w)$, one of the interpretations is $\mathcal{J}_c$, and the other $\mathcal{J}_w^\Sigma$. We consider the case where $c \in A_1^{\mathcal{J}_c}$ and $c \in A_2^{\mathcal{J}_w^\Sigma}$ (the other case being symmetric). We thus have $\mathcal{A}_c, \mathcal{T} \models A_1(c)$ and $\mathcal{A}_w^\Sigma, \mathcal{T} \models A_2(c)$, hence $\mathcal{A}_w, \mathcal{T} \models A_2(c)$, which also means $A_2(c) \in \mathcal{A}_w$ (by condition 2). By condition 4 of the compact witness definition, it follows from $\mathcal{A}_c, \mathcal{T} \models A_1(c)$ that $A_1(c) \in \mathcal{A}_w$. Then since $A_1(c), A_2(c) \in \mathcal{A}_w$ and $\mathcal{A}_w$ is closed under $\mathcal{T}$ (by condition 2 of the definition), we must also have $B(c) \in \mathcal{A}_w$. Again applying condition 4, but in the other direction, we obtain $\mathcal{A}_c, \mathcal{T} \models B(c)$. It follows that $c \in B^{\mathcal{J}_c}$, and hence $c \in B^\mathcal{K}$. Finally consider the case where $\exists r.B \sqsubseteq A \in \mathcal{T}$ and $c \in (\exists r.B)^\mathcal{K}$. As before, the only interesting case is when we need to combine elements of two different interpretations, i.e. $r(c, b)$ holds in one interpretation and $B(b)$ holds in another. This can only happen when $r(c, b) \in \mathcal{A}_w^\Sigma$ and $b \in B^{\mathcal{J}_b}$. From the latter, we obtain $\mathcal{A}_b, \mathcal{T} \models B(b)$. Applying condition 4 of the compact witness definition yields $B(b) \in \mathcal{A}_w$. So we have $\mathcal{A}_w, \mathcal{T} \models \exists r.B(c)$, and since $\mathcal{A}_w$ is closed under $\mathcal{T}$, also $A(c) \in \mathcal{A}_w$. Once again by condition 4, we get $\mathcal{A}_c, \mathcal{T} \models A(c)$, hence $c \in A^{\mathcal{J}_a} \subseteq A^\mathcal{K}$.

Now that we know that $\mathcal{K}$ is the canonical model of $\mathcal{T}, \mathcal{A}$, and that $A(f(t)) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$, it follows that we must have we also have that $f(t) \in A^\mathcal{K}$. By the definition of $\mathcal{K}$, we have either $f(t) \in A^{\mathcal{J}_{f(t)}}$ or $f(t) \in A^{\mathcal{J}_w^\Sigma}$. In the former case, we have $\mathcal{A}_{f(t)}, \mathcal{T} \models A(f(t))$, and hence by condition 4 of the compact witness definition, $A(f(t)) \in \mathcal{A}_w$. In the latter case, $\mathcal{A}_w^\Sigma, \mathcal{T} \models A(f(t))$, and hence by condition 2, $A(f(t)) \in \mathcal{A}_w$.

Condition 5: $r(t, t') \in s_0$ means $r(t, t') \in p_0$ and $f(t), f(t') \in \mathsf{Ind}(\mathcal{A}_w)$. Using the fact that Π satisfies condition 4 of the candidate match definition, we obtain $r(f(t), f(t')) \in \mathsf{close}(\mathcal{A}, \mathcal{T})$. We thus have $\mathcal{T}, \mathcal{A} \models r(f(t), f(t'))$, which implies $r(f(t), f(t')) \in \mathcal{A}_w$ since $f(t), f(t') \in \mathsf{Ind}(\mathcal{A}_w)$. As $g(t) = f(t) \in \mathsf{Ind}(\mathcal{A}_w)$ and $g(t') = f(t') = \mathsf{Ind}(\mathcal{A}_w)$, we obtain the desired $r(g(t), g(t')) \in \mathcal{A}_w$.

Condition 6: because of conditions 2 and 3 above, we know that all answer variables and constants in $\mathsf{term}(q_2')$ are mapped to $\mathsf{Ind}(\mathcal{A}_w)$ by f. Moreover, as Π satisfies condition 6 of the candidate match definition, the answer variables and constants in q_2' must belong to $\mathsf{term}(p_0) \cup \{t_1, \ldots, t_n\}$. Now consider an atom α in p_0 which contains an answer variable or constant t. Then since $f(t) \in \mathsf{Ind}(\mathcal{A}_w)$, α will either belong to s_0, or it will be the first atom in some s_i ($1 \leq i \leq k$), with t the first term. Applying condition 8 (see below), we can infer that t does not appear in any s_i' or $\hat{s}_i$. The other possibility is that we have an answer variable or constant t which does not appear in p_0 but does appear as

the first term in some p_i ($1 \leq i \leq n$). Since we know $f(t) \in \mathsf{Ind}(\mathcal{A}_w)$, it follows that p_i belongs to $\{s_1, \ldots, s_n\}$. Again applying condition 8, we obtain $t \notin \mathsf{term}(s'_i)$ and $t \notin \mathsf{term}(\hat{s}_j)$ for every $1 \leq i \leq k$ and $1 \leq j \leq l$. We have thus shown that $s'_1, \ldots, s'_k, \hat{s}_1, \ldots, \hat{s}_l$ may not contain any answer variables or constants, i.e. these queries contain only quantified variables.

Condition 7: consider some $s_i = r(u_i, y_i) \wedge q'_2|_{\mathsf{Reach}_{q'_2}(y_i)}$. Then we know that $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$, so $g(u_i) = f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$. We also know that $f(y_i) \notin \mathsf{Ind}(\mathcal{A}_w)$. Because of the structure of $\mathcal{A}$, $f(y_i)$ must belong to $\mathsf{Ind}(\mathcal{A}_{f(u_i)})$. As $q'_2|_{\mathsf{Reach}_{q'_2}(y_i)}$ is a tree-shaped query, all of its terms must be mapped by f to individuals in $\mathcal{A}_{f(u_i)}$, and hence will be mapped by g to $f(u_i)$. A similar argument shows that g maps all terms in a query $\hat{s}_i$ to unique individual in $\mathsf{Ind}(\mathcal{A}_w)$.

Condition 8: for each $1 \leq i \leq k$, let u_i and y_i be such that $s_i = r(u_i, y_i) \wedge q'_2|_{\mathsf{Reach}_{q'_2}(y_i)}$. We wish to show that the sets $\mathsf{term}(s_0) \cup \{u_1, \ldots, u_k\}$, $\mathsf{term}(s'_1), \ldots, \mathsf{term}(s'_k)$, $\mathsf{term}(\hat{s}_1), \ldots, \mathsf{term}(\hat{s}_l)$ are pairwise disjoint. We first consider a term $t \in \mathsf{term}(s_0)$. We know from the definition of s_0 that t belongs to $\mathsf{term}(p_0)$ and is mapped to $\mathsf{Ind}(\mathcal{A}_w)$ by f. Any atom $A(t) \in q'_2$ must thus belong to s_0. Atoms of the form $r(t, x)$ can either appear in s_0 (if $f(x) \in \mathsf{Ind}(\mathcal{A}_w)$) or will appear as the first atom of some s_i (if $f(x) \notin \mathsf{Ind}(\mathcal{A}_w)$). Finally, an atom of q'_2 of the form $r(x, t)$ must belong to p_0 (since otherwise t would appear in some p'_i or $\hat{p}_i$, contradicting condition 8 of Π). Hence condition 5 yields $r(f(x), f(t)) \in \mathcal{A}$, and the structure of $\mathcal{A}$ gives us $f(x) \in \mathsf{Ind}(\mathcal{A}_w)$. Then using the definition of Ψ, we conclude that the atom $r(x, t)$ belongs to s_0. A term u_i must either be such that $r(u_i, y_i) \in p_0$, $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$, and $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$, or such that $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$ and $s_i = r(u_i, y_i) \wedge q'_2|_{\mathsf{Reach}_{q'_2}(y_i)}$ is equal to some p_j. Atoms of the form $A(u_i)$ will belong to s_0 (since we must have $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$). Atoms of the form $r(u_i, x)$ belong either to p_0 or appear as the first atom of some p_i. As $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$, these atoms will appear either in s_0 or as the first atom of some s_j. Finally, consider an atom $r(x, u_i)$. As $u_i \in \mathsf{term}(p_0) \cup \{t_1, \ldots, t_n\}$, we know from condition 8 of Π that u_i does not appear in any p'_i or p_i. Thus, $r(x, u_i) \in p_0$, yielding $r(f(x), f(u_i)) \in \mathcal{A}_w$. We obtain $f(x) \in \mathsf{Ind}(\mathcal{A}_w)$, and hence $r(x, u_i) \in s_0$. We can thus conclude that no term in $\mathsf{term}(s_0) \cup \{u_1, \ldots, u_k\}$ may appear in any s'_i or $\hat{s}_i$. Finally, we remark that distinct s'_i cannot share terms since each such query is tree-shaped and contains all reachable terms. Distinct $\hat{s}_i$, and pairs $s_i, \hat{s}_j$, cannot share terms for similar reasons.

As q'_2 is a fork rewriting for q_2, and $\Psi = (s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l, g)$ is a $\vec{a}_w$-match candidate for q'_2 and $\mathcal{A}_w$, it must be the case that there is a s_i such that $C_{s_i} \in \nu(g(s_i))$ or a $\hat{s}_i$ such that $C^u_{\hat{s}_i} \in \nu(g(\hat{s}_i))$. We consider only the former case (the latter is similar). Let s_i be such that $C_{s_i} \in \nu(g(s_i))$, and let $a = g(s_i)$. We want to show that $\mathcal{T}, \mathcal{A}_a \models C_{s_i}(a)$, in order to obtain the desired contradiction. If $s_i = p_j$ for some j, then we are done, since then $f(p_j) \in \mathsf{Ind}(\mathcal{A}_w)$, $g(s_i) = f(p_j)$, and $\mathcal{A}_{f(p_j)}, \mathcal{T} \models C_{p_j}(f(p_j))$ (since $(\mathcal{A}_w, \vec{a}_w, q'_1, \Pi_w, \nu_w)$ is know to satisfy condition 4(c)). The other possibility is that $s_i = r(u_i, y_i) \wedge q'_2|_{\mathsf{Reach}_{q'_2}(y_i)}$ for some $r(u_i, y_i) \in q'_2$ with $f(u_i) \in \mathsf{Ind}(\mathcal{A}_w)$ and $f(y_i) \notin \mathsf{Ind}(\mathcal{A}_w)$. We will show by induction on the position of z in the tree structure of s_i that $\mathcal{T}, \mathcal{A}_a \models C_{s_i,z}(f(z))$ for all $z \in \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$. For the base case, consider a term $z \in \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$ which does not contain any children in $\mathsf{term}(s_i) \cap \mathsf{term}(p_0)$. We know that

$$C_{s_i,z} = \bigsqcap_{A(z) \in s_i} A \sqcap \bigsqcap_{r(z,w) \in s_i} \exists r.C_{s_i,w}$$

Each $A(z) \in s_i$ belongs to p_0, so condition 4 of match candidates yields $A(f(z)) \in \mathcal{A}$ and hence $\mathcal{T}, \mathcal{A}_a \models A(f(z))$. Next consider an atom $r(z, w)$ in s_i. As $w \notin \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$, the atom $r(z, w)$ does not appear in p_0, and hence it must appear as the first atom in some $p_i = r(z, w) \wedge p'_i$ with $f(p_i) = f(z)$. Because of our earlier assumption that $\mathcal{T}, \mathcal{A} \models C(b)$ for all concepts C from $\{C_{p_i} \mid f(p_i) = b\} \cup \{C^u_{\hat{p}_i} \mid f(\hat{p}_i) = b\}$, we know that $\mathcal{T}, \mathcal{A} \models C_{p_i}(f(z))$. As $C_{p_i} = \exists r.C_{s_i,w}$ and $f(z)$ only has successors in $\mathcal{A}_a$ (so $\mathcal{T}, \mathcal{A}_a \models C_{p_i}(f(z))$), we obtain the desired $\mathcal{T}, \mathcal{A}_a \models C_{s_i,z}(f(z))$. Now for the induction step, consider some term $z \in \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$ such that for every child w of z in $z \in \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$, we already know that $\mathcal{T}, \mathcal{A}_a \models C_{s_i,w}(f(w))$. For the same reasons as before, it is easy to see that $\mathcal{T}, \mathcal{A}_a \models A(f(z))$ whenever $A(z) \in s_i$. So consider some atom $r(z, w)$ in s_i. If $w \notin \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$, then we can proceed exactly as in the base case. Thus, let us suppose that $w \in \mathsf{term}(s_i) \cap \mathsf{term}(p_0)$. Then $r(z, w) \in p_0$, so $r(f(z), f(w)) \in \mathcal{A}$, hence $r(f(z), f(w)) \in \mathcal{A}_a$. Moreover, from the induction hypothesis, we know that $\mathcal{T}, \mathcal{A}_a \models C_{s_i,w}(f(w))$. We thus have $\mathcal{T}, \mathcal{A}_a \models \exists r.C_{s_i,w}(f(z))$. Since we have shown that all conjuncts of $C_{s_i,z}$ are entailed at $f(z)$, we obtain $\mathcal{T}, \mathcal{A}_a \models C_{s_i,z}(f(z))$. This completes the proof, since by taking $z = u_i$ we get $\mathcal{T}, \mathcal{A}_a \models C_{s_i}(a)$, yielding the desired contradiction. This proves $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{a}_w]$. ❑

To prove the other direction of Theorem 13, we start by showing that if there is an ABox witnessing non-containment, then we can find one which is forest-shaped and whose non-tree core is small.

Lemma 33 (Forest ABox witness)**.** *If $q_1 \not\subseteq_{\mathcal{T},\Sigma} q_2$, then there is a Σ-ABox $\mathcal{A}$ and a tuple $\vec{c}$ such that $\mathcal{T}, \mathcal{A} \models q_1[\vec{c}]$, $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{c}]$, and there exists $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_n \subseteq \mathcal{A}$ satisfying:*

- *$\mathcal{A}$ is the union of $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_n$*
- *$n \leq |\mathsf{Ind}(\mathcal{A}_0)|$ and $|\mathsf{Ind}(\mathcal{A}_0)| \leq |\mathsf{term}(q_1)|$*
- *for $1 \leq i \leq n$: $\mathcal{A}_i$ is a tree-shaped ABox rooted at some individual in $\mathsf{Ind}(\mathcal{A}_0)$*
- *for all $1 \leq i \neq j \leq n$: $\mathsf{Ind}(\mathcal{A}_i) \cap \mathsf{Ind}(\mathcal{A}_j) = \emptyset$ and $|\mathsf{Ind}(\mathcal{A}_i) \cap \mathsf{Ind}(\mathcal{A}_0)| = 1$*
- *if $a \in \mathsf{term}(q_1) \cap \mathsf{N_I}$ or $a \in \vec{c}$, then $a \in \mathsf{Ind}(\mathcal{A}_0)$*
- *if $\mathcal{T}, \mathcal{A} \models C(a)$ for $a \in \mathsf{Ind}(\mathcal{A}_0)$ and $C \in \mathsf{N_C} \cup \{C_p, C^u_p \mid p \in \mathsf{Trees}^*(q_1)\}$, then there is $\mathcal{A}_i$ rooted at a such that $\mathcal{A}_i, \mathcal{T} \models C(a)$*

Proof. Suppose $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$, and let $\mathcal{B}$ be a Σ-ABox such that $(\mathcal{T},\mathcal{B}) \models q_1[\vec{c}]$ and $(\mathcal{T},\mathcal{B}) \not\models q_2[\vec{c}]$. Let $\mathcal{I}$ be the canonical model of $(\mathcal{T},\mathcal{B})$. As $(\mathcal{T},\mathcal{B}) \models q_1[\vec{c}]$ and $(\mathcal{T},\mathcal{B}) \not\models q_2[\vec{c}]$, we know that $\mathcal{I} \models q_1[\vec{c}]$ and $\mathcal{I} \not\models q_2[\vec{c}]$. Consider some match π for the query $q_1[\vec{c}]$ in $\mathcal{I}$. Define a set M consisting of those elements $b \in \mathsf{Ind}(\mathcal{B})$ such that some element in the image of π appears in the subtree of $\mathcal{I}$ rooted at $b \in \Delta^{\mathcal{I}}$. Note in particular that every constant in $\mathsf{term}(q_1)$ will belong to M, as will every individual in $\vec{c}$.

For each $a \in M$, we let $\mathcal{B}_a$ be the unraveling of $\mathcal{A}$ with respect to a. More specifically, $\mathsf{Ind}(\mathcal{B}_a)$ consists of all finite words of the form $a_0 r_0 a_1 \dots r_{n-1} a_n$ such that $a_0 = a$ and $r_i(a_i, a_{i+1}) \in \mathcal{B}$ for every $0 \leq i < n$. Abusing notation, we will use $\mathsf{tail}(w)$ to denote the final individual in the word $w \in \mathsf{Ind}(\mathcal{B}_a)$. The ABox $\mathcal{B}_a$ consists of all assertions satisfying one of the following conditions:

- $A(w)$, where $w \in \mathsf{Ind}(\mathcal{B}_a)$ and $A(\mathsf{tail}(w)) \in \mathcal{B}$
- $s(w_1, w_2)$, where $w_1, w_2 \in \mathsf{Ind}(\mathcal{B}_a)$ and $w_2 = w_1 s b$ for some $b \in \mathsf{Ind}(\mathcal{B})$

By definition, each $\mathcal{B}_a$ is an infinite tree-shaped Σ-ABox. We wish to show that $\mathcal{B}_a, \mathcal{T} \models A(a)$ whenever $\mathcal{B}, \mathcal{T} \models A(a)$. To show this, we define an ABox simulation $S \subseteq \mathsf{Ind}(\mathcal{B}) \times \mathsf{Ind}(\mathcal{B}_a)$ as follows:

- $(a, a) \in S$
- if $(b, w) \in S$ and $r(b, c) \in \mathcal{B}$, then $(c, wrc) \in S$

It is easily verified that S is indeed a simulation and that $(\mathcal{B}, a) \leq (\mathcal{B}_a, a)$. It follows then by Lemma 35 that for all $\mathcal{EL}$ concepts C, we have $\mathcal{B}, \mathcal{T} \models C(a) \Rightarrow \mathcal{B}_a, \mathcal{T} \models C(a)$. In particular this holds for all concept names A and all concepts from $\{C_p, C_p^u \mid p \in \mathsf{Trees}^*(q_1)\}$. We can now apply compactness in order to find a finite subset $\mathcal{B}_a^f \subseteq \mathcal{B}_a$ such that $\mathcal{B}_a^f, \mathcal{T} \models A(a)$ for all $C \in \mathsf{N_C} \cup \{C_p, C_p^u \mid p \in \mathsf{Trees}^*(q_1)\}$ such that $\mathcal{B}, \mathcal{T} \models C(a)$. We can assume without loss of generality that no proper subset of $\mathcal{B}_a^f$ satisfies this condition, which means in particular that $\mathcal{B}_a^f$ is connected, hence is tree-shaped.

We let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be the ABoxes $\mathcal{B}_a^f$ ($a \in M$). We define the ABox $\mathcal{A}_0$ as follows:

$$\mathcal{A}_0 = \{s(a_1, a_2) \in \mathcal{B} \mid a_1, a_2 \in M\}$$

Finally, we let $\mathcal{A} = \mathcal{A}_0 \cup \dots \cup \mathcal{A}_n$. It follows easily from the definition of $\mathcal{A}$ that conditions 1-5 of the lemma are satisfied by $\mathcal{A}$. It remains to be shown that $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{c}]$, $\mathcal{T}, \mathcal{A} \models q_1[\vec{c}]$, and that condition 6 holds.

For the first point, we define a mapping $h : \mathsf{Ind}(\mathcal{A}) \to \mathsf{Ind}(\mathcal{B})$ by setting $h(w) = \mathsf{tail}(w)$. It is easily verified that h defines an ABox homomorphism, i.e. $A(a) \in \mathcal{A} \to A(h(a)) \in \mathcal{B}$ and $r(a, b) \in \mathcal{A} \Rightarrow r(h(a), h(b)) \in \mathcal{B}$. As $\mathcal{B}, \mathcal{T} \not\models q_2[\vec{c}]$, it follows that $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{c}]$.

For the second point, let $\mathcal{J}$ be the canonical model of $\mathcal{A}, \mathcal{T}$. We claim that the match π of $q_1[\vec{c}]$ in the canonical model $\mathcal{I}$ of $\mathcal{B}, \mathcal{T}$ used to define M is also a match of $q_1[\vec{c}]$ in $\mathcal{J}$. First note that if $A(t) \in q_1$ and $\pi(t) \in M$, then $A(\pi(t)) \in \mathcal{A}$, hence $\pi(t) \in A^{\mathcal{J}}$. Similarly, if $r(t_1, t_2) \in q_1[\vec{c}]$ and $\pi(t_1), \pi(t_2) \in M$, then $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{J}}$. All other atoms of q_1 contain a term which is mapped by π to the anonymous part of $\mathcal{I}$, so we need to ensure that the anonymous part of $\mathcal{J}$ contains the required portion of the anonymous part of $\mathcal{I}$. To do so, we define an ABox simulation $S \subseteq \mathsf{Ind}(\mathcal{B}) \times \mathsf{Ind}(\mathcal{A})$ as follows:

- for every $a \in M$, $(a, a) \in S$
- if $(b, w) \in S$ and $r(b, c) \in \mathcal{B}$, then $(c, wrc) \in S$

For each $a \in M$, we have $(\mathcal{A}, a) \leq (\mathcal{B}, a)$, and so, by Lemma 35, we know that for all $\mathcal{EL}$ concepts C and all $a \in M$: $\mathcal{B}, \mathcal{T} \models C(a) \Rightarrow \mathcal{T}, \mathcal{A} \models C(a)$. It follows that every path $a\, r_1\, C_1\, r_2 C_2 \cdots r_n\, C_n \in \Delta^{\mathcal{I}}$ with $a \in M$ also belongs to $\Delta^{\mathcal{J}}$. Moreover, the elements along this path satisfy all the atomic concepts that were satisfied by the corresponding elements of $\Delta^{\mathcal{I}}$. This means that π will satisfy the remaining atoms of q_1, and hence defines a match for $q_1[\vec{c}]$ in $\mathcal{J}$. By Lemma 31, we can conclude that $\mathcal{T}, \mathcal{A} \models q_1[\vec{c}]$.

For condition 6, we first note that $\mathcal{T}, \mathcal{A} \models C(a)$ implies $\mathcal{B}, \mathcal{T} \models C(a)$ for all $a \in \mathsf{Ind}(\mathcal{A})$ and $A \in \mathsf{N_C}$. This is because $g : \mathsf{Ind}(\mathcal{A}) \to \mathsf{Ind}(\mathcal{B})$ defined by $g(w) = \mathsf{tail}(w)$ defines an ABox homomorphism from $\mathcal{A}$ into $\mathcal{B}$. Thus, whenever $\mathcal{T}, \mathcal{A} \models C(a)$, we must also have $\mathcal{B}, \mathcal{T} \models C(a)$. Moreover, we showed above that $\mathcal{B}, \mathcal{T} \models C(a)$ implies $\mathcal{B}_a^f, \mathcal{T} \models C(a)$ for every $C \in \mathsf{N_C} \cup \{C_p, C_p^u \mid p \in \mathsf{Trees}^*(q_1)\}$. This concludes the proof since every $\mathcal{B}_a^f$ is equal to some $\mathcal{A}_i$ ($1 \leq i \leq n$). ❏

We now use the preceding lemma to prove the second direction of Theorem 13.

Lemma 34. *If $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$, then there is a compact witness for $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$.*

Proof. Suppose that $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$. By Lemma 33, we can find a Σ-ABox $\mathcal{A}$ and a tuple $\vec{c}$ such that $\mathcal{T}, \mathcal{A} \models q_1[\vec{c}]$, $\mathcal{T}, \mathcal{A} \not\models q_2[\vec{c}]$, and there exists $\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_n \subseteq \mathcal{A}$ such that (i) $\mathcal{A}$ is the union of $\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_n$, (ii) $n \leq |\mathsf{Ind}(\mathcal{A}_0)| \leq |\mathsf{vars}(q_1)|$, (iii) each $\mathcal{A}_i$ with $1 \leq i \leq n$ is a tree-shaped ABox rooted at some individual in $\mathsf{Ind}(\mathcal{A}_0)$, (iv) for all $1 \leq i \neq j \leq n$: $\mathsf{Ind}(\mathcal{A}_i) \cap \mathsf{Ind}(\mathcal{A}_j) = \emptyset$ and $|\mathsf{Ind}(\mathcal{A}_i) \cap \mathsf{Ind}(\mathcal{A}_0)| = 1$, (v) if $a \in \mathsf{term}(q_1) \cap \mathsf{N_I}$ or $a \in \vec{c}$, then $a \in \mathsf{Ind}(\mathcal{A}_0)$, and (vi) if $\mathcal{T}, \mathcal{A} \models C(a)$ for $a \in \mathsf{Ind}(\mathcal{A}_0)$ and $C \in \mathsf{N_C} \cup \{C_p, C_p^u \mid p \in \mathsf{Trees}^*(q_1)\}$, then there is $\mathcal{A}_i$ rooted at a such that $\mathcal{A}_i, \mathcal{T} \models C(a)$. Note that because of (iv), there is at most one $\mathcal{A}_i$ rooted at a for each $a \in \mathsf{Ind}(\mathcal{A}_0)$, We can assume w.l.o.g. that if $\mathcal{A}_i$ is rooted at a and $\mathcal{A}_i, \mathcal{T} \models A(a)$ for $A \in \Sigma \cap \mathsf{N_C}$, then $A(a) \in \mathcal{A}_i$. We let π be a match for $q_1[\vec{c}]$ in the canonical model $\mathcal{I}_{\mathcal{T},\mathcal{A}}$ of $\mathcal{T}, \mathcal{A}$.

We now wish to construct a compact witness for $q_1 \not\sqsubseteq_{\mathcal{T},\Sigma} q_2$. For the ABox, we take

$$\mathcal{A}_w = \mathcal{A}_0 \cup \{A(a) \mid a \in \mathsf{Ind}(\mathcal{A}_0), A \in \mathsf{N_C}, \mathcal{T}, \mathcal{A} \models A(a)\}$$

and for $\vec{a}_w$, we use $\vec{c}$. The query q_1' is defined as the finest fork rewriting of q_1 such that π is a match for q_1' in $\mathcal{I}_{\mathcal{T},\mathcal{A}}$.

For the $\vec{a}_w$-match candidate Π_w, the query p_0 consists of all atoms in q_1' whose terms are mapped by π to $\mathsf{Ind}(\mathcal{A}_0)$. For each term $t \in p_0$, we set $f(t) = \pi(t)$. The queries p_i consist of all queries of the form $r(t, x) \wedge q_1'|_{\mathsf{Reach}_{q_1'}(x)}$ such that $r(t, x) \in q_1'$, $\pi(t) \in \mathsf{Ind}(\mathcal{A}_0)$, and $\pi(x) \notin \mathsf{Ind}(\mathcal{A}_0)$. We define $f(p_i) = \pi(t)$ where t is the first term in p_i. For the

queries $\hat{p}_i$, we take all queries of the form $q_1'|_{\mathsf{Reach}_{q_1'}(t)}$ such that $t \in \mathsf{term}(q_1')$, $\pi(t) \notin \mathsf{Ind}(\mathcal{A}_0)$, and t is not reachable from any other term in q_1'. The function f maps the terms in $q_1'|_{\mathsf{Reach}_{q_1'}(t)}$ to the unique individual a such that $\pi(t)$ lies in the tree in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$ rooted at a. Given the structure of canonical models, it is not hard to see that $p_0, p_1, \ldots, p_n, \hat{p}_1, \ldots, \hat{p}_m$ partition the atoms of q_1'. Moreover, each of the queries p_i and $\hat{p}_i$ must be tree-shaped (given the forest structure of $\mathcal{I}_{\mathcal{A},\mathcal{T}}$), which together with the definition of the p_i and $\hat{p}_i$ gives us condition 1. Conditions 2 and 3 follow from the fact that all individual names in q_1' or in the answer tuple $\vec{a}_w$ belong to $\mathsf{Ind}(\mathcal{A}_0)$ (see (v) above). For condition 4, take some atom $A(t) \in p_0$. We know that $\pi(t) \in A^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$, and hence that $\mathcal{T}, \mathcal{A} \models A(\pi(t))$. As $f(t) = \pi(t) \in \mathsf{Ind}(\mathcal{A}_0)$, it follows that $A(f(t)) \in \mathcal{A}_w$. For condition 5, consider some atom $r(t, t') \in p_0$. We know that $(\pi(t), \pi(t')) \in r^{\mathcal{I}_{\mathcal{A},\mathcal{T}}}$, that $\pi(t), \pi(t') \in \mathsf{Ind}(\mathcal{A}_0)$, and that $f(t) = \pi(t), f(t') = \pi(t')$, which yields $r(f(t), f(t')) \in \mathcal{A}_w$. For condition 6, we use the fact that π maps the answer variables and constants in q_1' to elements of $\mathsf{Ind}(\mathcal{A}_0)$ (see property (v) above), whereas the terms in the queries p_i' and $\hat{p}_i$ are mapped into $\Delta^{\mathcal{I}_{\mathcal{A},\mathcal{T}}} \setminus \mathsf{Ind}(\mathcal{A}_0)$. Condition 7 follows immediately from our definition of f. Finally, for condition 8, we use the fact that the p_i' and $\hat{p}_i$ are mapped to disjoint subtrees of $\mathcal{I}_{\mathcal{T},\mathcal{A}}$.

For the final component of our compact witness, we need to define an $\vec{a}$-spoiler ν_w for q_2 w.r.t. $\mathcal{A}_w$. Consider some fork rewriting q_2' of q_2, and some $\vec{a}_w$-match candidate $\Psi = (s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l, g)$ for q_2' and $\mathcal{A}_w$. We remark that $\Psi = (s_0, s_1, \ldots, s_k, \hat{s}_1, \ldots, \hat{s}_l, g)$ is also a $\vec{a}_w$-match candidate for q_2' and $\mathsf{close}(\mathcal{A}, \mathcal{T})$, since $\mathcal{A}_w \subseteq \mathsf{close}(\mathcal{A}, \mathcal{T})$. As we know that $\mathcal{T}, \mathcal{A} \not\models q_2'[\vec{a}_w]$, it follows from Lemma 12 that $\mathcal{T}, \mathcal{A} \not\models C(a)$ for some concept C from $\{C_{s_i} \mid g(s_i) = a\} \cup \{C^u_{\hat{s}_i} \mid g(\hat{s}_i) = a\}$. It follows that $\mathcal{A}_w, \mathcal{T} \not\models C(a)$ for this same C and a. We include C in $\nu_w(a)$. By iterating over every fork rewriting of q_2' and every match candidate, we obtain a mapping ν_w which is an $\vec{a}$-spoiler for q_2 w.r.t. $\mathcal{A}_w$.

It remains to be shown that $(\mathcal{A}_w, \vec{a}_w, q_1', \Pi_w, \nu_w)$ satisfies conditions 1-4 of the definition of compact witnesses. For condition 1, we simply note that $\mathcal{T}, \mathcal{A} \not\models q_2'[\vec{a}_w]$ means $\mathcal{A}$ is consistent with $\mathcal{T}$. As $\mathcal{A}_w \subseteq \mathsf{close}(\mathcal{A}, \mathcal{T})$, $\mathcal{A}_w$ must also be consistent with $\mathcal{T}$. Point 2 follows from the definition of $\mathcal{A}_w$, since whenever $\mathcal{T}, \mathcal{A} \models A(a)$ for $a \in \mathsf{Ind}(\mathcal{A}_0) = \mathsf{Ind}(\mathcal{A}_w)$ and $A \in \mathsf{N_C}$, we have $A(a) \in \mathcal{A}_w$. Condition 3 follows from the fact that all role assertions in $\mathcal{A}_w$ belong to the Σ-ABox $\mathcal{A}$. For condition 4, fix some $a \in \mathsf{Ind}(\mathcal{A}_w)$, and let $\mathcal{A}_i$ be such that $\mathcal{A}_a$ be the unique $\mathcal{A}_i$ rooted at a. As $\mathcal{A}_a \subseteq \mathcal{A}$, we know that $\mathcal{A}_a$ is consistent with $\mathcal{T}$. For part (a), suppose $A(a) \in \mathcal{A}_w$ with $A \in \Sigma$. Then $\mathcal{T}, \mathcal{A} \models A(a)$, so by property (vi) above, $\mathcal{T}, \mathcal{A}_a \models A(a)$. Because of our assumption that all entailed Σ-concept assertions are realized at the roots of the $\mathcal{A}_i$, we obtain $A(a) \in \mathcal{A}_a$. For the other direction, we note that if $A(a) \in \mathcal{A}_a$ then $\mathcal{T}, \mathcal{A} \models A(a)$, hence $A(a) \in \mathcal{A}_w$. For (b), we use the following equivalences: $\mathcal{T}, \mathcal{A}_a \models A(a)$ if and only if $\mathcal{T}, \mathcal{A} \models A(a)$ if and only if $A(a) \in \mathcal{A}_w$. For (c), take some p_i such that $f(p_i) = a$. We know that there is a match for p_i rooted at a in $\mathcal{I}_{\mathcal{A},\mathcal{T}}$ and hence that $\mathcal{T}, \mathcal{A} \models C_{p_i}(a)$. Using property (vi), we can conclude that $\mathcal{T}, \mathcal{A}_a \models C_{p_i}(a)$. A similar argument applies for the concepts $C^u_{p_i}$. Finally, for item (d), consider some $C \in \nu_w(a)$. By the way we have defined ν_w, we know that $\mathcal{T}, \mathcal{A} \not\models C(a)$, hence $\mathcal{T}, \mathcal{A}_a \not\models C(a)$. ❑

Proposition 14. Given an $\mathcal{EL}_\bot$-TBox $\mathcal{T}$, an ABox signature Σ, a set of $\mathcal{EL}^u$-concepts Ψ_1, and a set of $\mathcal{EL}^u$-concepts Ψ_2, it is in EXPTIME to decide whether there is a tree-shaped Σ-ABox $\mathcal{A}$ with root a such that

1. $\mathcal{A}$ is consistent w.r.t. $\mathcal{T}$;
2. $\mathcal{T}, \mathcal{A} \models C(a)$ for all $C \in \Psi_1$;
3. $\mathcal{T}, \mathcal{A} \not\models C(a)$ for all $C \in \Psi_2$.

Proof. We first show that we can assume w.l.o.g. that Ψ_1 and Ψ_2 contain only concept names. Fix an $\mathcal{EL}_\bot$-TBox $\mathcal{T}$, a signature Σ, and a set Ψ of concepts of the form C or $\exists u.C$ with C an $\mathcal{EL}$ concept. For each $C \in \Psi$, we introduce a fresh concept name A_C, and for each $\exists u.C \in \Psi$, we create a fresh concept name A^u_C. Then we define a new TBox $\mathcal{T}'$ as follows:

$$\begin{aligned}\mathcal{T}' = \quad & \mathcal{T} \cup \{C \sqsubseteq A_C \mid C \in \Psi\} \cup \\ & \{C \sqsubseteq A^u_C, \exists r.A^u_C \sqsubseteq A^u_C \mid \exists u.C \in \Psi, r \in \mathsf{N_R}\}\end{aligned}$$

It is straightforward to show that for every tree-shaped Σ-ABox $\mathcal{A}$ with root a, the following properties hold:

- $\mathcal{A}$ is consistent with $\mathcal{T}$ iff $\mathcal{A}$ is consistent with $\mathcal{T}'$
- for every $C \in \Psi$: $\mathcal{T}, \mathcal{A} \models C(a)$ iff $\mathcal{T}, \mathcal{A}' \models A_C(a)$
- for every $\exists u.C \in \Psi$: $\mathcal{T}, \mathcal{A} \models \exists u.C(a)$ iff $\mathcal{T}, \mathcal{A}' \models A^u_C(a)$

Let $\mathcal{T}$ be $\mathcal{EL}_\bot$-TBox, Σ be the ABox signature, and Ψ_1 and Ψ_2 be sets of atomic concepts. Without loss of generality we assume and that all concept names in Ψ_1 and Ψ_2 as well as $\bot$ occur in $\mathcal{T}$. We use $\mathsf{sub}(\mathcal{T})$ to denote the set of all subconcepts of concepts occurring in $\mathcal{T}$ and set $\Gamma := \Sigma \cup \mathsf{sub}(\mathcal{T})$. A Σ-*type* is a finite set t of concept names that occur in Σ and such that $\bigcap t$ is satisfiable w.r.t. $\mathcal{T}$. A Γ-*type* is a subset t of Γ such that $\bigcap t$ is satisfiable w.r.t. $\mathcal{T}$. Given a Γ-type t, we use $\mathsf{cl}_\mathcal{T}(t)$ to denote the set $\{C \in \Gamma \mid \mathcal{T} \models \bigcap t \sqsubseteq C\}$. We use $\mathsf{ex}(\mathcal{T})$ to denote the number of concepts of the form $\exists r.C$ that occur in $\mathcal{T}$ (possibly as a subconcept). Define an automaton $\mathfrak{A} = (Q, \mathcal{F}, Q_f, \Delta)$ as follows:

- $\mathcal{F} = \{\langle t, r_1, \ldots, r_n\rangle \mid t \text{ a } \Sigma\text{-type}, r_1, \ldots, r_n \in \Sigma \cap \mathsf{N_R}, n < \mathsf{ex}(\mathcal{T})\}$ where each $\langle t, r_1, \ldots, r_n\rangle$ is of rank n;
- Q is the set of Γ-types;
- $Q_f = \{q \in Q \mid \Psi_1 \subseteq q \text{ and } q \cap \Psi_2 = \emptyset\}$;
- Θ consists of all rules $f(q_1, \ldots, q_n) \to q$ with $f = \langle t, r_1, \ldots, r_n\rangle$ such that

$$\begin{aligned}q = \mathsf{cl}_\mathcal{T}(t \cup \{&\exists r.C \in \mathsf{sub}(\mathcal{T}) \mid \\ & r = r_i \text{ and } C \in q_i \text{ for some } i \text{ with } 1 \leq i \leq n\}\end{aligned}$$

Since $\mathfrak{A}$ is single-exponentially large in $|\mathcal{T}| + |\Sigma|$ and the emptiness problem can be decided in polynomial time in the size of the automaton, it remains to prove the following claim to obtain the desired EXPTIME procedure.

Claim 1. $L(\mathfrak{A}) \neq \emptyset$ iff there is a tree-shaped Σ-ABox that satisfies Conditions 1-3 from Proposition 14.

Before we can establish Claim 1, we state the following technical result. A formal proof is not difficult and can be found in (Baader et al. 2010).

Claim 2. For any Σ-ABox $\mathcal{A}$ that is consistent w.r.t. $\mathcal{T}$ and all $a \in \mathsf{Ind}(\mathcal{A})$ and $C \in \Gamma$, we have $\mathcal{T}, \mathcal{A} \models C(a)$ iff $C \in \mathsf{cl}_{\mathcal{T}}(t_a)$ where $t_a = \{A \in \Sigma \mid A(a) \in \mathcal{A}\} \cup \{\exists r.C \in \mathsf{sub}(\mathcal{T}) \mid \exists r(a,b) \in \mathcal{A} : \mathcal{T}, \mathcal{A} \models C(b)\}$.

We now prove Claim 1, starting with the "if" direction. Let $\mathcal{A}$ be a tree-shaped Σ-ABox $\mathcal{A}$ with root a_0 that is consistent with $\mathcal{T}$ and satisfies Conditions 1-3 from Proposition 14. When $r(a,b) \in \mathcal{A}$, we call b a *successor* of a in $\mathcal{A}$. By Claim 2, we can assume that the number of successors of each $a \in \mathsf{Ind}(\mathcal{A})$ in $\mathcal{A}$ is bounded by $\mathsf{ex}(\mathcal{T})$: if it is not, choose for each $\exists r.C \in \mathsf{sub}(\mathcal{T})$ a $b \in \mathsf{Ind}(\mathcal{A})$ with $r(a,b) \in \mathcal{A}$ and $\mathcal{T}, \mathcal{A} \models C(b)$ (if such a b exists), and then drop all subtrees rooted at successors of a in $\mathcal{A}$ that have not been chosen. The resulting ABox is clearly still consistent w.r.t. $\mathcal{T}$. Moreover, it satisfies Conditions 2 and 3 of Proposition 14 due to Claim 2.

For each individual in $\mathcal{A}$, fix a total order on the successors. For $a \in \mathsf{Ind}(\mathcal{A})$, we use $\sigma_{\mathcal{A}}(a)$ to denote the set $\{A \in \Sigma \mid A(a) \in \mathcal{A}\}$. Define a tree $T = (V, E, \ell)$ as follows:

- $V = \mathsf{Ind}(\mathcal{A})$;
- $E = \{(a,b) \in V \times V \mid r(a,b) \in \mathcal{A}\}$ and the order of successor in T agrees with the chosen order on successors in $\mathcal{A}$;
- $\ell(a) = \langle \sigma_{\mathcal{A}}(a), r_1, \ldots, r_n \rangle$ where r_i is the (unique!) role such that $r_i(a, a_i) \in \mathcal{A}$, with a_i the i-th successor of a.

Define a mapping ρ that maps each $a \in \mathsf{Ind}(\mathcal{A})$ to the Γ-type $\rho(a) := \{C \in \Gamma \mid \mathcal{T}, \mathcal{A} \models C(a)\}$. We show that ρ is a run of $\mathfrak{A}$ on T. First, note that since $\mathcal{T}, \mathcal{A} \models C(a_0)$ for all $C \in \Psi_1$ and $\mathcal{A}, \mathcal{T} \not\models C(a_0)$ for all $C \in \Psi_2$, we have $\rho(a_0) \in Q_f$ (observe that a_0 is the root of T). By definition of Θ and T, it thus remains to show that for all $a \in \mathsf{Ind}(\mathcal{A})$, we have $\rho(a) = \mathsf{cl}_{\mathcal{T}}(t_a)$. This, however, is immediate by Claim 2.

For the "only if" direction of Claim 1, let $T = (V, E, \ell)$ be a tree accepted by $\mathcal{A}$, and ρ a run of $\mathfrak{A}$ on T. Define a Σ-ABox

$$\begin{aligned}\mathcal{A} \;:=\; & \{A(a_v) \mid v \in V \text{ and } \\ & \qquad \ell(v) = \langle t, r_1, \ldots, r_n \rangle \text{with } A \in t\} \cup \\ & \{r(a_v, a_{v_i}) \mid v_i \text{ is } i\text{-th successor of } v \text{ and } \\ & \qquad \ell(v) = \langle t, r_1, \ldots, r_n \rangle \text{ with } r_i = r\}.\end{aligned}$$

We want to show that $\mathcal{A}$ satisfies Conditions 1 to 3 of Proposition 14. We begin by proving the consistency of $\mathcal{A}$ with respect to $\mathcal{T}$. Let us define Ψ as the set of concepts C which are satisfiable w.r.t. $\mathcal{T}$ and such that $\exists r.C \in \Gamma$ for some role r. For each $C \in \Psi$, we let $\mathcal{J}_C$ be the canonical model of the ABox $\{B(b)\}$ and TBox $\mathcal{T} \cup \{B \equiv C\}$, and we use x_C to denote the element $b^{\mathcal{J}_C}$ of $\Delta^{\mathcal{J}_C}$. Suppose w.l.o.g. that the universes of the $\mathcal{J}_C$ are all disjoint. We use the interpretations $\mathcal{J}_C$ to construct a new interpretation $\mathcal{I}$ as follows:

$$\begin{aligned}\Delta^{\mathcal{I}} &= V \cup \bigcup_{C \in \Psi} \Delta^{\mathcal{J}_C} \\ A^{\mathcal{I}} &= \{v \in V \mid A \in \rho(v)\} \cup \bigcup_{C \in \Psi} A^{\mathcal{J}_C} \\ r^{\mathcal{I}} &= \{(v,w) \in E \mid w \text{ is } i\text{-th successor of } v \text{ and } \\ &\qquad \ell(v) = \langle t, r_1, ..., r_n \rangle \text{ where } r_i = r\} \\ &\qquad \cup \{(v, x_C) \mid v \in V \text{ and } \exists r.C \in \rho(v)\} \cup \bigcup_{C \in \Psi} r^{\mathcal{J}_C} \\ a_v^{\mathcal{I}} &= v\end{aligned}$$

It is easy to see that $\mathcal{I}$ is a model of $\mathcal{A}$. In order to show that it is also a model of $\mathcal{T}$, it is convenient to first establish the following claim. Proof details are in (Baader et al. 2010).

Claim 3.

1. $C \in \rho(v) \Rightarrow v \in C^{\mathcal{I}}$
2. $v \in C^{\mathcal{I}} \,\&\, C \in \Gamma \Rightarrow C \in \rho(v)$

Now let us suppose that $C \sqsubseteq D \in \mathcal{T}$ and $y \in C^{\mathcal{I}}$. The case where $y \in \Delta^{\mathcal{J}_E}$ for some $E \in \Psi$ is straightforward, so we concentrate on the case where $y \in V$. In this case, we know from Point 2 that $C \in \rho(y)$, which means that D must also belong to $\rho(y)$. It follows then from Point 1 that $y \in D^{\mathcal{I}}$, as desired. We have thus shown that $\mathcal{I}$ is a model of $\mathcal{A}$ and $\mathcal{T}$, so $\mathcal{A}$ is consistent with $\mathcal{T}$.

We now prove that Conditions 2 and 3 of Proposition 14 are also satisfied. For Condition 2, let $C \in \Psi_2$. Assume to the contrary of what is to be shown that $\mathcal{T}, \mathcal{A} \models C(a_\varepsilon)$, where ε is the root node of T. Consider the interpretation $\mathcal{I}$ constructed above to show satisfiability. By Point 2 of Claim 3, we have $C \in \rho(\varepsilon)$, in contradiction to $\rho(\varepsilon) \in Q_f$. For Condition 2, let $C \in \Psi_1$. Since $\rho(\varepsilon) \in Q_f$, we have $C \in \rho(\varepsilon)$. It thus suffices to establish the following claim.

Claim 4. For all $v \in V$ and $C \in \rho(v)$: $\mathcal{T}, \mathcal{A} \models C(a_v)$.

The proof is by induction on the co-depth of v. If v is a leaf and $C \in \rho(v)$, then the definition of Θ and $\mathcal{A}$ yields that $C \in \mathsf{cl}_{\mathcal{T}}(\sigma_{\mathcal{A}}(a_v))$, hence $\mathcal{T}, \mathcal{A} \models C(a_v)$. Now let v be a non-leaf with $\ell(v) = \langle t, r_1, \ldots, r_n \rangle$ and successors $v_1, \ldots, v_n$. Moreover, let $C \in \rho(v)$. Then $C \in \mathsf{cl}_{\mathcal{T}}(\sigma_{\mathcal{A}}(a_v) \cup \{\exists r.D \in \mathsf{sub}(\mathcal{T}) \mid r = r_i \text{ and } D \in \rho(a_{v_i}) \text{ for some } 1 \leq i \leq n\})$. By IH, we know that $D \in \rho(a_{v_i})$ implies $\mathcal{T}, \mathcal{A} \models D(a_{v_i})$. Thus, we also have $\mathcal{T}, \mathcal{A} \models C(a_v)$. This completes the proof of Claim 4 and of Claim 1. ❑

D Proofs for Section 5

We require a normal form (distinct from the normal form introduced previously for general $\mathcal{EL}$-TBoxes). Call a concept name A *non-conjunctive* in $\mathcal{T}$ if it does not occur in the form $A \equiv B_1 \sqcap \cdots \sqcap B_n$ in $\mathcal{T}$ with $n \geq 1$ and $B_1, \ldots, B_n$ concept names. A is *primitive* in $\mathcal{T}$ if it does not occur on the left-hand-side of any concept definition in $\mathcal{T}$. A classical TBox $\mathcal{T}$ is *normal* if it consists of definitions $A \equiv \exists r.B$

and $A \equiv B_1 \sqcap \cdots \sqcap B_n$ where $B, B_1, \ldots, B_n$ are concept names and $B_1, \ldots, B_n$ are non-conjunctive. For every classical $\mathcal{EL}$-TBox $\mathcal{T}$ one can construct in polynomial time a normal $\mathcal{EL}$-TBox $\mathcal{T}'$ using additional symbols such that $\mathcal{T}' \models \mathcal{T}$ and every interpretation satisfying $\mathcal{T}$ can be expanded to an interpretation satisfying $\mathcal{T}'$ (Konev, Walther, and Wolter 2008). From now on we only work with normal acyclic or classical TBoxes.

We introduce some notions and results required for the proofs. A relation S between two Σ-ABoxes $\mathcal{A}_1$ and $\mathcal{A}_2$ is called a *simulation* if

- $(a, b) \in S$ and $A(a) \in \mathcal{A}_1$ imply $A(b) \in \mathcal{A}_2$;
- $(a, b) \in S$ and $r(a, a') \in \mathcal{A}_1$ imply that there exists b' with $(a', b') \in S$ and $r(b, b') \in \mathcal{A}_2$.

We say that $(\mathcal{A}_1, a_1)$ is *simulated by* $(\mathcal{A}_2, a_2)$, in symbols $(\mathcal{A}_1, a_1) \leq (\mathcal{A}_2, a_2)$, if there exists a simulation S between $\mathcal{A}_1$ and $\mathcal{A}_2$ with $(a_1, a_2) \in S$. The following lemma is readily checked (cf. (Konev et al. 2011)).

Lemma 35. *For any two Σ-ABoxes $\mathcal{A}_1$ and $\mathcal{A}_2$, if $(\mathcal{A}_1, a_1) \leq (\mathcal{A}_2, a_2)$, then $(\mathcal{T}, \mathcal{A}_1) \models C(a_1)$ implies $(\mathcal{T}, \mathcal{A}_2) \models C(a_2)$, for all $\mathcal{EL}$-concepts C.*

We also require the following result about inclusions that follow from classical $\mathcal{EL}$-TBoxes (Konev, Walther, and Wolter 2008).

Lemma 36. *Let $\mathcal{T}$ be a normal $\mathcal{EL}$ TBox, r a role name, A a primitive concept name in $\mathcal{T}$ and D an $\mathcal{EL}$-concept.*

1. *If*
$$\mathcal{T} \models \bigsqcap_{1 \leq i \leq n} A_i \sqcap \bigsqcap_{1 \leq j \leq m} \exists r_j.C_j \sqsubseteq A$$
where A_i are concept names and C_j are $\mathcal{EL}$-concepts. Then there exists A_i, $1 \leq i \leq n$, such that $\mathcal{T} \models A_i \sqsubseteq A$.
2. *Assume now*
$$\mathcal{T} \models \bigsqcap_{1 \leq i \leq n} A_i \sqcap \bigsqcap_{1 \leq j \leq m} \exists r_j.C_j \sqsubseteq \exists r.D,$$
where A_j are concept names and C_j are $\mathcal{EL}$-concepts, then
   - *there exists A_i, $1 \leq i \leq n$, such that $\mathcal{T} \models A_i \sqsubseteq \exists r.D$ or*
   - *there exists r_j such that $r_j = r$ and $\mathcal{T} \models C_j \sqsubseteq D$.*

Theorem 37. *IQ-containment in $\mathcal{EL}$ with classical TBoxes is in* PTIME.

Proof. For a normal classical $\mathcal{EL}$-TBox $\mathcal{T}$ and a concept name A define

$$\text{non-conj}_{\mathcal{T}}(A) = \begin{cases} \{A\}, A \text{ is non-conjunctive in } \mathcal{T} \\ \{B_1, \ldots, B_n\}, A \equiv B_1 \sqcap \cdots \sqcap B_n \in \mathcal{T} \end{cases}$$

Now, given normal classical $\mathcal{EL}$-TBoxes $\mathcal{T}_1, \mathcal{T}_2$, IQs $A_1(x)$, $A_2(x)$ and a signature Σ, one can construct in polynomial time a Σ-ABox $\mathcal{A}_{\mathcal{T}_2,\Sigma}$ with individual names a_B for B non-conjunctive in $\mathcal{T}_2$ such that the following conditions are equivalent:

(a) $\mathcal{T}_1, A_1(x) \not\subseteq_\Sigma (\mathcal{T}_2, A_2(x))$;

(b) $\mathcal{T}_1, \mathcal{A}_{\mathcal{T}_2,\Sigma} \models A_1(a_B)$ for some $B \in \text{non-conj}_{\mathcal{T}_2}(A_2)$.

Intuitively, $\mathcal{A}_{\mathcal{T}_2,\Sigma}$ is the "most specific" Σ-ABox $\mathcal{A}$ such that $\mathcal{T}_2, \mathcal{A} \not\models B(a_B)$, for every $B \in \text{non-conj}_{\mathcal{T}_2}(A_2)$. Thus, $\mathcal{T}_2, \mathcal{A}_{\mathcal{T}_2,\Sigma} \not\models A_2(a_B)$ for every $B \in \text{non-conj}_{\mathcal{T}_2}(A_2)$. Then, Point 2 holds if, and only if, $\mathcal{A}_{\mathcal{T}_2,\Sigma}$ is a witness for $(\mathcal{T}_1, A_1(x)) \not\subseteq_\Sigma (\mathcal{T}_2, A_2(x))$. For the PTIME upper bound it remains to observe that $\mathcal{T}_1, \mathcal{A}_{\mathcal{T}_2,\Sigma} \models A_1(a_B)$ can be checked in polytime.

We give the construction of $\mathcal{A}_{\mathcal{T},\Sigma}$ and show the equivalence of (a) and (b). (The proof is almost identical to (Konev et al. 2011).) Fix individual names a_A, for A non-conjunctive in $\mathcal{T}$, and an additional individual name a_Σ. For A primitive in $\mathcal{T}$ we set

$$\begin{aligned} \mathcal{A}(A) \quad = \quad & \{E(a_A) \mid E \in \Sigma, \mathcal{T} \not\models E \sqsubseteq A\} \cup \\ & \{r(a_A, a_\Sigma) \mid r \in \Sigma\} \end{aligned}$$

For $A \equiv \exists r.B \in \mathcal{T}$ we set

$$\begin{aligned} \mathcal{A}(A) \quad = \quad & \{B(a_A) \mid B \in \Sigma, \mathcal{T} \not\models B \sqsubseteq A\} \\ & \cup \{s(a_A, a_\Sigma) \mid r \neq s \in \Sigma\} \\ & \cup \{r(a_A, a_E) \mid E \in \text{non-conj}_{\mathcal{T}}(B), \text{ if } r \in \Sigma\} \end{aligned}$$

We also let

$$\mathcal{A}_\Sigma = \{r(a_\Sigma, a_\Sigma) \mid r \in \Sigma\} \cup \{A(a_\Sigma) \mid A \in \Sigma\}.$$

Then

$$\mathcal{A}_{\mathcal{T},\Sigma} \quad = \quad \mathcal{A}_\Sigma \cup \{\mathcal{A}(A) \mid A \text{ non-conjunctive in } \mathcal{T}\}$$

One can readily prove

Claim 1. For every normal classical $\mathcal{EL}$-TBox $\mathcal{T}$, signature Σ, and concept name A_0 that is non-conjunctive in $\mathcal{T}$. $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A_0(a_{A_0})$ and the following conditions are equivalent for every Σ-ABox $\mathcal{A}$ and individual a:

- $(\mathcal{A}, a) \leq (\mathcal{A}_{\mathcal{T},\Sigma}, a_{A_0})$;
- $\mathcal{T}, \mathcal{A} \not\models A_0(a)$.

From Claim 1, one can easily prove the equivalence of (a) and (b) above. ❑

We now prove the correctness of the PSPACE algorithm for CQ-containment.

Lemma 38. *Let $n > 0$. Then $\mathsf{ABox}(\Gamma, n) = 1$ iff the conditions (b1)-(b5) hold for Γ and n.*

Proof. Assume the lemma has been proved for n. Assume $\mathsf{ABox}(\Gamma, n+1) = 1$. Take a tree-shaped Σ-ABox $\mathcal{A}$ with root a of depth $n+1$ with $\mathcal{T}, \mathcal{A} \models C(a)$ iff $C \in \Gamma$ (for all $C \in S$) and $\mathcal{T}, \mathcal{A} \not\models \exists u.C(a)$ for all $C \in \Psi_2^u$. We show conditions (b1)-(b5). Conditions (b1) to (b3) follow from the definition. For (b4) assume that $A \in \Gamma$ is primitive. We have to show there exists $B \in \Sigma \cap \Gamma$ such that $\mathcal{T} \models B \sqsubseteq A$. We have $\mathcal{T}, \mathcal{A} \models A(a)$. Since $\mathcal{A}$ is tree-shaped we have, by Lemma 36 that there exists $B(a) \in \mathcal{A}$ with $\mathcal{T} \models B \sqsubseteq A$, as required.

For (b5) assume $\exists r.C \in \Gamma$. Since $\mathcal{A}$ is tree-shaped we have, by Lemma 36 that (i) there exists $B(a) \in \mathcal{A}$ with $\mathcal{T} \models B \sqsubseteq A$, or (ii), there exists $r(a, b) \in \mathcal{A}$ with $\mathcal{T}, \mathcal{A}_b \models C(b)$, where $\mathcal{A}_b$ is the ABox induced by the subtree of $\mathcal{A}$ generated by b. Then $\Gamma' = \{G \in S \mid \mathcal{T}, \mathcal{A}_b \models G(b)\}$ is as required.

Conversely, assume that condition (b1)-(b5) hold for Γ, $n+1$. We show that there is a tree-shaped Σ-ABox $\mathcal{A}$ with root a of depth $n+1$ with $\mathcal{T}, \mathcal{A} \models C(a)$ iff $C \in \Gamma$ (for all $C \in S$) and $\mathcal{T}, \mathcal{A} \not\models \exists u.C(a)$ for all $C \in \Psi_2^u$.

We construct the required Σ-ABox $\mathcal{A}$ by taking for every $\exists r.C \in \Gamma$ such that there is no $B \in \Gamma \cap \Sigma$ with $\mathcal{T} \models B \sqsubseteq \exists r.C$ a $\Gamma' \subseteq S$ with the properties under (b5). For any such Γ' let $\mathcal{A}_{\Gamma'}$ be a Σ-ABox of depth $\leq n$ with root $a_{\Gamma'}$ such that $\mathcal{T}, \mathcal{A}_{\Gamma'} \models C(a)$ iff $C \in \Gamma'$ (for all $C \in S$) and $\mathcal{T}, \mathcal{A}_{\Gamma'} \not\models \exists u.C(a_{\Gamma'})$ for all $C \in \Psi_2^u$. $\mathcal{A}$ is the union of those ABoxes $\mathcal{A}_{\Gamma'}$ with all $\{r(a, a_{\Gamma'})\}$ and $\{A(a) \mid A \in \Sigma \cap \Gamma\}$. ❑

We now consider the proof of the PSPACE lower bound for single TBox CQ-containment. The following lemma has been shown in the main part of the paper already:

Lemma 39. *φ is valid iff $L_0(x) \not\subseteq_{\mathcal{T},\Sigma} \bigsqcup_{C \in \mathcal{C}} C(x)$.*

We now encode the disjunction on the right-hand-side of this containment problem into an extension of $L_0(a)$ to a conjunctive query. Assume that $\mathcal{C} = \{C_1, \ldots, C_k\}$. Consider an additional role name s, set $\Sigma' = \Sigma \cup \{s\}$, and define concepts D_i, $1 \leq i \leq k$, inductively:

$$\begin{array}{rcl} D_i & := & \exists s.(C_i \sqcap D_{i+1}), \text{ for } 1 \leq i \leq k-1, \\ D_k & := & \exists s.C_k \end{array}$$

Now set

$$\begin{array}{rcl} q_1 & = & \{L_0(a_k)\} \cup \{s(a_i, a_{i+1}), s(a_i, a_k) \mid 0 \leq i < k\} \cup \\ & & \{C_i(a_i), D_{i+1}(a_k) \mid 1 \leq i < k\}. \end{array}$$

(Note that in the definition of q_1 we use the expressions $C_i(a)$ and $D_i(a)$ as abbreviations for the corresponding tree-shaped conjunctive queries.) To enforce that all concepts $C_1, \ldots, C_k$ are false at a_k, we choose $q_2 = D_1(a_0)$.

Lemma 40. *The following conditions are equivalent:*

1. *there exists a Σ'-ABox $\mathcal{A}$ with $\mathcal{T}, \mathcal{A} \models q_1$ and $\mathcal{T}, \mathcal{A} \not\models q_2$;*
2. *φ is valid.*

Proof. For the direction (1) implies (2), let $\mathcal{A}$ be a Σ'-ABox such that $\mathcal{T}, \mathcal{A} \models q_1$ and $\mathcal{T}, \mathcal{A} \not\models q_2$. By Lemma 39, it is sufficient to show that $\mathcal{T}, \mathcal{A} \not\models C(a_k)$ for any $C \in \mathcal{C}$. Let $\mathcal{I}$ be a model of $\mathcal{T}$ and $\mathcal{A}$ with $a_0^{\mathcal{I}} \notin D_1^{\mathcal{I}}$. Since $\mathcal{T}, \mathcal{A} \models q_1$ and $\neg D_1 \equiv \forall s.(\neg C_1 \sqcup \neg D_2)$, we have that $\neg C_1$ is true at a_k and $\neg D_2$ is true at a_1. One can repeat this argument $k-1$ times and obtain that $\neg C_1, \ldots, \neg C_k$ are true at a_k. Hence $\mathcal{T}, \mathcal{A} \not\models \bigsqcup_{C \in \mathcal{C}} C(a_k)$.

For the direction (2) implies (1), assume that φ is valid. Consider the Σ'-ABox $\mathcal{A}_0$ depicted in Figure 3, where the bottom half corresponds to the validation tree for φ. It is readily checked that $\mathcal{T}, \mathcal{A}_0 \models q_1$ and, by Lemma 39, $\mathcal{T}, \mathcal{A}_0 \not\models q_2$, as required. ❑

We modify the proof above for the full signature case. Assume φ is given. We define $\mathcal{T}$ as before and set $\mathcal{T}_1 = \mathcal{T}$. Introduce new concept names V_i', $1 \leq 0 \leq n$ and L_i', $0 \leq i \leq n$, and let $\mathcal{T}_2$ be the TBox obtained from $\mathcal{T}_1$ by replacing

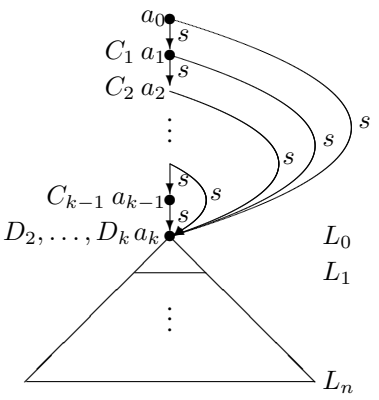

Figure 3: Σ'-ABox $\mathcal{A}_0$.

all V_i and L_i by V_i' and L_i', respectively. Let $\mathcal{D} = \mathcal{C} \cup \mathcal{C}'$, where

$$\begin{array}{rcl} \mathcal{C}' & := & \{\exists r^j.L_i \mid 0 \leq j \leq n, 0 \leq i \leq n\} \cup \\ & & \{\exists r^j.V_i \mid 0 \leq j \leq n, 1 \leq i \leq n\} \end{array}$$

We obtain the following variant of Lemma 39.

Lemma 41. *The following conditions are equivalent:*

1. *there exists an ABox $\mathcal{A}$ with $\mathcal{T}_1, \mathcal{A} \models L_0(a)$ and $\mathcal{T}_2, \mathcal{A} \not\models C(a)$, for every $C \in \mathcal{D}$;*
2. *φ is valid.*

Proof. Can be done by reduction to Lemma 39. Note that the condition $\mathcal{T}_2, \mathcal{A} \not\models C(a)$, for every $C \in \mathcal{D}$, means that the ABox $\mathcal{A}$ is essentially a Σ-ABox: for no b reachable from a by an r-path of length $\leq n$ do we have $B(b) \in \mathcal{A}$ for $B = L_i$, $0 \leq i \leq n$, or $B = V_i$, $1 \leq i \leq n$. ❑

The reduction to conjunctive queries is now done in almost the same way as above. Assume that $\mathcal{D} = \{C_1, \ldots, C_k\}$. Consider a role name s and define concepts D_i, $1 \leq i \leq k$, inductively:

$$\begin{array}{rcl} D_i & := & \exists s.(C_i \sqcap D_{i+1}), \text{ for } 1 \leq i \leq k-1, \\ D_k & := & \exists s.C_k \end{array}$$

and set, as before,

$$\begin{array}{rcl} q_1 & = & \{L_0(a_k)\} \cup \{s(a_i, a_{i+1}), s(a_i, a_k) \mid 0 \leq i < k\} \cup \\ & & \{C_i(a_i), D_{i+1}(a_k) \mid 1 \leq i < k\}. \end{array}$$

Again, $q_2 = D_1(a_0)$.

Lemma 42. *The following conditions are equivalent:*

1. *there exists an ABox $\mathcal{A}$ with $\mathcal{T}_1, \mathcal{A} \models q_1$ and $\mathcal{T}_2, \mathcal{A} \not\models q_2$;*
2. *φ is valid.*

E Proofs for Section 6

Let $\mathcal{T}$ be a TBox and Σ a signature. To prove the results of this section, we introduce an interpretation $\mathcal{I}_{\mathcal{T},\Sigma}$, called the *canonical* Σ*-model of* $\mathcal{T}$. It is defined by setting:

$$\begin{array}{lcl} \Delta^{\mathcal{I}_{\mathcal{T},\Sigma}} & = & \mathfrak{T}_{\mathcal{T},\Sigma} \\ A^{\mathcal{I}_{\mathcal{T},\Sigma}} & = & \{t \in \mathfrak{T}_{\mathcal{T},\Sigma} \mid A \in t\} \\ r^{\mathcal{I}_{\mathcal{T},\Sigma}} & = & \{(t,t') \in \mathfrak{T}_{\mathcal{T},\Sigma} \times \mathfrak{T}_{\mathcal{T},\Sigma} \mid \\ & & \text{for all } \exists r.C \in \mathsf{cl}(\mathcal{T},\Sigma) : C \in t' \Rightarrow \exists r.C \in t\} \end{array}$$

Observe that $\mathcal{A}_{\mathcal{T},\Sigma}$ can be defined using $\mathcal{I}_{\mathcal{T},\Sigma}$ as follows:

$$\begin{array}{lcl} \mathcal{A}_{\mathcal{T},\Sigma} & = & \{A(a_t) \mid t \in A^{\mathcal{I}_{\mathcal{T},\Sigma}} \wedge A \in \Sigma\} \cup \\ & & \{r(a_t, a_{t'}) \mid (t,t') \in r^{\mathcal{I}_{\mathcal{T},\Sigma}} \wedge r \in \Sigma\} \end{array}$$

We state the main property of canonical ABoxes as a theorem:

Theorem 43.
We have $A_1(x) \sqsubseteq_{\mathcal{T},\Sigma} A_2(x)$ *iff* $\mathsf{cert}_{\mathcal{T}}(A_1(x), \mathcal{A}_{\mathcal{T},\Sigma}) \subseteq \mathsf{cert}_{\mathcal{T}}(A_2(x), \mathcal{A}_{\mathcal{T},\Sigma})$.

To prove Theorem 43, we start by establishing three helpful lemmas.

Lemma 44. *$\mathcal{I}_{\mathcal{T},\Sigma}$ is a model of $\mathcal{T}$ and $\mathcal{A}_{\mathcal{T},\Sigma}$.*

Proof. It is straightforward to prove by induction on the structure of C that for all $C \in \mathsf{cl}(\mathcal{T},\Sigma)$, we have $C \in t$ iff $t \in C^{\mathcal{I}_{\mathcal{T},\Sigma}}$. By definition of types, $C \sqsubseteq D \in \mathcal{T}$ and $C \in t$ implies $D \in t$. Thus, $\mathcal{I}_{\mathcal{T},\Sigma}$ is clearly a model of $\mathcal{T}$. It is an immediate consequence of the definition of $\mathcal{A}_{\mathcal{T},\Sigma}$ that $\mathcal{I}_{\mathcal{T},\Sigma}$ is also a model of $\mathcal{A}_{\mathcal{T},\Sigma}$. ❑

Definition 45. Let $\mathcal{A}$ and $\mathcal{A}'$ be literal ABoxes. An *ABox homomorphism* from $\mathcal{A}$ to $\mathcal{A}'$ is a total map $h : \mathsf{Ind}(\mathcal{A}) \to \mathsf{Ind}(\mathcal{A}')$ such that the following conditions are satisfied:

- $A(a) \in \mathcal{A}$ implies $A(h(a)) \in \mathcal{A}'$;
- $r(a,b) \in \mathcal{A}$ implies $r(h(a), h(b)) \in \mathcal{A}'$.

Lemma 46. *If $\mathcal{T}$ is an $\mathcal{ALCI}$-TBox, $\mathcal{A}$ a Σ-ABox, $A_0 \in \mathsf{N_C}$, $a_0 \in \mathsf{Ind}(\mathcal{A})$, and h an ABox homomorphism from $\mathcal{A}$ to $\mathcal{A}'$, then $\mathcal{T}, \mathcal{A} \models A_0[a_0]$ implies $\mathcal{T}, \mathcal{A}' \models A_0[h(a_0)]$.*

Proof. We prove the contrapositive. Thus assume that $\mathcal{T}, \mathcal{A}' \not\models A_0[h(a_0)]$. Then there is a model $\mathcal{I}'$ of $\mathcal{T}$ and $\mathcal{A}'$ such that $\mathcal{I}' \not\models A_0[h(a_0)]$. Define a model $\mathcal{I}$ by starting with $\mathcal{I}'$ and reinterpreting the individual names in $\mathsf{Ind}(\mathcal{A})$ by setting $a^{\mathcal{I}} = h(a)^{\mathcal{I}'}$ for each $a \in \mathsf{Ind}(\mathcal{A})$. Clearly, $\mathcal{I}$ is still a model of $\mathcal{T}$. It is also a model of $\mathcal{A}$: if $A(a) \in \mathcal{A}$, then $A(h(a)) \in \mathcal{A}'$ by definition of ABox homomorphisms; since $\mathcal{I}'$ is a model of $\mathcal{A}'$ and by definition of $\mathcal{I}$, it follows that $a^{\mathcal{I}} \in A^{\mathcal{I}}$. The case $r(a,b) \in \mathcal{A}$ is analogous. Finally, $\mathcal{I}' \not\models A_0[a_0]$ and the definition of $\mathcal{I}$ yield $\mathcal{I} \not\models A_0[a_0]$. We have thus shown that $\mathcal{T}, \mathcal{A} \not\models A_0[a_0]$. ❑

Lemma 47. *Let $\mathcal{T}$ be an $\mathcal{ALCI}$-TBox, $\mathcal{A}$ a Σ-ABox that is consistent w.r.t. $\mathcal{T}$, $A_0 \in \mathsf{N_C}$, and $a_0 \in \mathsf{Ind}(\mathcal{A})$ such that $\mathcal{T}, \mathcal{A} \not\models A_0[a_0]$. Then there is an ABox homomorphism h from $\mathcal{A}$ to $\mathcal{A}_{\mathcal{T},\Sigma}$ such that $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A_0(h(a_0))$.*

Proof. Let $\mathcal{I}$ be a model of $\mathcal{T}$ and $\mathcal{A}$ such that $a^{\mathcal{I}} \notin A_0^{\mathcal{I}}$. For each $a \in \mathsf{Ind}(\mathcal{A})$, define $t_a^{\mathcal{I}} = \{C \in \mathsf{cl}(\mathcal{T},\Sigma) \mid a^{\mathcal{I}} \in C^{\mathcal{I}}\}$. Define h by setting $h(a) = a_t$ with $t = t_a^{\mathcal{I}}$ for all $a \in \mathsf{Ind}(\mathcal{A})$. Using the definition of $\mathcal{A}_{\mathcal{T},\Sigma}$, it is easy to see that h is indeed an ABox homomorphism. Let $t = t_{a_0}^{\mathcal{I}}$. By choice of $\mathcal{I}$, we have $A_0 \notin t$, hence $t \notin A_0^{\mathcal{I}_{\mathcal{T},\Sigma}}$. Since $h(a_0) = a_t$, the definition of $\mathcal{A}_{\mathcal{T},\Sigma}$ yields $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A_0(h(a_0))$ as required. ❑

We are now ready to prove Theorem 43.

Proof. (of Theorem 43) The "only if" direction is trivial. For the "if" direction, let $\mathsf{cert}_{\mathcal{T}}(A_1(x), \mathcal{A}_{\mathcal{T},\Sigma}) \subseteq \mathsf{cert}_{\mathcal{T}}(A_2(x), \mathcal{A}_{\mathcal{T},\Sigma})$. To show that $A_1(x) \sqsubseteq_{\mathcal{T},\Sigma} A_2(x)$, take a Σ-ABox $\mathcal{A}$ that is consistent with $\mathcal{T}$. Assume to the contrary of what is to be shown that there is an $a_0 \in \mathsf{Ind}(\mathcal{A})$ such that $\mathcal{T}, \mathcal{A} \models A_1[a_0]$ and $\mathcal{T}, \mathcal{A} \not\models A_2[a_0]$. By Lemma 47, there is an ABox homomorphism h from $\mathcal{A}$ to $\mathcal{A}_{\mathcal{T},\Sigma}$ such that $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A_2[h(a_0)]$, and it follows from Lemma 46 that we also have $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \models A_1[h(a_0)]$, which gives the desired contradiction. ❑

Theorem 20. IQ-query containment in $\mathcal{ALCI}$ is in $\mathrm{P}^{\mathrm{NExp}}$.

Proof. By Theorem 3, it is sufficient to consider single TBox containment. We show that non-containment is in $\mathrm{NP}^{\mathrm{NExp}}$ and derive from $\mathrm{NP}^{\mathrm{NExp}} \subseteq \mathrm{P}^{\mathrm{NExp}}$ the desired result (Hemachandra 1987).

Let $\mathcal{T}$ be a consistent $\mathcal{ALCI}$-TBox, Σ an ABox signature, and $A_1(x)$, $A_2(x)$ IQs such that it is to be decided whether $A_1(x) \sqsubseteq_{\mathcal{T},\Sigma} A_2(x)$ is *not* the case. The algorithm guesses $a \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma})$, and then checks (1) $a \in \mathsf{cert}_{\mathcal{T}}(A_1(x), \mathcal{A}_{\mathcal{T},\Sigma})$ and (2) $a \notin \mathsf{cert}_{\mathcal{T}}(A_2(x), \mathcal{A}_{\mathcal{T},\Sigma})$, by calling a NEXPTIME oracle that decides the following problem: given an $\mathcal{ALCI}$-TBox $\mathcal{T}'$, signature Σ', individual $a' \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T}',\Sigma'})$ and IQ $A(x)$, does $a' \notin \mathsf{cert}_{\mathcal{T}'}(A(x), \mathcal{A}_{\mathcal{T}',\Sigma'})$ hold? The algorithm accepts if checks (1) and (2) both succeed, and otherwise, it rejects.

Assuming the described NEXPTIME oracle exists, the above algorithm clearly runs in $\mathrm{NP}^{\mathrm{NExp}}$, and its correctness follows immediately from Theorem 43. Thus, to complete the proof, we only need to demonstrate the existence of the described oracle:

Claim. Given an $\mathcal{ALCI}$-TBox $\mathcal{T}'$, signature Σ', individual $a' \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T}',\Sigma'})$ and IQ $A(x)$, it can be decided in NEXPTIME whether $a' \notin \mathsf{cert}_{\mathcal{T}'}(A(x), \mathcal{A}_{\mathcal{T}',\Sigma'})$.

Proof of claim. We give a non-deterministic exponential time procedure for deciding $a' \notin \mathsf{cert}_{\mathcal{T}'}(A(x), \mathcal{A}_{\mathcal{T}',\Sigma'})$. The first step is to compute in deterministic single-exponential time the canonical ABox $\mathcal{A}_{\mathcal{T}',\Sigma'}$. Next the procedure guesses a map $\pi : \mathsf{Ind}(\mathcal{A}_{\mathcal{T}',\Sigma'}) \to \mathfrak{T}_{\mathcal{T}',\Sigma'}$. Finally the procedure verifies that the following conditions are satisfied: (i) $A \notin \pi(a')$, (ii) $C(c) \in \mathcal{A}_{\mathcal{T}',\Sigma'}$ implies $C \in \pi(c)$, and (iii) $r(b,c) \in \mathcal{A}_{\mathcal{T}',\Sigma'}$, $C \in \pi(c)$, and $\exists r.C \in \mathsf{cl}(\mathcal{T}',\Sigma')$ implies $\exists r.C \in \pi(b)$. It accepts if the verification succeeds, and rejects otherwise.

We need to show that the above procedure yields the desired result. First, suppose that the algorithm accepts. Then there is a mapping $\pi : \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma}) \to \mathfrak{T}_{\mathcal{T},\Sigma}$ that satisfies

conditions (i) to (iii) above. Since every type in $\mathfrak{T}_{\mathcal{T},\Sigma}$ is satisfiable w.r.t. $\mathcal{T}$, we can take models of $\pi(b)$ and $\mathcal{T}$ for each $b \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma})$, and assemble them into a model $\mathcal{I}_{a'}$ of $\mathcal{T}$ and $\mathcal{A}_{\mathcal{T},\Sigma}$ such that $a'^{\mathcal{I}_a} \notin A^{\mathcal{I}_a}$. Thus $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A(a')$.

For the second direction, suppose $a' \notin \mathsf{cert}_{\mathcal{T}'}(A(x), \mathcal{A}_{\mathcal{T}',\Sigma'})$. Since $\mathcal{T}, \mathcal{A}_{\mathcal{T},\Sigma} \not\models A(a')$, we find some model $\mathcal{I}_a$ of $\mathcal{T}$ and $\mathcal{A}_{\mathcal{T},\Sigma}$ such that $a'^{\mathcal{I}_a} \notin A^{\mathcal{I}_a}$. So we can guess the mapping π such that $\pi(b)$ is the type of $b^{\mathcal{I}_a}$ in the model $\mathcal{I}_a$ for each $b \in \mathsf{Ind}(\mathcal{A}_{\mathcal{T},\Sigma})$, i.e. $\pi(b) = \{C \mid b^{\mathcal{I}_a} \in C^{\mathcal{I}_a} \text{ and } C \in \mathsf{cl}(\mathcal{T}, \Sigma)\}$. By construction, π satisfies Conditions (i) to (iii) and thus the verification results in acceptance. ❑

F Proofs for Section 7

We start by making more precise how a CQ q can be viewed as an ABox $\mathcal{A}_q$. For every variable $x \in \mathsf{N_V}$, we use $\mathsf{ind}(x)$ to denote the ABox individual $a_x \in \mathsf{N_I}$ and for every $a \in \mathsf{N_I}$, set $\mathsf{ind}(a) = a$. Then every atom $C(t) \in q$ gives rise to an ABox assertion $C(\mathsf{ind}(t)) \in \mathcal{A}_q$ and likewise for every atom $r(t, t') \in q$ and the ABox assertion $r(\mathsf{ind}(t), \mathsf{ind}(t')) \in \mathcal{A}_q$. These are the only atoms in $\mathcal{A}_q$.

Lemma 48. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox and q_1, q_2 CQs with answer variables $x_1, \dots, x_n$ that are consistent with $\mathcal{T}$. Then $q_1 \subseteq_{\mathcal{T}} q_2$ iff $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \dots, a_{x_n}]$.*

Proof. "if". Assume $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \dots, a_{x_n}]$ and let $\mathcal{I}_{q_1}$ be the canonical model of $\mathcal{A}_{q_1}$ and $\mathcal{T}$. By Lemma 29, there is an $a_{x_1}, \dots, a_{x_n}$-match π of q_2 in $\mathcal{I}_{q_1}$. Take an ABox $\mathcal{A}$ which is consistent with $\mathcal{T}$ and a tuple $a_1, \dots, a_n$ such that $\mathcal{T}, \mathcal{A} \models q_1[a_1, \dots, a_n]$. We have to show that $\mathcal{T}, \mathcal{A} \models q_2[a_1, \dots, a_n]$. Let $\mathcal{I}$ be the canonical model of $\mathcal{A}$ and $\mathcal{T}$ and τ an $a_1, \dots, a_n$-match of q_1 in $\mathcal{I}$. Consider the following function h:

- $h(a_x) = \tau(x)$ for all variables $x \in \mathsf{term}(q_1)$;
- $h(a) = a$ for all individual names $a \in \mathsf{term}(q_1)$;
- $h(\mathsf{ind}(t)c_{R_1} \cdots c_{R_n}) = h(\mathsf{ind}(t))c_{R_1} \cdots c_{R_n}$ for all $t \in \mathsf{term}(q_1)$ and $c_{R_1} \cdots c_{R_n}$, $n \geq 1$, with $\mathsf{ind}(t)c_{R_1} \cdots c_{R_n} \in \Delta^{\mathcal{I}_{q_1}}$.

Using the construction of $\mathcal{I}_{q_1}$ and $\mathcal{I}$, it can be shown that the range of h is contained in $\Delta^{\mathcal{I}}$ and that h is a homomorphism from $\mathcal{I}_{q_1}$ to $\mathcal{I}$. Moreover, since $\tau(x_i) = a_i$ for each answer variable x_i, we have $h(a_{x_i}) = a_i$. The composition of the match τ and the homomorphism h thus yields an $a_1, \dots, a_n$-match of q_2 in $\mathcal{I}$. It follows that $\mathcal{T}, \mathcal{A} \models q_2[a_1, \dots, a_n]$, thus Lemma 29 yields $\mathcal{T}, \mathcal{A} \models q_2[a_1, \dots, a_n]$ as required.

"only if". $q_1 \subseteq_{\mathcal{T}} q_2$ implies $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \dots, a_{x_n}]$ since, clearly, $\mathcal{T}, \mathcal{A}_{q_1} \models q_1[a_{x_1}, \dots, a_{x_n}]$. ❑

Theorem 23. Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox and q_1, q_2 CQs such that $q_1 \equiv_{\mathcal{T}} q_2$ and q_1, q_2 are $\mathcal{T}$-minimal w.r.t. set inclusion and consistent with $\mathcal{T}$. Then $\#q_1 = \#q_2$.

Proof. We show that $\#q_1 \leq \#q_2$. The converse direction is symmetric. Let $x_1, \dots, x_n$ be the answer variables in q_1 and q_2. By Lemma 48, $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \dots, a_{x_n}]$. Let $\mathcal{I}_{q_1}$ be the canonical model of $\mathcal{A}_{q_1}$ and $\mathcal{T}$. By Lemma 29, there is an $a_{x_1}, \dots, a_{x_n}$-match π of q_2 in $\mathcal{I}_{q_1}$.

For what follows, it is convenient to view $\mathcal{I}_{q_1}$ as an ABox that contains an assertion $A(d)$ whenever $d \in A^{\mathcal{I}_{q_1}}$ and an assertion $r(d, e)$ whenever $(d, e) \in r^{\mathcal{I}_{q_1}}$. By construction of $\mathcal{I}_{q_1}$, for each of the resulting assertions as we find a *single* atom $c(\mathsf{as}) \in q_1$ such that $c(\mathsf{as}) \in q$ 'causes' the presence of as in $\mathcal{I}_{q_1}$ in the sense that as is part of the canonical model of $\{c(\mathsf{as})\}$ and $\mathcal{T}$ (there may be multiple candidates for $c(\mathsf{as})$, in which case we choose an arbitrary one).

For each atom $\mathsf{at} \in q_2$, the match π identifies a corresponding assertion $\pi(\mathsf{at})$ in $\mathcal{I}_{q_1}$:

- if $\mathsf{at} = A(t)$ and $\pi(t) = d$, then $\pi(\mathsf{at}) = A(d) \in \mathcal{I}_{q_1}$;
- if $\mathsf{at} = r(t, t')$, $\pi(t) = d$, and $\pi(t') = e$, then $\pi(\mathsf{at}) = r(d, e) \in \mathcal{I}_{q_1}$.

Let q_1' denote the query $\{c(\pi(\mathsf{at})) \mid \mathsf{at} \in q_2\} \subseteq q_1$. Since $\#q_1' \leq \#q_2$, to show that $\#q_1 \leq \#q_2$ it suffices to show that $q_1' = q_1$. Assume that this is not the case, i.e., $q_1' \subsetneq q_1$. Let $\mathcal{I}_{q_1'}$ be the canonical model of $\mathcal{A}_{q_1'}$ and $\mathcal{T}$. By construction of q_1', $\mathcal{A}_{q_1'}$, and $\mathcal{I}_{q_1'}$, we have that $\pi(\mathsf{at}) \in \mathcal{I}_{q_1'}$ for all $\mathsf{at} \in q_2$. It follows that π is an $a_{x_1}, \dots, a_{x_n}$-match of q_2 in $\mathcal{I}_{q_1'}$ and, by Lemma 29, $\mathcal{T}, \mathcal{A}_{q_1'} \models q_2[a_{x_1}, \dots, a_{x_n}]$. By Lemma 48, we obtain $q_1' \subseteq_{\mathcal{T}} q_2$ and the transitivity of "$\subseteq_{\mathcal{T}}$" yields $q_1' \subseteq_{\mathcal{T}} q_1$, in contradiction to q_1 being $\mathcal{T}$-minimal w.r.t. set inclusion. ❑

Note that, from here on, we are working with queries and ABoxes that can contain atoms of the form $\exists r(t)$ and assertions of the form $\exists r(a)$. The definition of canonical models can be extended to this case in a straightforward way and analogues of Lemmas 29 and 48 are easily established. The notion of a query match in a model is also easily generalized.

To establish Theorems 25 and 26, we first show that our minimization strategy preserves rootedness of queries. This is the object of the following lemma. For convenience, we introduce the notation $\mathsf{qvar}(q)$ to denote the set of quantified variables in a query q.

Lemma 49. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox, q a rooted CQ, and $\widehat{q}$ the query resulting from applying the minimization strategy to q. Then $\widehat{q}$ is either rooted or contains no atoms.*

Proof. Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox, q_0 a rooted CQ, and $\widehat{q}_0$ the query resulting from applying the minimization strategy to q_0. It is easy to see that Step 1 preserves rootedness, since no atoms are removed. This means we begin Step 2 with a rooted query which is equivalent to the input query q_0 w.r.t. $\mathcal{T}$. Note that during Step 2, the set of answer variables and individuals names does not change. To see why, note that whenever an atom $r(t, t')$ is removed, the query will still contain concept atoms $\exists r(t)$ and $\exists r^-(t')$ because of Step 1. It is possible though for the query to become disconnected during Step 2. The following claim shows however that there is always one connected subquery which is equivalent to q_0 w.r.t. $\mathcal{T}$ and contains all of the answer variables and individual names in q_0.

Claim 1. Suppose a query q possesses a rooted subquery q_s such that $q \equiv_{\mathcal{T}} q_s$ and $\mathsf{term}(q) \setminus \mathsf{qvar}(q) \subseteq \mathsf{term}(q_s)$. If $r(t, t') \in q$ is such that $q \equiv_{\mathcal{T}} q \setminus \{r(t, t')\}$, then $q \setminus \{r(t, t')\}$ also possesses a rooted subquery q_s' such that $q \equiv_{\mathcal{T}} q_s'$ and $\mathsf{term}(q) \setminus \mathsf{qvar}(q) \subseteq \mathsf{term}(q_s')$.

Proof of claim: Let q and q_s be as in the claim. We can suppose without loss of generality that q_s is a maximally connected subquery of q. If $r(t,t') \notin q_s$, then q_s is the desired subquery. Let us then consider the more interesting case where $r(t,t') \in q_s$. Let $q' = q \setminus \{r(t,t')\}$, and let $q_1, \ldots, q_n$ be the maximally connected subqueries of q'. As $q' \equiv_{\mathcal{T}} q$, by Lemma 48, $\mathcal{T}, \mathcal{A}_{q'} \models q[a_{x_1}, \ldots, a_{x_n}]$, where $x_1, \ldots, x_n$ are the answer variables of q' and q. Then by Lemma 29, there exists an $a_{x_1}, \ldots, a_{x_n}$-match π of q in the canonical model $\mathcal{I}_{q'}$ of $\mathcal{A}_{q'}$ and $\mathcal{T}$. As $q' = \cup_i q_i$ and the terms of the q_i are pairwise disjoint, it follows from the definition of canonical models that $\mathcal{I}_{q'}$ is precisely the disjoint union of the canonical models of the $\mathcal{A}_{q_i}$ and $\mathcal{T}$. More precisely, we have:

- $\Delta^{\mathcal{I}_{q'}} = \cup_i \Delta^{\mathcal{I}_{q_i}}$
- $A^{\mathcal{I}_{q'}} = \cup_i A^{\mathcal{I}_{q_i}}$
- $r^{\mathcal{I}_{q'}} = \cup_i r^{\mathcal{I}_{q_i}}$

As $q_s \subseteq q$ is rooted and contains at least one atom, it must be connected. Thus, there must be a single q_i such that every term in q_s (hence all answer variables and individual names in q) is mapped by π to $\Delta^{\mathcal{I}_{q_i}}$. It follows that $q_i \subseteq_{\mathcal{T}} q_s$. Then we can use the fact that $q_s \subseteq_{\mathcal{T}} q$ to infer $q_i \subseteq_{\mathcal{T}} q$, hence $q_i \equiv_{\mathcal{T}} q$. The query q_i is thus the desired subquery of q'. This completes the proof of the claim.

At the beginning of Step 2, we start with a query q satisfying the conditions of Claim 1 (simply take q itself as the rooted subquery). We can thus use Claim 1 to show that at the end of Step 2, we obtain a query q' which has a rooted subquery q_s such that $q_s \equiv_{\mathcal{T}} q' \equiv_{\mathcal{T}} q_0$ and $\mathsf{term}(q') \setminus \mathsf{qvar}(q') = \mathsf{term}(q_0) \setminus \mathsf{qvar}(q_0) \subseteq \mathsf{term}(q_s)$. We remark that if $r(t,t') \in q'$, then we must have $r(t,t') \in q_s$, since otherwise we would have $q' \equiv_{\mathcal{T}} q' \setminus \{r(t,t')\}$, contradicting the fact that q' has no redundant role atoms. Thus, all role atoms in q' must belong to q_s. We next show that the minimized query $\widehat{q}_0$ resulting from Step 3 is a subset of q_s. For this, we start by proving the following claim:

Claim 2. Suppose a query q possesses a rooted subquery q_s such that $q \equiv_{\mathcal{T}} q_s$ and $\mathsf{term}(q) \setminus \mathsf{qvar}(q) \subseteq \mathsf{term}(q_s)$. If $C(t) \in q_s$ is such that $q \equiv_{\mathcal{T}} q \setminus \{C(t)\}$, then $q \equiv_{\mathcal{T}} q_s \setminus \{C(t)\}$.

Proof of claim: Let q and q_s be as in the claim, and suppose without loss of generality that q_s is a maximally connected subquery of q. Take $C(t) \in q_s$ with $q \equiv_{\mathcal{T}} q \setminus \{C(t)\}$. Set $q^- = q \setminus \{C(t)\}$ and $q_s^- = q \setminus \{C(t)\}$. Since $q \equiv_{\mathcal{T}} q^-$, by Lemma 48, $\mathcal{T}, \mathcal{A}_{q^-} \models q[a_{x_1}, \ldots, a_{x_n}]$. Then by Lemma 29, there exists an $a_{x_1}, \ldots, a_{x_n}$-match π of q in the canonical model $\mathcal{I}_{q^-}$ of $\mathcal{A}_{q^-}$ and $\mathcal{T}$. Because of our assumption that q_s is maximally connected, we can partition q^- into q_s^- and q_r, such that $\mathsf{term}(q_s^-) \cap \mathsf{term}(q_r) = \emptyset$. From the way canonical models are constructed, we know that $\mathcal{I}_{q^-}$ is the disjoint union of the canonical models $\mathcal{I}_{q_s^-}$ and $\mathcal{I}_{q_r}$. Because $q_s \subseteq q$ is connected and contains at least one answer variable or individual name, it follows that π maps all terms in q_s to $\Delta^{q_s^-}$. It follows that $q_s \setminus \{C(t)\} = q_s^- \subseteq_{\mathcal{T}} q_s$, which when combined with $q_s \setminus q$ yields the desired $q \equiv_{\mathcal{T}} q_s \setminus \{C(t)\}$.

We now return to showing $\widehat{q}_0 \subseteq q_s$. Suppose for a contradiction that $\widehat{q}_0 \setminus q_s \neq \emptyset$. First consider the case where $\widehat{q}_0 \cap q_s \neq \emptyset$. Note that at the start of Step 3, the query q' and subquery q_s satisfy all the conditions of Claim 2. We can thus repeatedly apply Claim 2 to show that at the end of Step 3, we have $\widehat{q}_0 \cap q_s \equiv_{\mathcal{T}} \widehat{q}_0$. Since $\widehat{q}_0 \setminus q_s$ contains only concept atoms (see above), this implies that $\widehat{q}_0$ contains a redundant concept atom, a contradiction.

Now consider the other case in which we have both $\widehat{q}_0 \setminus q_s \neq \emptyset$ and $\widehat{q}_0 \cap q_s \neq \emptyset$. As no role assertions are removed during Step 3, it follows that q_s consists of concept atoms for some term t. Moreover, we must have $\{t\} = \mathsf{term}(q_0) \setminus \mathsf{qvar}(q_0)$, since all answer variables and individuals in q_0 belong to q_s. Let $C(t)$ be the last atom in q_s removed during Step 3. Since t is not a quantified variable, the removal of $C(t)$ can only occur if $\mathcal{T} \models \top \sqsubseteq C$. Because of Claim 2, we know that $C(t) \subseteq_{\mathcal{T}} q' \setminus q_s$. First suppose that t is a constant (i.e. q_0 is Boolean). Then by Lemmas 48 and 29, there is a match for $q' \setminus q_s$ in the canonical model $\mathcal{I}_{C(t)}$ of $\{C(t)\}$ and $\mathcal{T}$. Now take any ABox $\mathcal{A}$ consistent with $\mathcal{T}$, and let a be some individual appearing in $\mathcal{A}$. Then $\mathcal{T}, \mathcal{A} \models C(a)$, hence there is a homomorphism from $\mathcal{I}_{C(t)}$ to $\mathcal{I}_{\mathcal{T},\mathcal{A}}$. It follows that there is a match for $q' \setminus q_s$ in $\mathcal{I}_{\mathcal{T},\mathcal{A}}$, which yields $\mathcal{T}, \mathcal{A} \models q' \setminus q_s$. Thus, we have shown that for every ABox $\mathcal{A}$, we have $\mathcal{T}, \mathcal{A} \models q' \setminus q_s$. It follows that every atom in $\widehat{q}$ is redundant, a contradiction. Finally, suppose instead that t is an answer variable (hence the unique answer variable in q_0). Then by Lemmas 48 and 29, there is an a_t-match for $q' \setminus q_s$ in $\mathcal{I}_{C(t)}$, the canonical model of $\{C(a_t)\}$ and $\mathcal{T}$ (here $q' \setminus q_s$ still has t as answer variable, even though no atoms mention t). Using similar arguments to above, we can show that for any ABox $\mathcal{A}$ consistent with $\mathcal{T}$, and any individual a appearing in $\mathcal{A}$, we have $\mathcal{T}, \mathcal{A} \models q' \setminus q_s[a]$. It follows that $q' \setminus q_s$ is equivalent to the empty query with one answer variable, contradicting the assumed minimality of $\widehat{q}$.

We have thus shown that $\widehat{q}_0 \subseteq q_s$. It follows that either $\widehat{q}_0$ has no atoms, or else it is a connected query containing some answer variable or individual name, hence rooted.

❑

Our next step will be to characterize the result of applying the CQ minimization strategy given in the paper in a more abstract way. For CQs q_1 and q_2, we write $q_1 \prec q_2$ if q_1 can be obtained from q_2 by one of the following operations:

- drop a concept atom $C(t)$;
- replace a concept atom $C(t)$ with $D(t)$ when $\mathcal{T} \models C \equiv D$ and $D < C$;
- replace a concept atom $C(t)$ with $D(t)$ when $\mathcal{T} \models D \sqsubseteq C$, but not $\mathcal{T} \models C \sqsubseteq D$;
- drop role atoms, and then add zero or more concept atoms $C(t)$ with $\mathsf{sig}(C) \subseteq \mathsf{sig}(\mathcal{T})$ and $t \in \mathsf{term}(q_2)$.

We say that a CQ q_1 is *strongly $\mathcal{T}$-minimal w.r.t. set inclusion* if there is no CQ q_2 such that $q_2 \prec q_1$ and $q_1 \equiv_{\mathcal{T}} q_2$. It is not hard to show that any CQ produced according to our minimization strategy is strongly $\mathcal{T}$-minimal w.r.t. set inclusion.

For the next lemma, we extend the notion of a homomorphism h from a CQ q to a CQ q' to handle atoms of the form $\exists R(t)$. We use the following condition for an atom $\exists R(t)$: if $\exists R(t) \in q$, then $\exists R(h(t)) \in q'$.

Lemma 50. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox and q_1, q_2 rooted CQs that are strongly $\mathcal{T}$-minimal w.r.t. set inclusion and consistent with $\mathcal{T}$, and such that $q_1 \equiv_{\mathcal{T}} q_2$. Then there is a bijective homomorphism from q_2 to q_1.*

Proof. By Lemma 48, $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \ldots, a_{x_n}]$, where $x_1, \ldots, x_n$ are the answer variables of q_1 and q_2. By Lemma 29, there thus is an $a_{x_1}, \ldots, a_{x_n}$-match π of q_2 in the canonical model $\mathcal{I}_{q_1}$ of $\mathcal{A}_{q_1}$ and $\mathcal{T}$.

Let $\mathsf{hit}(\pi)$ be the set of all $a \in \mathsf{Ind}(\mathcal{A}_{q_1})$ such that some element $ac_{R_1} \cdots c_{R_k} \in \Delta^{\mathcal{I}_{q_1}}$, $k \geq 0$, is in the range of π.

Claim 1. $\mathsf{Ind}(\mathcal{A}_{q_1}) \subseteq \mathsf{hit}(\pi)$.

Proof of Claim 1: By definition of matches, we have $\pi(x_i) = a_{x_i}$ for each answer variable x_i. Moreover, $q_1 \equiv_{\mathcal{T}} q_2$ and strong $\mathcal{T}$-minimality of q_1 and q_2 implies that $\mathsf{term}(q_1)$ and $\mathsf{term}(q_2)$ contain the same individual names and, by definition of matches, we have $\pi(a) = a$ for each such individual name a. As a consequence, it remains to show that all $a_x \in \mathsf{Ind}(\mathcal{A}_{q_1})$ with x a quantified variable are in $\mathsf{hit}(\pi)$. Note that since q_1 is rooted, x occurs in at least one role atom in q_1. Let q_1' be the restriction of q_1 to the terms $\mathsf{term}(q_1) \setminus \{x\}$ extended with

1. the concept atom $A(t)$ whenever $t \in \mathsf{term}(q_1) \setminus \{x\}$, $A \in \mathsf{N_C}$ occurs in $\mathcal{T}$, and $\mathsf{ind}(t) \in A^{\mathcal{I}_{q_1}}$;
2. the concept atom $\exists r(t)$ whenever $t \in \mathsf{term}(q_1) \setminus \{x\}$, r occurs in $\mathcal{T}$, and $\mathsf{ind}(t) \in (\exists r)^{\mathcal{I}_{q_1}}$

Since q_1' is obtained from q_1 by dropping at least one role atom, and possibly also some concept atoms, and then adding new concept atoms on the remaining variables, we must have $q_1' \prec q_1$.

We show that π is an $a_{x_1}, \ldots, a_{x_n}$-match of q_2 also in the canonical model $\mathcal{I}_{q_1'}$ of $\mathcal{A}_{q_1'}$ and $\mathcal{T}$. By choice of q_1' and construction of canonical models, we have $\pi(t) \in \Delta^{\mathcal{I}_{q_1'}}$ for each $t \in \mathsf{term}(q_2)$. It remains to show that all atoms in q_2 are satisfied:

- $C(t) \in q_2$.
  First assume $\pi(t) \in \mathsf{Ind}(\mathcal{A}_{q_1})$. Then there is a $t' \in \mathsf{term}(q_1')$ with $\mathsf{ind}(t') = \pi(t)$. Since π is a match in $\mathcal{I}_{q_1}$, we have $\pi(t) \in C^{\mathcal{I}_{q_1}}$. Thus, the extension step in the construction of q_1' ensures that $C(t') \in q_1'$, which yields $\pi(t) \in C^{\mathcal{I}_{q_1'}}$ as required. Next suppose that $\pi(t) \notin \mathsf{Ind}(\mathcal{A}_{q_1})$, i.e., $\pi(t)$ has the form $ac_{R_1} \cdots c_{R_k}$ with $k > 0$. There is a $t' \in \mathsf{term}(q_1')$ with $\mathsf{ind}(t') = a$. By construction of $\mathcal{I}_{q_1}$, we find in q_1 an atom at of the form $C(t')$, $r(\widehat{t}, t')$, or $r(t', \widehat{t})$ such that $ac_{R_1} \cdots c_{R_k}$ belongs to the canonical model of $\{\mathsf{at}\}$ and $\mathcal{T}$ and is an instance of C in this model. By the extension of q_1', there is an atom $C(t') \in q_1'$ such that $ac_{R_1} \cdots c_{R_k}$ belongs to the canonical model of $\{C(\mathsf{ind}(t'))\}$ and $\mathcal{T}$ and is an instance of C in this model. By construction of $\mathcal{I}_{q_1'}$, we thus have $\pi(t) = ac_{R_1} \cdots c_{R_k} \in C^{\mathcal{I}_{q_1'}}$ as required.
- $r(t_1, t_2) \in q_2$.
  First assume $\pi(t_1), \pi(t_2) \in \mathsf{Ind}(\mathcal{A}_{q_1})$. There is a $t_i' \in \mathsf{term}(q_1')$ with $\mathsf{ind}(t_i') = \pi(t_i)$ for each $i \in \{1, 2\}$. Since π is a match in $\mathcal{I}_{q_1}$, we have $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{q_1}}$. By construction of $\mathcal{I}_{q_1}$, this implies $r(t_1', t_2') \in q_1$. It follows that $r(t_1', t_2') \in q_1'$, which yields $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{q_1'}}$ as required.
  Now assume that at least one of $\pi(t_1), \pi(t_2)$ is not in $\mathsf{Ind}(\mathcal{A}_{q_1})$. By definition of $\mathcal{A}_{q_1}$ and construction of $\mathcal{I}_{q_1}$, there is a term $t' \in \mathsf{term}(q_1)$ such that $\mathsf{ind}(t') \in \mathsf{hit}(\pi)$ and an atom at of the form $C(t')$, $s(\widehat{t}, t')$, or $s(t', \widehat{t})$ such that $\pi(t_1)$ and $\pi(t_2)$ both belong to the canonical model of $\{\mathsf{at}\}$ and $\mathcal{T}$ and $\pi(t_1)$ and $\pi(t_2)$ are related by r in this model. If x appears in no role atoms, then at belongs to q_1', and so we are done. Otherwise, we use the fact that because of the addition of new concept atoms, there must exist some atom $D(t') \in q_1'$ with the same property as at and thus, by construction of $\mathcal{I}_{q_1'}$, we have $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{q_1'}}$ as required.

By Lemmas 29 and 48, we have $q_1' \subseteq_{\mathcal{T}} q_2$, which implies $q_1' \equiv_{\mathcal{T}} q_1$, in contrast to the strong $\mathcal{T}$-minimality of q_1. This finishes the proof of the Claim 1.

By two applications of Claim 1, we obtain the following property which will prove useful later in the proof:

Property 2. $|\mathsf{term}(q_1)| = |\mathsf{term}(q_2)|$.

Claim 1 established that every individual a in $\mathcal{A}_{q_1}$ is such that some $ac_{r_1} \ldots c_{r_k}$ is in the range of the match π of q_2 in the canonical model $\mathcal{I}_{q_1}$. The next claim shows that only individuals, not paths, appear in the range of π. We use $\mathsf{ran}(\pi)$ to denote the range of π.

Claim 3. $\mathsf{ran}(\pi) \subseteq \mathsf{Ind}(\mathcal{A}_{q_1})$.

Proof of Claim 3: By definition of matches, we have $\mathsf{ran}(\pi) \not\subseteq \mathsf{Ind}(\mathcal{A}_{q_1})$ iff there is a quantified variable $x \in \mathsf{term}(q_2)$ with $\pi(x) \notin \mathsf{Ind}(\mathcal{A}_{q_1})$. Assume to the contrary of what is to be shown that there is such an x. Then $\pi(x) = ac_{R_1} \ldots c_{R_k}$ with $k \geq 1$. Since q_2 is rooted, we know that there is some answer variable or constant $t \in \mathsf{term}(q_2)$. Rootedness also ensures that q_2 is connected, which means there is a sequence of role atoms in q_2 connecting t and x. Since $\pi(t) = t \in \mathsf{Ind}(\mathcal{A}_{q_1})$, it follows that there is some $t' \in \mathsf{term}(q)$ such that $\pi(t') = a$ (we may have $t = t'$). Thus, at least two terms in $\mathsf{term}(q_2)$ are mapped to paths starting by a, namely x and t'. But Property 2 tells us that $|\mathsf{term}(q_1)| = |\mathsf{term}(q_2)|$, so there must be some $b \in \mathsf{Ind}(\mathcal{A}_{q_1})$ which does not belong to $\mathsf{hit}(\pi)$, contradicting Claim 1. As we have reached a contradiction, it follows that there can be no such x, yielding the desired $\mathsf{ran}(\pi) \subseteq \mathsf{Ind}(\mathcal{A}_{q_1})$.

It follows from Claim 3 that $\mathsf{hit}(\pi) = \mathsf{ran}(\pi)$. Combining Claims 1 and 3, we thus get $\mathsf{Ind}(\mathcal{A}_{q_1}) = \mathsf{ran}(\pi)$. By Property 2, we have $|\mathsf{term}(q_1)| = |\mathsf{term}(q_2)|$ and thus the function h defined by setting $h(t) = t'$ when $\pi(t) = \mathsf{ind}(t')$ for all $t \in \mathsf{term}(q_2)$ is a bijection from $\mathsf{term}(q_2)$ to $\mathsf{term}(q_1)$. To show that h is a bijective homomorphism from q_2 to q_1, it thus remains to establish the following:

- $A(t) \in q_2$ implies $A(h(t)) \in q_1$.
  $A(t) \in q_2$ implies $\pi(t) \in A^{\mathcal{I}_{q_1}}$. As $\pi(t) \in \mathsf{Ind}(\mathcal{A}_{q_1})$, by definition of $\mathcal{I}_{q_1}$ one of the following three cases must hold:

– $C(h(t)) \in q_1$ for some concept C with $\mathcal{T} \models C \sqsubseteq A$.
– $r(h(t), t') \in q_1$ for some t' and r with $\mathcal{T} \models \exists r \sqsubseteq A$.
– $r(t', h(t)) \in q_1$ for some t' and r with $\mathcal{T} \models \exists r^- \sqsubseteq A$.

However, we will now show that in fact only the first case is possible. Suppose we are in the second case: $r(h(t), t') \in q_1$ for some t' and r with $\mathcal{T} \models \exists r \sqsubseteq A$. Let u be such that $h(u) = t'$. We claim that $r(t, u) \in q_2$. Suppose for a contradiction that this is not the case. Let q_1' be the query $q_1 \setminus \{r(h(t), t')\}$, extended with

1. the concept atom $B(t)$ whenever $B \in \mathsf{N_C}$ occurs in $\mathcal{T}$ and $\mathsf{ind}(t) \in B^{\mathcal{I}_{q_1}}$;
2. the concept atom $\exists S(t)$ whenever S occurs in $\mathcal{T}$ and $\mathsf{ind}(t) \in (\exists S)^{\mathcal{I}_{q_1}}$.

Then it is easily verified that π is a match of q_2 in $\mathcal{I}_{q_1'}$, which means that $q_1' \subseteq_{\mathcal{T}} q_2$, hence $q_1' \equiv_{\mathcal{T}} q_1$. This contradicts the strong $\mathcal{T}$-minimality of q_1. So we indeed have $r(t, u) \in q_2$. Now, however, $A(t) \in q_2$ yields a contradiction against the strong $\mathcal{T}$-minimality of q_2. The impossibility of the third case is shown analogously.

Thus, we must have $C(h(t)) \in q_1$ for some concept C with $\mathcal{T} \models C \sqsubseteq A$. Strong $\mathcal{T}$-minimality of q_1 and $C(h(t)) \in q_1$ yields that C is $<$-minimal among all concepts that are equivalent to C w.r.t. $\mathcal{T}$. Strong $\mathcal{T}$-minimality of q_2 and $A(t) \in q_2$ yields the same for A. It thus remains to show that A and C are equivalent w.r.t. $\mathcal{T}$. Assume they are not. Then we can replace $A(t)$ by $C(t)$ in q_2 and argue that the resulting query is $\mathcal{T}$-equivalent to q_2 (as it still has a match in $\mathcal{I}_{q_1}$), in contradiction to its strong $\mathcal{T}$-minimality. Thus, A and C are equivalent w.r.t. $\mathcal{T}$, and since they are both minimal, they must be identical, yielding $A(h(t)) \in q_1$.

- $\exists r(t) \in q_2$ implies $\exists r(h(t)) \in q_1$.
  Argument proceeds analogously to the previous case.
- $r(t_1, t_2) \in q_2$ implies $r(h(t), h(t_2)) \in q_1$.
  Clear since $r(t_1, t_2) \in q_2$ implies $(\pi(t_1), \pi(t_2)) \in r^{\mathcal{I}_{q_1}}$, which yields $r(h(t_1), h(t_2)) \in q_1$ by definition of $\mathcal{I}_{q_1}$ and the fact that $\pi(t_1), \pi(t_2) \in \mathsf{Ind}(\mathcal{A}_{q_1})$.

❑

For Theorem 25, we first note that if one of the minimized queries is empty, then the other must be empty too. Thus, the only interesting case is when both of the minimized queries are non-empty, hence rooted (by Lemma 49). Because minimized queries are always strongly $\mathcal{T}$-minimal, it suffices to prove the following.

Theorem 51. *Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox and q_1, q_2 rooted CQs such that $q_1 \equiv_{\mathcal{T}} q_2$ and q_1, q_2 are strongly $\mathcal{T}$-minimal w.r.t. set inclusion and consistent w.r.t. $\mathcal{T}$. Then q_1 and q_2 are isomorphic.*

Proof. By Lemma 50, there is a bijective homomorphism h from q_1 to q_2. It suffices to show that the inverse of h is a homomorphism from q_2 to q_1. Let q_2' be the restriction of q_2 to the atoms hit by h, i.e.,

$$q_2' = \{A(h(t)) \mid A(t) \in q_1\} \cup \{r(h(t_1), h(t_2)) \mid r(t_1, t_2) \in q_1\}.$$

It suffices to show that $q_2' = q_2$. Clearly, $|q_2'| \leq |q_2|$. Since h is injective, q_2' is isomorphic to q_1, and thus $|q_1| = |q_2'|$. It follows that $|q_1| \leq |q_2|$ and we can argue symmetrically using the fact that there is an injective homomorphism also from q_2 to q_1 to show that $|q_2| \geq |q_1|$. Thus $|q_2| = |q_2'|$, which yields $q_2 = q_2'$ as required. ❑

We now prove Theorem 26 from the paper.

Theorem 26. Let $\mathcal{T}$ be a DL-Lite$_{\mathsf{core}}$-TBox, q_1, q_2 rooted CQs with $q_1 \equiv_{\mathcal{T}} q_2$ that are consistent with $\mathcal{T}$, and let $\widehat{q}_1$ be obtained from q_1 by the minimization strategy. Then $\widehat{q}_1^{\,r}$ is isomorphic to a subquery of q_2.

Proof. Let $\mathcal{T}$, $q_1, q_2, \widehat{q}_1$ be as in the statement of the theorem. If $\widehat{q}_1$ is empty, then the theorem trivially holds. So suppose that $\widehat{q}_1$ is non-empty (hence rooted by Lemma 49), and let $\widehat{q}_2$ be obtained from q_2 by the minimization strategy. Clearly $\widehat{q}_2$ must also be non-empty, hence rooted. Thus, we can apply Theorem 25 to show that $\widehat{q}_1$ and $\widehat{q}_2$ are isomorphic. It follows that $\widehat{q}_1^{\,r}$ and $\widehat{q}_2^{\,r}$ are isomorphic. To complete the proof, we simply note that $\widehat{q}_2^{\,r} \subseteq q_2$. ❑

The proof of the following lemma is exactly analogous to the proof of the DL-Lite version (Lemma 48), based on canonical models for $\mathcal{EL}$ and Lemma 31. Details are left to the reader.

Lemma 52. *Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox and q_1, q_2 CQs with answer variables $x_1, \ldots, x_n$. Then $q_1 \subseteq_{\mathcal{T}} q_2$ iff $\mathcal{T}, \mathcal{A}_{q_1} \models q_2[a_{x_1}, \ldots, a_{x_n}]$.*

Lemma 27. Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox and q a CQ that is not acyclic, but such that $q \equiv_{\mathcal{T}} p$ for some acyclic CQ p. Then q contains a fork whose elimination yields a query q' with $q \equiv_{\mathcal{T}} q'$.

Proof. Let $\mathcal{I}_p$ be the canonical model of $\mathcal{A}_p$ and $\mathcal{T}$. By Lemma 52, q has an $a_{x_1}, \ldots, a_{x_n}$-match π in $\mathcal{I}_p$, where $x_1, \ldots, x_n$ are the answer variables of q and p. Since q contains a cycle and $\mathcal{I}_p$ is acyclic by construction, there must be a t_1, t_2-fork in q such that $\pi(t_1) = \pi(t_2)$. Let $b_1, \ldots, b_k$ be the individual names in $\mathsf{term}(q)$. Since $a_{x_1}^{\mathcal{I}_p}, \ldots, a_{x_n}^{\mathcal{I}_p}, b_1^{\mathcal{I}_p}, \ldots, b_k^{\mathcal{I}_p}$ are pairwise distinct by definition of $\mathcal{I}_p$, at least one of t_1, t_2 is a quantified variable. Define q' as q, but with t_2 replaced by t_1 if t_2 is a quantified variable, and t_1 replaced by t_2 otherwise. Clearly, q and q' have the same answer variables. We show that $q' \equiv_{\mathcal{T}} q$. Since $q' \subseteq_{\mathcal{T}} q$ is obvious, it is enough to show $q \subseteq_{\mathcal{T}} q'$.

Let $\mathcal{A}$ be an ABox and assume that $\mathcal{T}, \mathcal{A} \models q[a_1, \ldots, a_n]$. Then $\mathcal{T}, \mathcal{A} \models p[a_1, \ldots, a_n]$. We have to show that $\mathcal{T}, \mathcal{A} \models q'[a_1, \ldots, a_n]$. Let $\mathcal{I}$ be the canonical model of $\mathcal{A}$ and $\mathcal{T}$ and τ an $a_1, \ldots, a_n$-match of p in $\mathcal{I}$. Consider the following function h:

- $h(a_x) = \tau(x)$ for all variables $x \in \mathsf{term}(p)$;
- $h(a) = a$ for all individual names $a \in \mathsf{term}(p)$;
- $h(\mathsf{ind}(t) r_1 C_1 r_2 \cdots r_\ell C_\ell) = h(\mathsf{ind}(t)) r_1 C_1 r_2 \cdots r_\ell C_\ell$ for all $t \in \mathsf{term}(p)$ and $\mathsf{ind}(t) r_1 C_1 r_2 \cdots r_\ell C_\ell \in \Delta^{\mathcal{I}_p}$.

Using the construction of $\mathcal{I}_p$ and $\mathcal{I}$, it can be shown that the range of h is contained in $\Delta^{\mathcal{I}}$ and that h is a homomorphism from $\mathcal{I}_p$ to $\mathcal{I}$. Moreover, since $\tau(x_i) = a_i$ for each

answer variable x_i, we have $h(a_{x_i}) = a_i^{\mathcal{I}}$ for each i. The composition of π and h thus yields an $a_1, \ldots, a_n$-match π' of q in $\mathcal{I}$. Since $\pi(x_1) = \pi(x_2)$, we have $\pi'(x_1) = \pi'(x_2)$. It follows that π' is a match for q', thus Lemma 52 yields $\mathcal{T}, \mathcal{A} \models q'[a_1, \ldots, a_n]$ as required. ❑

We now detail the second step of the minimization procedure for $\mathcal{EL}$ which was mentioned in the paper. For a CQ q, the *$\mathcal{T}$-decoration* of q is the query $\widehat{q}$ obtained from q by adding for every subconcept D in $\mathcal{T}$ and all $t \in \mathsf{term}(q)$, the atom $D(t)$ if $\mathcal{A}_q, \mathcal{T} \models D(\mathsf{ind}(t))$. The second step of our overall query minimization strategy for $\mathcal{EL}$ consists in first switching from q to its $\mathcal{T}$-decoration $\widehat{q}$, and then exhaustively dropping quantified variables such that $\mathcal{T}$-equivalence is preserved, in any order. Here, *dropping* a variable x from a CQ p means to replace p with the restriction $p|_{\overline{x}}$ of p to the terms $\mathsf{term}(p) \setminus \{x\}$.

The following lemma shows that, when following the described strategy, we obtain an equivalent query with a minimum number of variables. We say that a CQ p is *minimal regarding variable dropping* if $p \not\equiv_{\mathcal{T}} p|_{\overline{x}}$ for any quantified variable $x \in \mathsf{term}(p)$.

Lemma 53. *Let $\mathcal{T}$ be an $\mathcal{EL}$-TBox and q_1, q_2 CQs such that $q_1 \equiv_{\mathcal{T}} q_2$ and $\widehat{q}_1$ is minimal regarding variable dropping. Then $|\mathsf{term}(\widehat{q}_1)| \leq |\mathsf{term}(q_2)|$.*

Proof. Let q_1 and q_2 be as in the lemma and assume to the contrary of what is to be shown that $|\mathsf{term}(q_2)| < |\mathsf{term}(\widehat{q}_1)|$. Let $\mathcal{I}_{\widehat{q}_1}$ be the canonical model of $\mathcal{A}_{\widehat{q}_1}$ and $\mathcal{T}$. By Lemma 52, we find an $a_{x_1}, \ldots, a_{x_n}$-match π of q_2 in $\mathcal{I}_{\widehat{q}_1}$, where $x_1, \ldots, x_n$ are the answer variables of q_1, $\widehat{q}_1$, and q_2.

Let $c : \mathsf{term}(q_2) \to \mathsf{term}(\widehat{q}_1)$ be defined by setting, for all $t \in \mathsf{term}(q_2)$:

- if $\pi(t) = a_x$ with x a variable from $\mathsf{term}(\widehat{q}_1)$, then set $c(t) = x$;
- if $\pi(t) = a$ with a an individual name from $\mathsf{term}(\widehat{q}_1)$, then set $c(t) = a$;
- if $\pi(t) = \mathsf{ind}(t')r_1C_1r_2 \cdots r_\ell C_\ell$ with $t' \in \mathsf{term}(\widehat{q}_1)$, then set $c(t) = t'$.

Let $\widehat{q}_1'$ denote the restriction of $\widehat{q}_1$ to the terms in the range of c. Since $|\mathsf{term}(q_2)| < |\mathsf{term}(\widehat{q}_1)|$, we clearly have $|\mathsf{term}(\widehat{q}_1')| < |\mathsf{term}(\widehat{q}_1)|$, i.e., $\widehat{q}_1'$ is obtained from $\widehat{q}_1$ by dropping one or more quantified variables (note that all answer variables and individual names from $\mathsf{term}(\widehat{q}_1)$ are in the range of c). To obtain a contradiction, it thus suffices to show that $\widehat{q}_1' \equiv_{\mathcal{T}} \widehat{q}_1$. Since $\widehat{q}_1 \subseteq_{\mathcal{T}} \widehat{q}_1'$ is obvious, it remains to show $\widehat{q}_1' \subseteq_{\mathcal{T}} \widehat{q}_1$, or equivalently, $\widehat{q}_1' \subseteq_{\mathcal{T}} \widehat{q}_2$.

By Lemma 52, it suffices to prove that q_2 has an $a_{x_1}, \ldots, a_{x_n}$-match in the canonical model $\mathcal{I}_{\widehat{q}_1'}$ of $\mathcal{A}_{\widehat{q}_1'}$ and $\mathcal{T}$. We do this by showing that the match π of q_2 in $\mathcal{I}_{\widehat{q}_1}$ is also a match of q_2 in $\mathcal{I}_{\widehat{q}_1'}$. We begin by proving that the anonymous tree below each element $\mathsf{ind}(t)$ with $t \in \mathsf{term}(\widehat{q}_1')$ is identical in $\mathcal{I}_{\widehat{q}_1}$ and $\mathcal{I}_{\widehat{q}_1'}$. For this, it is sufficient to show that for every subconcept C in $\mathcal{T}$ and every term t in $\widehat{q}_1'$, we have $\mathcal{T}, \mathcal{A}_{\widehat{q}_1} \models C(\mathsf{ind}(t))$ if and only if $\mathcal{T}, \mathcal{A}_{\widehat{q}_1'} \models C(\mathsf{ind}(t))$. As $\widehat{q}_1'$ is a subquery of $\widehat{q}_1$, the second implication is trivial, so we concentrate on the first. Suppose that $\mathcal{T}, \mathcal{A}_{\widehat{q}_1} \models C(\mathsf{ind}(t))$, where C is a subconcept of $\mathcal{T}$ and $t \in \mathsf{term}(\widehat{q}_1')$. By the definition of $\mathcal{T}$-decoration, we must have $C(t) \in \widehat{q}_1$. As $t \in \mathsf{term}(\widehat{q}_1')$, it follows that $C(t) \in \widehat{q}_1'$, which yields $\mathcal{T}, \mathcal{A}_{\widehat{q}_1'} \models C(\mathsf{ind}(t))$.

Now we return to showing that the match π of q_2 in $\mathcal{I}_{\widehat{q}_1}$ defines a match of q_2 in $\mathcal{I}_{\widehat{q}_1'}$. First we note that for every $t \in \mathsf{term}(q_2)$, $\pi(t)$ must belong to $\Delta^{\mathcal{I}_{\widehat{q}_1'}}$, since $\pi(t)$ is a path whose first element is $\mathsf{ind}(t')$ for some $t' \in \mathsf{term}(\widehat{q}_1')$, and by the above arguments, the anonymous tree below such an element $\mathsf{ind}(t')$ is the same in $\mathcal{I}_{\widehat{q}_1}$ and $\mathcal{I}_{\widehat{q}_1'}$. Next consider some atom $A(t) \in q_2$. Since π is a match of q_2 in $\mathcal{I}_{\widehat{q}_1}$, we must have $\pi(t) \in A^{\mathcal{I}_{\widehat{q}_1}}$. If A is a subconcept of $\mathcal{T}$, then by above, we must have $\pi(t) \in A^{\mathcal{I}_{\widehat{q}_1'}}$. If A does not appear in $\mathcal{T}$, then $\pi(t) \in A^{\mathcal{I}_{\widehat{q}_1}}$ implies that $\pi(t) = \mathsf{ind}(t')$ for some term t' in $\widehat{q}_1$ such that $A(t') \in \widehat{q}_1$. This means we also have $A(t') \in \widehat{q}_1'$, yielding $\mathsf{ind}(t') = \pi(t) \in A^{\mathcal{I}_{\widehat{q}_1'}}$. Now consider some term $r(t, t') \in q_2$. If both $\pi(t)$ and $\pi(t')$ are mapped to the core of $\mathcal{I}_{\widehat{q}_1}$, then it follows from the definition of canonical models, the ABox $\mathcal{A}_{\widehat{q}_1}$, and the function c that $r(c(t), c(t')) \in \widehat{q}_1$. This yields $r(c(t), c(t')) \in \widehat{q}_1'$, and hence $(\pi(t), \pi(t')) \in r^{\mathcal{I}_{\widehat{q}_1'}}$. Otherwise, it must be the case that $\pi(t)$ and $\pi(t')$ both belong to the anonymous tree below some element $\mathsf{ind}(t')$ with t' in the image of c. Because $\mathcal{I}_{\widehat{q}_1'}$ contains exactly the same anonymous tree, it follows that $(\pi(t), \pi(t')) \in r^{\mathcal{I}_{\widehat{q}_1'}}$. This completes our proof that π is a match of q_2 in $\mathcal{I}_{\widehat{q}_1'}$, which yields the desired contradiction by the argument above. ❑

Finally we note that dropping variables will never make an acyclic query cyclic. It follows that by applying the two steps of the minimization procedure, we obtain a $\mathcal{T}$-equivalent query with a minimal number of variables which is acyclic whenever a $\mathcal{T}$-equivalent acyclic query exists.