Query Answering in DL-Lite with Datatypes: A Non-Uniform Approach

André Hernich and **Julio Lemos** and **Frank Wolter**
University of Liverpool
Department of Computer Science
{hernich, jlemos, wolter}@liverpool.ac.uk

Abstract

Adding datatypes to ontology-mediated queries (OMQs) often makes query answering hard. As a consequence, the use of datatypes in OWL 2 QL has been severely restricted. In this paper we propose a new, non-uniform, way of analyzing the data-complexity of OMQ answering with datatypes. Instead of restricting the ontology language we aim at a classification of the patterns of datatype atoms in OMQs into those that can occur in non-tractable OMQs and those that only occur in tractable OMQs. To this end we establish a close link between OMQ answering with datatypes and constraint satisfaction problems over the datatypes. In a case study we apply this link to prove a P/coNP-dichotomy for OMQs over DL-Lite extended with the datatype $(\mathbb{Q}, \leq)$. The proof employs a recent dichotomy result by Bodirsky and Kára for temporal constraint satisfaction problems.

1 Introduction

In recent years, querying data using ontologies has become one of the main applications of description logics (DLs). The general idea is that an ontology is used to enrich incomplete and heterogenous data with a semantics and with background knowledge and thereby serves as an interface for querying data that also allows the derivation of additional facts. In this area called ontology-based data management (OBDM) one of the main research problems is to identify ontology languages and queries for which query answering scales to large amounts of data (Calvanese et al. 2007; Bienvenu and Ortiz 2015). In DL, ontologies take the form of a TBox, data is stored in an ABox, and the most important class of queries are (unions of) conjunctive queries, or simply (U)CQs. A basic observation regarding this setup is that even for DLs from the DL-Lite family that have been designed for tractable OBDM the addition of datatypes to the TBoxes or the UCQs easily leads to non-tractable query answering problems (Artale, Ryzhikov, and Kontchakov 2012; Savkovic and Calvanese 2012). As a consequence of this, the use of datatypes in TBoxes and query languages for OBDM has been severely restricted (Motik and Horrocks 2008; Motik et al. 2009). In applications, however, there is clearly a need for expressive datatypes both in TBoxes and in queries.

The aim of this paper is to revisit OBDM with expressive datatypes from a new, non-uniform perspective. Instead of the standard approach that aims at the definition of DLs $\mathcal{L}$ and query languages $\mathcal{Q}$ such that for any TBox $\mathcal{T}$ in $\mathcal{L}$ and any query q in $\mathcal{Q}$, answering q under $\mathcal{T}$ is tractable in data-complexity we now aim at describing the complexity of query answering with datatypes at a more fine-grained level by taking into account the way in which datatype atoms can occur in TBoxes and in queries. To this end, we establish a close link between the complexity of query answering and of constraint satisfaction problems (CSPs) over the datatype. This link enables us to transfer complexity results from the CSP world to the world of OBDM and leads, in some cases, to complete classifications of the complexity of query answering into PTime and coNP-complete classes.

In more detail, we consider TBoxes in the DL DL-Lite$_\mathcal{R}$ underpinning the OWL profile OWL 2 QL extended with concept inclusions that contain attribute restrictions qualified by unary datatype atoms on their right hand side and UCQs that contain datatype atoms of arbitrary arity. If $\mathcal{T}$ is such a TBox over datatype $\mathcal{D}$ and q such a UCQ over $\mathcal{D}$, then $Q = (\mathcal{T}, q)$ is called an ontology-mediated query (OMQ) over $\mathcal{D}$. We aim at understanding the complexity of query answering for this very broad class of OMQs. A first observation is that query answering becomes undecidable for many important datatypes $\mathcal{D}$ including $(\mathbb{Q}, <)$, $(\mathbb{Z}, <)$ and $(\mathbb{Z}, \leq)$. To restore decidability and enable a polynomial reduction to the complement of CSPs over $\mathcal{D}$ we introduce the bounded match depth property (BMDP), a new property of OMQs that ensures that answers to OMQs can be determined based on a bounded subset of the standard chase of a DL-Lite$_\mathcal{R}$ knowledge base. This property is a generalization of the bounded derivation depth property in (Calì, Gottlob, and Lukasiewicz 2012). Many practical OMQs have the BMDP. For example, all OMQs with either TBoxes whose chase always terminates (which is often the case in practice (Grau et al. 2013)) or with rooted UCQs whose variables are all connected via non-datatype variables to answer variable (which covers a broad class of UCQs). If the datatype $\mathcal{D}$ is homogeneous (as is the case for $(\mathbb{Q}, <)$ and $(\mathbb{Q}, \leq)$), then the latter condition can be relaxed even further to certain Boolean UCQs. As the CSP of many important datatypes is in NP, it follows that query answering for OMQs with the BMDP over such datatypes is in coNP, a significant im-

provement compared to the undecidable OMQs without the BMDP. To sharpen the link between OMQ answering and CSP further, we also provide a converse polynomial reduction of CSPs over a datatype $\mathcal{D}$ to the complement of answering OMQs with BMDP over $\mathcal{D}$. This converse reduction can thus be used to transfer NP-hardness results from the CSP world to coNP-hardness results for OMQ answering. More importantly, however, we now have a framework for transfering *complexity classification results* from CSP to OMQ answering.

We illustrate the power of this framework for the datatype $(\mathbb{Q}, \leq)$. Note first that even without qualified attribute restrictions OMQs over datatype $(\mathbb{Q}, \leq)$ can express many interesting queries.

Example 1.1. Let $\mathcal{T} = \{\exists p \sqsubseteq \exists U, \exists p^- \sqsubseteq \exists U\}$ be a TBox, where p is a role name and U an attribute. Then the Boolean CQ

$$q \leftarrow p(x, y) \wedge U(x, u) \wedge U(y, v) \wedge u \leq v$$

is entailed by a KB $(\mathcal{T}, \mathcal{A})$ over $(\mathbb{Q}, \leq)$ such that $\mathcal{A}$ is an ABox containing no assertions using U iff $\mathcal{A}$ contains a p-cycle. Thus, answering the OMQ $(\mathcal{T}, q)$ is NLogSpace-complete. We also construct OMQs over $(\mathbb{Q}, \leq)$ that are PTime-complete and, respectively, coNP-complete.

Our main result is a P/coNP-dichotomy for OMQs over $(\mathbb{Q}, \leq)$ with the BMDP. To formulate the dichotomy we associate with every OMQ $Q = (\mathcal{T}, q)$ a datatype pattern $\mathsf{dtype}(Q) = (\theta_\mathcal{T}, \theta_q)$ such that $\theta_\mathcal{T}$ contains the datatype atoms in $\mathcal{T}$ and θ_q contains the datatype atoms in q. In Example 1.1, $\theta_\mathcal{T} = \emptyset$ and $\theta_q = \{u \leq v\}$. Then, based on the framework introduced above and a recent P/NP-dichotomy result for temporal CSPs (Bodirsky and Kára 2010a) we show that for any datatype pattern θ exactly one of the following two conditions holds (unless P=coNP):

- Evaluating OMQs Q with BMDP and $\mathsf{dtype}(Q) = \theta$ is always in PTime.
- There exists an OMQ Q with BMDP and $\mathsf{dtype}(Q) = \theta$ whose evaluation problem is coNP-hard.

In addition, our dichotomy comes with a purely syntactic description of the datatype patterns that lead to OMQs that are in PTime. For example, the datatype pattern in Example 1.1 will always lead to an OMQ in PTime.

Related Work Expressive DLs with datatypes (or concrete domains) have been introduced in (Baader and Hanschke 1991) and studied extensively (Lutz 2002). In the context of tractable DLs, reasoning with datatypes has been studied in (Baader, Brandt, and Lutz 2005; Magka, Kazakov, and Horrocks 2011) for $\mathcal{EL}$ and in (Poggi et al. 2008; Savkovic and Calvanese 2012; Artale, Ryzhikov, and Kontchakov 2012) for DL-Lite. These works focus on finding ontology languages for which typical reasoning tasks are tractable. In contrast, here we start with ontology languages for which query answering is intractable in general, and aim at a complexity classification of query answering guided by the datatype pattern. Our methodology is closely related to recent work relating OBDM to constraint satisfaction problems (Lutz and Wolter 2012; Bienvenu et al. 2014; Hernich et al. 2015). However, here we classify datatype patterns according to the data-complexity of evaluating the OMQs containing them, whereas in (Lutz and Wolter 2012) TBoxes are classified according to the data-complexity of OMQs containing them, and in (Bienvenu et al. 2014) OMQs themselves are classified according to their data-complexity. Consequently, here we establish a link to temporal constraint satisfaction (Bodirsky and Kára 2010a) whereas the work mentioned above establishes a link to standard constraint satisfaction and the Feder-Vardi conjecture (Feder and Vardi 1993; Bulatov, Jeavons, and Krokhin 2005).

Detailed proofs are provided in the full version of this paper.

2 Preliminaries

A *datatype* is a tuple $\mathcal{D} = (D, R_1, R_2, \dots)$, where D is a non-empty set and $R_1, R_2, \dots$ are relations on D. We call $\mathrm{dom}(\mathcal{D}) := D$ the *domain* of $\mathcal{D}$. To simplify the presentation, we will not distinguish between a relation R_i and its name (i.e., we use R_i both as a relation and as the name of the relation R_i).

A *primitive positive (PP) formula* over $\mathcal{D}$ is a first-order formula φ built from atomic formulas over $\mathcal{D}$ using conjunction and existential quantification. If φ has no free variables, then φ is called a *PP sentence* over $\mathcal{D}$. By $\mathsf{CSP}(\mathcal{D})$ we denote the problem of deciding whether a given PP sentence over $\mathcal{D}$ is satisfied in $\mathcal{D}$. PP formulas over $\mathcal{D}$ do not use domain elements of $\mathcal{D}$ as constants. If we admit such constants we speak about PP formulas or sentences over $\mathcal{D}$ *with constants* and denote the satisfiction problem for such PP sentences by $\mathsf{CSP}_c(\mathcal{D})$.

We assume countably infinite and mutually disjoint sets of *concept names*, *role names*, *attribute names*, and *individual names*. *Roles* r and *basic concepts* B are defined by the rules

$$r := p \mid p^- \quad B := A \mid \exists r \mid \exists U$$

where p ranges over all role names, A ranges over all concept names, and U ranges over all attribute names. A *DL-Lite*$_\mathcal{R}^{attrib}(\mathcal{D})$ *TBox* is a finite set of *role inclusions* $r_1 \sqsubseteq r_2$, *attribute inclusions* $U_1 \sqsubseteq U_2$, and *concept inclusions* $B_1 \sqsubseteq B_2$ and $B_1 \sqsubseteq \neg B_2$, where r_1, r_2 range over roles, U_1, U_2 over attributes, and B_1, B_2 over basic concepts. In TBoxes of the extension *DL-Lite*$_\mathcal{R}^{qattrib}(\mathcal{D})$ of *DL-Lite*$_\mathcal{R}^{attrib}(\mathcal{D})$ one can use, in addition, *qualified attribute restrictions* on the right hand side of inclusions:

$$B \sqsubseteq \exists U.\varphi, \quad B \sqsubseteq \forall U.\varphi$$

where B is a basic concept, U is an attribute, and φ is a PP formula over $\mathcal{D}$ with constants and a single free variable x.

Let $\mathcal{D}$ be a datatype. A *$\mathcal{D}$-ABox* consists of *assertions* of the form $A(a)$, $p(a, b)$, and $U(a, u)$, where A is a concept name, p is a role name, U is an attribute name, a, b are individual names, and $u \in \mathrm{dom}(\mathcal{D})$. A *$\mathcal{D}$-knowledge base ($\mathcal{D}$-KB)* is a pair $(\mathcal{T}, \mathcal{A})$ consisting of a *DL-Lite*$_\mathcal{R}^{qattrib}(\mathcal{D})$ TBox $\mathcal{T}$ and a $\mathcal{D}$-ABox $\mathcal{A}$.

An *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ over a datatype $\mathcal{D}$ consists of a non-empty domain $\Delta^{\mathcal{I}} = \Delta^{\mathcal{I}}_{\mathsf{ind}} \cup \mathrm{dom}(\mathcal{D})$ and an interpretation function $\cdot^{\mathcal{I}}$ that assigns to each concept name A a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}_{\mathsf{ind}}$, to each role name p a relation $p^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}_{\mathsf{ind}} \times \Delta^{\mathcal{I}}_{\mathsf{ind}}$, and to each attribute name U a relation $U^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}_{\mathsf{ind}} \times \mathrm{dom}(\mathcal{D})$. The elements in $\Delta^{\mathcal{I}}_{\mathsf{ind}}$ are called *individuals*, whereas the elements in $\mathrm{dom}(\mathcal{D})$ are called *data values*. We assume that $\Delta^{\mathcal{I}}_{\mathsf{ind}}$ and $\mathrm{dom}(\mathcal{D})$ are disjoint. Throughout this paper, we make the *standard name assumption*: if $\mathcal{I}$ is an interpretation, then we set $a^{\mathcal{I}} := a$ for all individual names a. We also set $u^{\mathcal{I}} := u$ for each $u \in \mathrm{dom}(\mathcal{D})$, and $R^{\mathcal{I}} := R$ for each relation R of $\mathcal{D}$. The interpretation $\mathcal{I}$ induces the interpretations $B^{\mathcal{I}}$ and $r^{\mathcal{I}}$ for each basic concept B and role r in the standard way (Artale et al. 2009). The qualified attribute restrictions are interpreted as follows:

$$(\exists U.\varphi)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}}_{\mathsf{ind}} \mid \exists s\, ((a,s) \in U^{\mathcal{I}} \,\&\, \mathcal{D} \models \varphi(s))\},$$
$$(\forall U.\varphi)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}}_{\mathsf{ind}} \mid \forall s\, ((a,s) \in U^{\mathcal{I}} \Rightarrow \mathcal{D} \models \varphi(s))\}.$$

The interpretation $\mathcal{I}$ is called *model* of a $\mathcal{D}$-ABox $\mathcal{A}$ if $a \in A^{\mathcal{I}}$, $(a,b) \in p^{\mathcal{I}}$, and $(a,u) \in U^{\mathcal{I}}$ for all assertions $A(a)$, $p(a,b)$, and $U(a,u)$ in $\mathcal{A}$. It is called *model* of a *DL-Lite*$^{qattrib}_{\mathcal{R}}(\mathcal{D})$ TBox $\mathcal{T}$ if $X^{\mathcal{I}} \subseteq Y^{\mathcal{I}}$ for every inclusion $X \sqsubseteq Y \in \mathcal{T}$.

An interpretation is a model of a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$ if it is a model of both $\mathcal{A}$ and $\mathcal{T}$. A $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$ is *satisfiable* if it has a model; in this case we say that $\mathcal{A}$ is *satisfiable relative to* $\mathcal{T}$.

Conjunctive queries (CQs) q over $\mathcal{D}$ take the form $q(\bar{x}) \leftarrow \varphi$, where $\bar{x}$ is the tuple of *answer variables* of q, and φ is a conjunction of atomic formulas of the form $A(y)$, $p(y,z)$, $U(y,u)$, or $R(u_1,\ldots,u_k)$, where A, p, U, and R range over concept names, role names, attribute names, and relation names in $\mathcal{D}$, respectively; each y, z and each $u, u_1, \ldots, u_k$ is a *variable* and the variables $u, u_1, \ldots, u_k$ are called *data variables*. As usual, all variables of $\bar{x}$ must occur in some atom of φ. A *match* of q in an interpretation $\mathcal{I}$ is a mapping μ from the variables of φ to $\Delta^{\mathcal{I}}$ such that for each atom $X(t_1,\ldots,t_k)$ of φ we have $(\mu(t_1),\ldots,\mu(t_k)) \in X^{\mathcal{I}}$. A tuple $\bar{c}$ of individual names and data values is an *answer* to q in an interpretation $\mathcal{I}$ if there is a match μ of q in $\mathcal{I}$ such that $\mu(\bar{x}) = \bar{c}$. We denote this by $\mathcal{I} \models q(\bar{c})$.

A *union of conjunctive queries (UCQ)* q over $\mathcal{D}$ takes the form $q_1(\bar{x}), \ldots, q_n(\bar{x})$, where each $q_i(\bar{x})$ is a CQ over $\mathcal{D}$. The q_i are called *disjuncts* of q. A tuple $\bar{c}$ of individual names and data values is an *answer* to q in an interpretation $\mathcal{I}$, denoted by $\mathcal{I} \models q(\bar{c})$, if $\bar{c}$ is an answer to some disjunct of q in $\mathcal{I}$.

Given a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$, a UCQ q over $\mathcal{D}$, and a tuple $\bar{c}$ of individual names and data values, we write $\mathcal{T}, \mathcal{A} \models q(\bar{c})$ if $\bar{c}$ is an answer to q in every model of $(\mathcal{T}, \mathcal{A})$.

An *ontology-mediated query (OMQ)* over $\mathcal{D}$ takes the form $Q = (\mathcal{T}, q)$, where $\mathcal{T}$ is a *DL-Lite*$^{qattrib}_{\mathcal{R}}(\mathcal{D})$ TBox and q is a UCQ over $\mathcal{D}$. Given a $\mathcal{D}$-ABox $\mathcal{A}$ and a tuple $\bar{c}$, we write $\mathcal{A} \models Q(\bar{c})$ if $\mathcal{T}, \mathcal{A} \models q(\bar{c})$.

When studying the OMQ answering problem we focus on data complexity, unless otherwise stated.

3 Universal Pre-Models

Many ontology languages for which query evaluation is tractable admit the construction of universal models of KBs using some variant of the chase procedure. This applies to DL-Lite$_{\mathcal{R}}$ as well. In contrast, KBs in its extension *DL-Lite*$^{qattrib}_{\mathcal{R}}(\mathcal{D})$ often do not have universal models and the aim of this section is to introduce instead models with placeholders for data values that play a similar role as universal models, but modulo the assignment of data values to the placeholders. The following examples illustrates why universal models do not always exist.

Example 3.1. Consider the KB $(\mathcal{T}, \mathcal{A})$ over the datatype $(\mathbb{Q}, \leq)$ with $\mathcal{T} = \{A \sqsubseteq \exists U_1, A \sqsubseteq \exists U_2\}$. and $\mathcal{A} = \{A(a)\}$. Consider the OMQs $Q_i = (\mathcal{T}, q_i)$, for $i \in \{1,2\}$, where

$$q_1(x) \leftarrow U_1(x,u_1) \wedge U_2(x,u_2) \wedge u_1 \leq u_2,$$
$$q_2(x) \leftarrow U_1(x,u_1) \wedge U_2(x,u_2) \wedge u_2 \leq u_1.$$

Clearly $\mathcal{A} \not\models Q_1(a)$ since for the interpretation $\mathcal{I}$ with $U_1^{\mathcal{I}} = \{(a,2)\}$ and $U_2^{\mathcal{I}} = \{(a,1)\}$ we have $\mathcal{I} \not\models q_1(a)$. Also, $\mathcal{A} \not\models Q_2(a)$ since for the interpretation $\mathcal{J}$ with $U_1^{\mathcal{J}} = \{(a,1)\}$ and $U_2^{\mathcal{J}} = \{(a,2)\}$ we have $\mathcal{J} \not\models q_2(a)$. However, there does not exist a model $\mathcal{I}$ of $\mathcal{T}$ and $\mathcal{A}$ such that $\mathcal{I} \not\models q_i(a)$ for both $i = 1$ and $i = 2$.

The reason that universal models do not exist is that distinct interpretations of attributes can be required to refute the entailment of CQs or UCQs. The notion of pre-interpretation formalizes this intuition: it fixes the interpretation of concept and roles names but leaves the interpretation of attributes open by adding placeholders for data values (called data nulls) to the set of possible values of attributes.

Fix a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$. A *pre-interpretation* $\mathcal{J}$ over $\mathcal{D}$ is the same as an interpretation over $\mathcal{D}$ with the exception that attribute names U are now interpreted as sets $U^{\mathcal{J}} \subseteq \Delta^{\mathcal{J}}_{\mathsf{ind}} \times (\mathrm{dom}(\mathcal{D}) \cup \Delta^{\mathcal{J}}_{\mathsf{null}})$, where $\Delta^{\mathcal{J}}_{\mathsf{null}}$ is a set of *data nulls* disjoint from $\Delta^{\mathcal{J}}_{\mathsf{ind}} \cup \mathrm{dom}(\mathcal{D})$, and that for each $u \in \Delta^{\mathcal{J}}_{\mathsf{null}}$ we fix a set $Z^{\mathcal{J}}(u)$ of PP formulas φ that occur in qualified attribute restrictions of $\mathcal{T}$. Intuitively, the formulas in $Z^{\mathcal{J}}(u)$ restrict the possible data values in $\mathrm{dom}(\mathcal{D})$ that can be assigned to u. The definitions of the interpretations $C^{\mathcal{J}}$ of a concept C and $S^{\mathcal{J}}$ of a role or attribute S are extended from interpretations to pre-interpretations in the straightforward way. A *pre-model* of a KB is a pre-interpretation that satisfies all assertions and inclusions in the KB.

Pre-interpretations $\mathcal{J}$ can be completed to interpretations by assigning data values to data nulls. A *completion function* f for $\mathcal{J}$ is a mapping $f\colon \Delta^{\mathcal{J}}_{\mathsf{null}} \to \mathrm{dom}(\mathcal{D})$ such that for all $u \in \Delta^{\mathcal{J}}_{\mathsf{null}}$ and $\varphi \in Z^{\mathcal{J}}(u)$ we have $\mathcal{D} \models \varphi(f(u))$. The *completion* $f(\mathcal{J})$ of $\mathcal{J}$ by f is the interpretation $\mathcal{I}$ obtained from $\mathcal{J}$ by replacing each data null u in $\mathcal{J}$ by $f(u)$, and by dropping the sets $Z^{\mathcal{J}}(u)$.

Using a straightforward modification of the standard chase procedure for *DL-Lite*$_{\mathcal{R}}$ (see, e.g., (Calvanese et al. 2007)) one can construct a pre-model $\mathrm{can}(\mathcal{T}, \mathcal{A})$ of any satisfiable $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$ such that for any UCQ q over $\mathcal{D}$ and

any $\bar{c}$:

$$(\mathcal{T},\mathcal{A}) \models q(\bar{c}) \iff f(\text{can}(\mathcal{T},\mathcal{A})) \models q(\bar{c}),$$
for all completion functions f.

We call $\text{can}(\mathcal{T},\mathcal{A})$ with this property a *universal pre-model* of $\mathcal{T}$ and $\mathcal{A}$.

Lemma 3.2. *For every satisfable $\mathcal{D}$-KB $(\mathcal{T},\mathcal{A})$ there exists a universal pre-model can$(\mathcal{T},\mathcal{A})$ of $\mathcal{T}$ and $\mathcal{A}$.*

The following example illustrates the construction of universal pre-models.

Example 3.3. Extend the TBox $\mathcal{T}$ from Example 3.1 to $\mathcal{T}'$ by adding $A \sqsubseteq \forall U_1.x \geq 0$ and $A \sqsubseteq \forall U_1.x \leq 1$. The universal pre-model $\text{can}(\mathcal{T}',\mathcal{A})$ is given by setting $\Delta^{\text{can}(\mathcal{T},\mathcal{A})} = \{a,u_1,u_2\}$; $A^{\text{can}(\mathcal{T},\mathcal{A})} = \{a\}$; $U_1^{\text{can}(\mathcal{T},\mathcal{A})} = \{(a,u_1)\}$; $U_2^{\text{can}(\mathcal{T},\mathcal{A})} = \{(a,u_2)\}$; $Z^{\text{can}(\mathcal{T},\mathcal{A})}(u_1) = \{0 \leq x,\, x \leq 1\}$, and $Z^{\text{can}(\mathcal{T},\mathcal{A})}(u_2) = \emptyset$. A completion function f for $\text{can}(\mathcal{T},\mathcal{A})$ maps u_1 to a rational number $f(u_1)$ with $0 \leq f(u_1) \leq 1$ and u_2 to some rational number and defines a completion $f(\text{can}(\mathcal{T},\mathcal{A}))$ in which U_1 is interpreted as $(a, f(u_1))$ and U_2 is interpreted as $(a, f(u_2))$.

4 Query Evaluation and CSP

Universal pre-models can be infinite and their completions can have a very complicated structure. In fact, we begin this section with the observation that there are many important and simple datatypes $\mathcal{D}$ such that evaluating OMQs over $\mathcal{D}$ is undecidable, and this holds even for OMQs with TBoxes without qualified attribute restrictions. We then introduce the bounded match depth property (BMDP) of OMQs which entails that evaluating OMQs can only be undecidable if a CSP over $\mathcal{D}$ is undecidable. Even more importantly, the BMDP allows us to establish mutual polynomial reductions between evaluating OMQs over a datatype $\mathcal{D}$ and CSPs over $\mathcal{D}$. We also establish the BMDP for a large class of OMQs.

Theorem 4.1. *Let $\mathcal{D} \in \{(\mathbb{Z},\neq),(\mathbb{Z},<),(\mathbb{Z},\leq),(\mathbb{Q},\neq),(\mathbb{Q},<)\}$. Then the query evaluation problem for OMQs over $\mathcal{D}$ with DL-Lite$_{\mathcal{R}}^{attrib}(\mathcal{D})$ TBoxes is undecidable in combined complexity.*

For $(\mathbb{Z},\neq)$ and $(\mathbb{Q},\neq)$ the proof is by reduction of the $\mathbb{N}\times\mathbb{N}$ tiling problem and very similar to known undecidability proofs for query evaluation with either UCQs with inequality or TBoxes with key constraints (Toman and Weddell 2008; Rosati 2007; Gutiérrez-Basulto et al. 2015). As $x \neq y$ iff $x < y$ or $y < x$, undecidability of query evaluation with inequality in the datatype entails undecidability of query evaluation with $<$ in the datatype. Finally, undecidability of query evaluation for $(\mathbb{Z},\leq)$ can be proved by reduction of query evaluation for $(\mathbb{Z},<)$. It remains open whether query evaluation is decidable for OMQs over $(\mathbb{Q},\leq)$ and whether undecidability holds for data complexity.

The proof of Theorem 4.1 makes heavy use of TBoxes whose chase does not terminate and of UCQs that are not *safe* or *rooted*. Given a CQ q over $\mathcal{D}$, we call any two variables x,y *connected* in q if x and y are connected via a path of atoms in q that does not contain data variables. Then q is *rooted* if any non-data variable is connected to a non-data answer variable and q is *safe* if any two non-data variables of q are either connected in q or are both connected to non-data answer variables. A UCQ is rooted (safe) if all its disjuncts are rooted (and safe, respectively) and an OMQ is rooted (safe) if its UCQ is rooted (and safe, respectively). Thus, every rooted query is safe, but the converse does not hold. Note that safe queries do not allow us to compare attribute values of individuals that are arbitrarily far apart in $\text{can}(\mathcal{T},\mathcal{A})$.

Example 4.2. The CQ in Example 1.1 is safe but not rooted, the CQs in Example 3.1 are rooted. In contrast, the Boolean CQ given by $q \leftarrow U_1(x_1,u_1) \wedge U_2(x_2,u_2) \wedge u_1 \leq u_2$ is not safe.

An important shared property of all OMQs with TBoxes with a terminating chase, all rooted OMQs, and all safe OMQs over homogeneous datatypes[1] is that they can be answered on a finite initial portion of $\text{can}(\mathcal{T},\mathcal{A})$ whose size only depends on the OMQ. Given an integer $d \geq 0$, let $\text{can}^d(\mathcal{T},\mathcal{A})$ be the subinterpretation of $\text{can}(\mathcal{T},\mathcal{A})$ induced by the set of domain elements that are reachable from ABox elements in at most d steps. An OMQ $Q = (\mathcal{T},q)$ over datatype $\mathcal{D}$ enjoys the *bounded match depth property (BMDP)* if there exists a $d \geq 0$ such that for all $\mathcal{D}$-ABoxes $\mathcal{A}$ and all tuples $\bar{c}$,

$$\mathcal{T},\mathcal{A} \models q(\bar{c}) \iff f(\text{can}^d(\mathcal{T},\mathcal{A})) \models q(\bar{c}), \qquad (\star)$$
for all completion functions f.

This property closely resembles the bounded derivation depth property in (Calì, Gottlob, and Lukasiewicz 2012), with the key difference being that it crucially depends on the notion of completion.

It is not difficult to verify that all OMQs with a TBox with a terminating chase and all rooted OMQs $(\mathcal{T},q)$ have the BMDP. In fact, in the latter case $(\star)$ holds when d is the maximum number of atoms in a disjunct of q. A much more sophisticated argument shows that safe OMQs over homogeneous datatypes enjoy the BMDP.

Lemma 4.3. *The following OMQs enjoy the BMDP: safe OMQs over homogeneous datatypes; rooted OMQs; and OMQs using a TBox $\mathcal{T}$ such that $\text{can}(\mathcal{T},\mathcal{A})$ is finite for each ABox $\mathcal{A}$.*

We now show that every OMQ $Q = (\mathcal{T},q)$ over $\mathcal{D}$ that enjoys the BMDP can be reduced to a constraint satisfaction problem $\mathsf{CSP}_c(\Gamma)$ whose constraint language Γ is defined solely by the patterns of formulas over $\mathcal{D}$ that occur in Q. We describe these patterns using the notion of the *datatype pattern* of Q. The *datatype pattern* of Q is defined as $\mathsf{dtype}(Q) = (\theta_{\mathcal{T}},\theta_q)$, where $\theta_{\mathcal{T}}$ is the set of all PP formulas that occur in qualified attribute restrictions of $\mathcal{T}$, and θ_q contains, for each disjunct q' of q, the conjunction of all atoms of q' over $\mathcal{D}$. Note that $\theta_{\mathcal{T}}$ is a set of PP formulas with constants and a single free variable, whereas θ_q is a set of PP formulas without constants. We refer to datatype patterns of OMQs over $\mathcal{D}$ as *datatype patterns over $\mathcal{D}$*.

[1] A datatype $\mathcal{D}$ is homogeneous if every isomorphism between finite induced substructures extends to an automorphism on $\mathcal{D}$. $(\mathbb{Q},<)$ and $(\mathbb{Q},\leq)$ are examples of homogeneous datatypes (Chang and Keisler 1998).

Example 4.4. Let $\mathcal{T} = \{A \sqsubseteq \exists U_1.1 \leq x \wedge x \leq 3\}$ be a $\mathcal{D}$-TBox, for $\mathcal{D} = (\mathbb{Q}, \leq)$, and let q be the UCQ over $\mathcal{D}$ given by $q_1(x), q_2(x)$, where

$$q_1(x) \leftarrow \textstyle\bigwedge_{i=1}^{3} U_i(x, z_i) \wedge z_1 \leq z_2 \wedge z_1 \leq z_3,$$
$$q_2(x) \leftarrow U_1(x, z_1') \wedge U_2(x, z_2') \wedge z_2' \leq z_1'.$$

Then $\mathsf{dtype}(\mathcal{T}, q) = (\theta_\mathcal{T}, \theta_q)$, where $\theta_\mathcal{T} = \{1 \leq x \wedge x \leq 3\}$ and $\theta_q = \{z_1 \leq z_2 \wedge z_1 \leq z_3, z_2' \leq z_1'\}$.

Now, with each datatype pattern $\theta = (\theta_\mathcal{T}, \theta_q)$ over $\mathcal{D}$, where $\theta_\mathcal{T} = \{\varphi_1, \ldots, \varphi_m\}$ and $\theta_q = \{\psi_1, \ldots, \psi_n\}$, we associate the constraint language

$$\Gamma_\theta = (\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \ldots, R_{\varphi_m}, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n}),$$

where for each formula $\varphi(x_1, \ldots, x_k)$ over $\mathcal{D}$, we let $R_\varphi = \{\bar{a} \in \mathrm{dom}(\mathcal{D})^k \mid \mathcal{D} \models \varphi(\bar{a})\}$.

Theorem 4.5. *Let θ be a datatype pattern over $\mathcal{D}$.*

If $Q = (\mathcal{T}, q)$ is an OMQ over $\mathcal{D}$ with $\mathsf{dtype}(Q) = \theta$ and Q enjoys the BMDP, then evaluating Q is polynomially reducible to the complement of $\mathsf{CSP}_c(\Gamma_\theta)$.

Conversely, there is a rooted OMQ Q over $\mathcal{D}$ with $\mathsf{dtype}(Q) = \theta$ such that the complement of $\mathsf{CSP}_c(\Gamma_\theta)$ is polynomially reducible to evaluating Q.

Proof. Let $\theta = (\theta_\mathcal{T}, \theta_q)$, where $\theta_\mathcal{T} = \{\varphi_1, \ldots, \varphi_m\}$ and $\theta_q = \{\psi_1, \ldots, \psi_n\}$. Assume Q enjoys the BMDP, and that q is given as $q_1(\bar{x}), \ldots, q_n(\bar{x})$. Let $\mathcal{A}$ be a $\mathcal{D}$-ABox and let $\bar{c}$ be a tuple of individual names and data values of the same length as $\bar{x}$. It is not difficult to see that satisfiability of $\mathcal{D}$-ABoxes $\mathcal{A}$ relative to $\mathcal{T}$ is polynomially reducible to $\mathsf{CSP}_c(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \ldots, R_{\varphi_m})$. Thus, we can assume that $\mathcal{A}$ is satisfiable relative to $\mathcal{T}$. Consider the pre-model $\mathcal{I} := \mathrm{can}^d(\mathcal{T}, \mathcal{A})$, where d is an integer that satisfies $(\star)$ but is independent of $\mathcal{A}$. A *pre-match* of q_i in $\mathcal{I}$ is a match of the abstract part of q_i (i.e. q_i stripped off of all atoms over $\mathcal{D}$) in $\mathcal{I}$. Let X_i be the set of all pre-matches μ of q_i in $\mathcal{I}$ with $\mu(\bar{x}) = \bar{c}$, and let $\bar{u} = u_1, \ldots, u_k$ be a repetition-free enumeration of all the data nulls that occur in the image of some pre-match in $X_1 \cup \cdots \cup X_n$. Then the following is an instance of $\mathsf{CSP}_c(\Gamma_\theta)$:

$$\Phi := \exists \bar{u} \left(\bigwedge_{i=1}^{k} \bigwedge_{\varphi \in Z^\mathcal{I}(u_i)} R_\varphi(u_i) \wedge \bigwedge_{i=1}^{n} \bigwedge_{\mu \in X_i} R_{\neg\psi_i}(\mu(\bar{z}_i)) \right),$$

where $u_1, \ldots, u_k$ are identified with individual variables in Φ. It is easy to check that $\mathcal{T}, \mathcal{A} \not\models q(\bar{c})$ iff $\Gamma_\theta \models \Phi$.

For the converse, we encode each instance Φ of $\mathsf{CSP}_c(\Gamma_\theta)$ by an ABox $\mathcal{A}_\Phi$ as follows. We use a distinguished individual a_Φ to denote the root of $\mathcal{A}_\Phi$. For each atom $\alpha = R_{\neg\psi_i}(x_1, \ldots, x_k)$ in Φ there is an individual b_α connected to a_Φ via an assertion $r_i(a_\Phi, b_\alpha)$. For each $j \in \{1, \ldots, k\}$, this individual is connected to individual c_{x_j} via an assertion $s_j(b_\alpha, c_{x_j})$. Finally, for each variable x that occurs in Φ we include the assertion $A(c_x)$; if $R_{\varphi_i}(x)$ occurs in Φ we additionally include $A_{\varphi_i}(c_x)$. Furthermore, for each element $u \in \mathrm{dom}(\mathcal{D})$ that occurs in Φ we include the assertion $U(c_u, u)$. Figure 1 illustrates this construction for the case $m = 1$, $n = 3$, and Φ being

$$\exists x, y \, (R_{\varphi_1}(x) \wedge R_{\neg\psi_1}(0, x) \wedge R_{\neg\psi_1}(x, y) \wedge R_{\neg\psi_3}(y, 1, x)) \, .$$

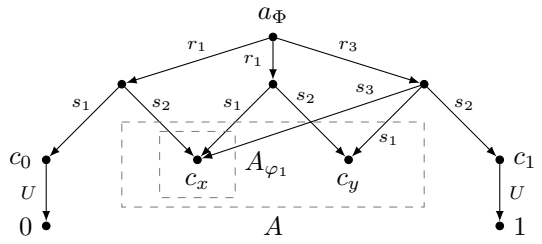

Figure 1: ABox $\mathcal{A}_\Phi$ from the proof of Theorem 4.5

Let $\mathcal{T} = \{A \sqsubseteq \exists U\} \cup \{A_{\varphi_i} \sqsubseteq \forall U.\varphi_i \mid 1 \leq i \leq m\}$ and let q be the UCQ given by $q_1(x), \ldots, q_n(x)$, where $q_i(x)$ is

$$r_i(x, y) \wedge \bigwedge_{1 \leq j \leq k} (s_j(y, z_j) \wedge U(z_j, u_j)) \wedge \psi_i(u_1, \ldots, u_k)$$

and k is the number of free variables of ψ_i. Clearly, $Q = (\mathcal{T}, q)$ is a rooted OMQ over $\mathcal{D}$ with $\mathsf{dtype}(Q) = \theta$. It is straightforward to verify that $\Gamma_\theta \models \Phi$ if and only if $\mathcal{T}, \mathcal{A}_\Phi \not\models q(a_\Phi)$. □

Note that the TBoxes constructed in the converse direction of the proof of Theorem 4.5 are very simple. In particular, the chase terminates on any given ABox.

It follows immediately from Theorem 4.5 that for the datatypes $\mathcal{D}$ given in Theorem 4.1 the query evaluation problem for OMQs with the BMDP over $\mathcal{D}$ is in coNP. In the next section we show how Theorem 4.5 can be used to understand the OMQs with the BMDP whose query evaluation problem is not only in coNP but in PTime.

5 The Datatype $(\mathbb{Q}, \leq)$

The results of the previous section establish a tight link between evaluating OMQs and solving CSPs. In particular, they allow us to transfer complexity classification results from CSP to OMQ answering. We now demonstrate the power of this link for the datatype $(\mathbb{Q}, \leq)$. Our main result is the following P/coNP-dichotomy of evaluating OMQs over $(\mathbb{Q}, \leq)$ that enjoy the BMDP based on their datatype patterns. To simplify the presentation, we assume w.l.o.g. that for all datatype patterns $\theta = (\theta_\mathcal{T}, \theta_q)$ over $(\mathbb{Q}, \leq)$ the formulas in θ_q are acyclic.[2]

We use *min-pattern* and *max-pattern* to refer to formulas of the form $x_0 \leq x_1 \wedge \cdots \wedge x_0 \leq x_k$ and $x_1 \leq x_0 \wedge \cdots \wedge x_k \leq x_0$, for $k \geq 0$, respectively.

Theorem 5.1. *Let $\theta = (\theta_\mathcal{T}, \theta_q)$ be a datatype pattern over $(\mathbb{Q}, \leq)$, where $\theta_q = \{\psi_1, \ldots, \psi_n\}$.*

If each ψ_i is a min-pattern or each ψ_i is a max-pattern, then evaluating OMQs Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ and the BMDP is in PTime.

Otherwise, there is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ such that evaluating Q is coNP-complete.

[2] Cycles $x_1 \leq x_2 \leq \cdots \leq x_n \leq x_1$ that occur in a formula in θ_q can always be eliminated by removing all their atoms and replacing each x_i with x_1.

The simple and purely syntactic characterization of the tractable cases of Theorem 5.1 makes it very easy to verify whether evaluating a given OMQ over $(\mathbb{Q}, \leq)$ with the BMDP is guaranteed to be tractable or possibly coNP-complete. For instance, Theorem 5.1 implies that the OMQ $(\mathcal{T}, q)$ in Example 4.4 and in general all OMQs over $(\mathbb{Q}, \leq)$ that have the same datatype pattern and enjoy the BMDP can be evaluated in PTime. On the other hand, if we consider the datatype pattern that has $z_1 \leq z_2 \wedge z_2 \leq z_3$ in place of $z_1 \leq z_2 \wedge z_1 \leq z_3$, then there are rooted OMQs over $(\mathbb{Q}, \leq)$ with that datatype pattern for which evaluation is coNP-complete.

We now take a closer look at the proof of Theorem 5.1. The main device is a recent dichotomy for *temporal CSPs* by Bodirsky and Kára (2010a). Temporal CSPs are defined as $\mathsf{CSP}(\Gamma)$, where $\Gamma = (\mathbb{Q}, R_1, R_2, \dots)$ and each R_i is definable by a first-order formula $\Phi_i(\bar{x})$ over $(\mathbb{Q}, <)$ without constants, i.e., $R_i = \{\bar{a} \mid (\mathbb{Q}, <) \models \Phi_i(\bar{a})\}$. The datatypes Γ of a temporal CSP are called *temporal constraint languages*. Bodirsky and Kára (2010a) prove that for every temporal constraint language Γ, $\mathsf{CSP}(\Gamma)$ is either in PTime or NP-complete. Whether or not a given temporal constraint language defines a tractable CSP depends only on which functions preserve its relations. Bodirsky and Kára consider a set $\mathcal{F}$ of five types of functions $f\colon \mathbb{Q}^2 \to \mathbb{Q}$ together with the set *dual-*$\mathcal{F}$ consisting of the duals *dual-*$f(x, y) := -f(-x, -y)$ of all functions f in $\mathcal{F}$.[3] A function $f \in \mathcal{F} \cup$ *dual-*$\mathcal{F}$ *preserves* a relation $R \subseteq \mathbb{Q}^n$ if for all $(a_1, \dots, a_n), (b_1, \dots, b_n) \in R$ we have $(f(a_1, b_1), \dots, f(a_n, b_n)) \in R$. We also say that f preserves a temporal constraint language Γ if f preserves all relations in Γ.

Theorem 5.2. (Bodirsky and Kára 2010a) *Let Γ be a temporal constraint language. If Γ is preserved by a function in $\mathcal{F} \cup$ dual-$\mathcal{F}$, then $\mathsf{CSP}(\Gamma)$ is in PTime. Otherwise, $\mathsf{CSP}(\Gamma)$ is NP-complete.*

Together with Theorem 4.5 and a constant-elimination technique, this leads to a basic P/coNP-dichotomy for evaluating OMQs Q over $(\mathbb{Q}, \leq)$ with the BMDP based on the preservation properties of $\Gamma_{\mathsf{dtype}(Q)}$.

Theorem 5.3. *Let $\theta = (\theta_{\mathcal{T}}, \theta_q)$ be a datatype pattern over $(\mathbb{Q}, \leq)$, where $\theta_q = \{\psi_1, \dots, \psi_n\}$.*

If there is a function in $\mathcal{F} \cup$dual-$\mathcal{F}$ that preserves $R_{\neg\psi_i}$ for each $i \in \{1, \dots, n\}$, then evaluating OMQs Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ and the BMDP is in PTime.

Otherwise, there is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ such that evaluating Q is coNP-complete.

To obtain a purely syntactic characterization of the datatype patterns $\theta = (\theta_{\mathcal{T}}, \theta_q)$ over $(\mathbb{Q}, \leq)$ that lead to tractable OMQs, we further analyze the relations $R_{\neg\psi}$ that are defined by the formulas $\psi \in \theta_q$ and are preserved under some function in $\mathcal{F} \cup$ *dual-*$\mathcal{F}$. Note that the negation $\neg\psi$ of each formula $\psi \in \theta_q$ is equivalent to a disjunction Ψ of atomic formulas $x < y$, with x and y variables. Moreover, since ψ is acyclic (by assumption), the directed graph with the variables of Ψ as its vertices, and edges (y, x) for each atomic formula $x < y$ of Ψ is acyclic. We call such formulas Ψ *acyclic disjunctive* formulas. The following lemma is the combinatorial core of our analysis.

Lemma 5.4. *Let $R \subseteq \mathbb{Q}^n$ be defined by an acyclic disjunctive formula Ψ over $(\mathbb{Q}, <)$. If R is preserved by a function in $\mathcal{F} \cup$ dual-$\mathcal{F}$, then Ψ has the form $\bigvee_{i=1}^{k} x_i < x_0$ if $f \in \mathcal{F}$, and $\bigvee_{i=1}^{k} x_0 < x_i$ if $f \in$ dual-$\mathcal{F}$.*

Altogether, this leads to a proof of Theorem 5.1.

Proof of Theorem 5.1. By Theorem 5.3, it suffices to show that the following are equivalent:

1. Each $\psi \in \theta_q$ is a min-pattern, or each $\psi \in \theta_q$ is a max-pattern.
2. There is a function $f \in \mathcal{F} \cup$ *dual-*$\mathcal{F}$ such that each $R_{\neg\psi}$, $\psi \in \theta_q$, is preserved under f.

If each $\psi \in \theta_q$ is a min-pattern, then for each $\psi \in \theta_q$ the relation $R_{\neg\psi}$ is defined by a formula of the form $\bigvee_{i=1}^{n} x_0 > x_i$. Similarly, if each $\psi \in \theta_q$ is a max-pattern, then for each $\psi \in \theta_q$ the relation $R_{\neg\psi}$ is defined by a formula of the form $\bigvee_{i=1}^{n} x_i > x_0$. It is known (Bodirsky and Kára 2010b, Proposition 3.5) that relations defined by such formulas are preserved under a function in $\mathcal{F} \cup$ *dual-*$\mathcal{F}$.

Conversely, let f be a function in $\mathcal{F} \cup$ *dual-*$\mathcal{F}$ such that each $R_{\neg\psi}$ with $\psi \in \theta_q$ is preserved under f. Let $\psi \in \theta_q$. Then, $R_{\neg\psi}$ is defined by an acyclic disjunctive formula Ψ. By Lemma 5.4, Ψ has the form $\bigvee_{i=1}^{n} x_i < x_0$, if $f \in \mathcal{F}$, and $\bigvee_{i=1}^{n} x_0 < x_i$, if $f \in$ *dual-*$\mathcal{F}$. This implies that each $\psi \in \theta_q$ is a min-pattern, or each $\psi \in \theta_q$ is a max-pattern. □

We now refine the analysis of the datatype patterns that lead to OMQs with an evaluation problem in PTime further by presenting a dichotomy between those that can be used in PTime-hard OMQs and those that always lead to OMQs in NLogSpace. Example 1.1 shows that there are rooted OMQs Q over $(\mathbb{Q}, \leq)$ and with $\mathsf{dtype}(Q) = (\emptyset, \{x \leq y\})$ such that evaluating Q is NLogSpace-complete. It turns out that the NLogSpace upper bound holds for all OMQs Q whose datatype pattern contains atomic formulas only.

Theorem 5.5. *Evaluating OMQs Q over $(\mathbb{Q}, \leq)$ with the BMDP and $\mathsf{dtype}(Q) = (\theta_{\mathcal{T}}, \theta_q)$ such that each formula in θ_q is of the form $x_0 \leq x_1$ is in NLogSpace.*

The proof of Theorem 5.5 is a straightforward application of Part 1 of Theorem 4.5, which allows us to reduce evaluating OMQs as in Theorem 5.5 to the complement of $\mathsf{CSP}_c(\mathbb{Q}, <, \leq)$ (it is not difficult to see that this reduction can be carried out in logarithmic space), and the observation that $\mathsf{CSP}_c(\mathbb{Q}, <, \leq)$ is in NLogSpace via a simple reachability test. The following result entails that the NLogSpace upper bound cannot be generalised to further tractable datatype patterns.

Theorem 5.6. *There is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = (\emptyset, \{x \leq y \wedge x \leq z\})$ such that evaluating Q is PTime-complete.*

The proof of Theorem 5.6 is by reduction of the alternating reachability problem.

[3] The set $\mathcal{F}$ consists of the functions min, mi, mx, and ll, defined in (Bodirsky and Kára 2010a), and all constant functions.

6 Conclusion

We have presented a framework for analyzing the non-uniform complexity of OMQ answering with expressive datatypes by establishing a close link to CSPs. We have illustrated the power of this framework by transfering a P/coNP dichotomy result for CSPs over $(\mathbb{Q}, \leq)$ to OMQ answering over $(\mathbb{Q}, \leq)$. Many research questions arise within this framework, including the following: (1) The framework should be applied to analyze the complexity of OMQ answering for other important datatypes based on the rationals, the integers, strings, or other structures used in spatial and temporal reasoning. On the CSP side there has been very significant progress in understanding the complexity of such structures (Bodirsky 2015). (2) We have established the BMDP for many relevant OMQs, but a more systematic investigation covering additional datatypes and TBoxes would be useful. (3) We have focused on establishing worst-case complexity results and dichotomies for OMQ answering. It would be of great interest to exploit the CSP reduction further and develop practical query answering algorithms, in particular, to use constraint solvers as part of practical query engines.

Acknowledgments. This work was supported by the EPSRC under the grant EP/M012646/1 "iTract: Islands of Tractability in Ontology-Based Data Access".

References

Artale, A.; Calvanese, D.; Kontchakov, R.; and Zakharyaschev, M. 2009. The *DL-Lite* family and relations. *J. Artif. Intell. Res. (JAIR)* 36:1–69.

Artale, A.; Ryzhikov, V.; and Kontchakov, R. 2012. *DL-Lite* with attributes and datatypes. In *ECAI*, 61–66.

Baader, F., and Hanschke, P. 1991. A scheme for integrating concrete domains into concept languages. In *IJCAI 1991*, 452–457.

Baader, F.; Brandt, S.; and Lutz, C. 2005. Pushing the $\mathcal{EL}$ envelope. In *IJCAI-05*, 364–369.

Bienvenu, M., and Ortiz, M. 2015. Ontology-mediated query answering with data-tractable description logics. In *Reasoning Web 2015*, 218–307.

Bienvenu, M.; ten Cate, B.; Lutz, C.; and Wolter, F. 2014. Ontology-based data access: A study through disjunctive datalog, csp, and MMSNP. *ACM Trans. Database Syst.* 39(4):33:1–33:44.

Bodirsky, M., and Kára, J. 2010a. The complexity of temporal constraint satisfaction problems. *J. ACM* 57(2):9:1–9:41.

Bodirsky, M., and Kára, J. 2010b. A fast algorithm and datalog inexpressibility for temporal reasoning. *ACM Trans. Comput. Log.* 11(3):15:1–15:21.

Bodirsky, M. 2015. The complexity of constraint satisfaction problems (invited talk). In *STACS 2015*, 2–9.

Bulatov, A. A.; Jeavons, P.; and Krokhin, A. A. 2005. Classifying the complexity of constraints using finite algebras. *SIAM J. Comput.* 34(3):720–742.

Calì, A.; Gottlob, G.; and Lukasiewicz, T. 2012. A general datalog-based framework for tractable query answering over ontologies. *J. Web Sem.* 14:57–83.

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. Autom. Reasoning* 39(3):385–429.

Chang, C., and Keisler, H. 1998. *Model Theory*. Elsevier.

Feder, T., and Vardi, M. Y. 1993. Monotone monadic SNP and constraint satisfaction. In *Proc. of the ACM Symposium on Theory of Computing*, 612–622.

Grau, B. C.; Horrocks, I.; Krötzsch, M.; Kupke, C.; Magka, D.; Motik, B.; and Wang, Z. 2013. Acyclicity notions for existential rules and their application to query answering in ontologies. *J. Artif. Intell. Res. (JAIR)* 47:741–808.

Gutiérrez-Basulto, V.; Ibáñez-García, Y. A.; Kontchakov, R.; and Kostylev, E. V. 2015. Queries with negation and inequalities over lightweight ontologies. *J. Web Sem.* 35:184–202.

Hernich, A.; Lutz, C.; Ozaki, A.; and Wolter, F. 2015. Schema.org as a description logic. In *IJCAI 2015*, 3048–3054.

Immerman, N. 1999. *Descriptive complexity*. Graduate texts in computer science. Springer.

Lutz, C., and Wolter, F. 2012. Non-uniform data complexity of query answering in description logics. In *KR 2012*.

Lutz, C. 2002. Description logics with concrete domains-a survey. In *Advances in Modal Logic 4*, 265–296.

Magka, D.; Kazakov, Y.; and Horrocks, I. 2011. Tractable extensions of the description logic $\mathcal{EL}$ with numerical datatypes. *J. Autom. Reasoning* 47(4):427–450.

Motik, B., and Horrocks, I. 2008. OWL datatypes: Design and implementation. In *ISWC 2008*, 307–322.

Motik, B.; Grau, B. C.; Horrocks, I.; Wu, Z.; Fokoue, A.; and Lutz, C. 2009. Owl 2 web ontology language: Profiles. World Wide Web Consortium, Working Draft WD-owl2-profiles-20081202.

Poggi, A.; Lembo, D.; Calvanese, D.; De Giacomo, G.; Lenzerini, M.; and Rosati, R. 2008. Linking data to ontologies. *J. Data Semantics* 10:133–173.

Rosati, R. 2007. The limits of querying ontologies. In Schwentick, T., and Suciu, D., eds., *Database Theory - ICDT 2007*, volume 4353 of *Lecture Notes in Computer Science*, 164–178. Springer.

Savkovic, O., and Calvanese, D. 2012. Introducing datatypes in *DL-Lite*. In *ECAI 2012*, 720–725.

Toman, D., and Weddell, G. E. 2008. On keys and functional dependencies as first-class citizens in description logics. *J. Autom. Reasoning* 40(2-3):117–132.

Tutte, W. T. 1982. The method of alternating paths. *Combinatorica* 2(3):325–332.

A Formal Definitions and Proofs for Section 3

We start by giving formal definitions of pre-interpretations and pre-models of a KB. We then present the chase procedure to construct a universal pre-model for any satisfiable $\mathcal{D}$-KB. It also checks whether the KB is satisfiable.

Definition A.1. A *pre-interpretation $\mathcal{I}$ over a datatype $\mathcal{D}$* is a triple $(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}}, Z^{\mathcal{I}})$ such that

- $\Delta^{\mathcal{I}} = \Delta^{\mathcal{I}}_{\mathsf{ind}} \cup \Delta^{\mathcal{I}}_{\mathsf{data}}$ is the domain of $\mathcal{I}$, where $\Delta^{\mathcal{I}}_{\mathsf{ind}}$ is a non-empty set of individuals, $\Delta^{\mathcal{I}}_{\mathsf{data}} = \mathrm{dom}(\mathcal{D}) \cup \Delta^{\mathcal{I}}_{\mathsf{null}}$ is a set of *data elements*, and $\Delta^{\mathcal{I}}_{\mathsf{null}}$ is a set of *data nulls*;
- $\cdot^{\mathcal{I}}$ assigns to each concept name A a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$, to each role name P a relation $P^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}_{\mathsf{ind}} \times \Delta^{\mathcal{I}}_{\mathsf{ind}}$, and to each attribute name U a relation $U^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}_{\mathsf{ind}} \times \Delta^{\mathcal{I}}_{\mathsf{data}}$;
- $Z^{\mathcal{I}}$ maps each $u \in \Delta^{\mathcal{I}}_{\mathsf{data}}$ to a set $Z^{\mathcal{I}}(u)$ of PP formulas $\varphi(x)$ with constants over $\mathcal{D}$ such that (i) $\mathcal{D} \models \exists x(\bigwedge_{\varphi(x) \in Z^{\mathcal{I}}(u)} \varphi(x))$ holds for every $u \in \Delta^{\mathcal{I}}_{\mathsf{data}}$ and (ii) if $d \in D$, then $(x = d) \in Z^{\mathcal{I}}(d)$.

We extend the definition of the extension $C^{\mathcal{I}}$ of a concept C from interpretations to pre-interpretations by defining the extension of the concepts C not involving attributes in the straightforward way and by setting for any pre-interpretation $\mathcal{I}$ for $\mathcal{D}$

- $(\exists U)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}}_{\mathsf{ind}} \mid \text{there exists } v \in \Delta^{\mathcal{I}}_{\mathsf{data}} \text{ with } (a, v) \in U^{\mathcal{I}}\}$;
- $(\exists U.\varphi)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}}_{\mathsf{ind}} \mid \text{there exists } v \in \Delta^{\mathcal{I}}_{\mathsf{data}} \text{ with } (a, v) \in U^{\mathcal{I}} \text{ and } \varphi \in Z^{\mathcal{I}}(v)\}$;
- $(\forall U.\varphi)^{\mathcal{I}} = \{a \in \Delta^{\mathcal{I}}_{\mathsf{ind}} \mid \text{for all } v \in \Delta^{\mathcal{I}}_{\mathsf{data}} \text{ with } (a, v) \in U^{\mathcal{I}} \text{ we have } \varphi \in Z^{\mathcal{I}}(v)\}$;

For any inclusion $C \sqsubseteq Y$ we set $\mathcal{I} \models X \sqsubseteq Y$ if $X^{\mathcal{I}} \subseteq Y^{\mathcal{I}}$. In this case we say that $\mathcal{I}$ satisfies $X \sqsubseteq Y$. $\mathcal{I}$ is a *pre-model* of a TBox if it satisfies all its inclusions. $\mathcal{I}$ is a *pre-model* of an ABox $\mathcal{A}$ if $a \in A^{\mathcal{I}}$ for all $A(a) \in \mathcal{A}$, $(a, b) \in P^{\mathcal{I}}$ for all $P(a, b) \in \mathcal{A}$, and $(a, d) \in U^{\mathcal{I}}$ for all $U(a, d) \in \mathcal{A}$. $\mathcal{I}$ is a *pre-model* of a KB $(\mathcal{T}, \mathcal{A})$ if it is a pre-model of $\mathcal{T}$ and $\mathcal{A}$.

A completion of a pre-interpretation replaces data nulls by suitable values from $\mathrm{dom}(\mathcal{D})$. Let $\mathcal{J}$ be a pre-interpretation over $\mathcal{D}$. Then

1. a *completion function* f for $\mathcal{J}$ is a mapping $f \colon \Delta^{\mathcal{J}}_{\mathsf{null}} \to D$ such that $\mathcal{D} \models \varphi(f(u))$ for all $u \in \Delta^{\mathcal{J}}_{\mathsf{null}}$ and $\varphi(x) \in Z^{\mathcal{J}}(u)$.
2. The *completion* of $\mathcal{J}$ given by f, written $f(\mathcal{J})$, is the interpretation $\mathcal{I}$ defined by setting:
   - $\Delta^{\mathcal{I}}_{\mathsf{ind}} = \Delta^{\mathcal{J}}_{\mathsf{ind}}$;
   - $A^{\mathcal{I}} = A^{\mathcal{J}}$ for all $A^{\mathcal{J}} \subseteq \Delta^{\mathcal{J}}_{\mathsf{ind}}$;
   - $r^{\mathcal{I}} = r^{\mathcal{J}}$ for all $r^{\mathcal{J}} \subseteq \Delta^{\mathcal{J}}_{\mathsf{ind}} \times \Delta^{\mathcal{J}}_{\mathsf{ind}}$;
   - $U^{\mathcal{I}} = \big(U^{\mathcal{J}} \cap (\Delta^{\mathcal{J}}_{\mathsf{ind}} \times \mathrm{dom}(\mathcal{D}))\big) \cup \{(b, f(v)) \mid (b, v) \in U^{\mathcal{J}}, v \in \Delta^{\mathcal{J}}_{\mathsf{null}}\}$.

The following lemma can be proved in a straightforward way.

Lemma A.2. *If $\mathcal{J}$ is a pre-model of a KB $(\mathcal{T}, \mathcal{A})$, then $f(\mathcal{J})$ is a model of $(\mathcal{T}, \mathcal{A})$.*

Definition A.3. Let $\mathcal{I}$ be a pre-model of a KB $(\mathcal{T}, \mathcal{A})$ such that for any UCQ q and tuple $\bar{c}$, $(\mathcal{T}, \mathcal{A}) \models q(\bar{c})$ if, and only if, $f(\mathcal{I}) \models q(\bar{c})$ for all completion functions f. Then $\mathcal{I}$ is called a *universal pre-model of* $(\mathcal{T}, \mathcal{A})$.

We now show that any pre-model of a KB that is initial regarding homomorphisms is a universal pre-model of the KB. To this end, we first define homomorphisms between a pre-interpretation for $\mathcal{D}$ and an interpretation for $\mathcal{D}$. Let $\mathcal{I}$ be a pre-interpretation and $\mathcal{J}$ be an interpretation for $\mathcal{D}$. A *homomorphism* from $\mathcal{I}$ to $\mathcal{J}$ is a mapping $h \colon \Delta^{\mathcal{I}} \to \Delta^{\mathcal{J}}$ such that all invididual names and all $d \in \mathrm{dom}(\mathcal{D})$ are mapped to themselves and

(a) $h[\Delta^{\mathcal{I}}_{\mathsf{ind}}] \subseteq \Delta^{\mathcal{J}}_{\mathsf{ind}}$ such that $h(a) \in A^{\mathcal{J}}$ if $a \in A^{\mathcal{I}}$, and $(h(a), h(a')) \in r^{\mathcal{J}}$ if $(a, a') \in r^{\mathcal{I}}$;

(b) $h[\Delta^{\mathcal{I}}_{\mathsf{data}}] \subseteq \mathrm{dom}(\mathcal{D})$ such that if $(a, v) \in U^{\mathcal{I}}$, then $(h(a), h(v)) \in U^{\mathcal{J}}$ and $\mathcal{D} \models \varphi(h(v))$ for all $\varphi(x) \in Z^{\mathcal{I}}(v)$.

Observe that if $\mathcal{I}$ is actually an interpretation (rather than a pre-interpretation only) then we obtain a homomorphism in the standard sense and it follows that for any UCQ q and tuple $\bar{c}$ of individual names and data values in $\mathcal{I}$ we have $\mathcal{J} \models q(\bar{c})$ whenever $\mathcal{I} \models q(\bar{c})$.

Definition A.4. A pre-model $\mathcal{I}$ of a KB $(\mathcal{T}, \mathcal{A})$ is hom-initial for $(\mathcal{T}, \mathcal{A})$ if for all models $\mathcal{I}'$ of $(\mathcal{T}, \mathcal{A})$, there is a homomorphism from $\mathcal{I}$ to $\mathcal{I}'$.

Theorem A.5. *Let $\mathcal{I}$ be hom-initial for a KB $(\mathcal{T}, \mathcal{A})$. Then $\mathcal{I}$ is a universal pre-model of $(\mathcal{T}, \mathcal{A})$.*

Proof. Let $\mathcal{I}$ be a hom-initial pre-model of a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$. We show that $\mathcal{I}$ satisfies Definition A.3, that is, for any UCQ q and tuple $\bar{c}$, $(\mathcal{T}, \mathcal{A}) \models q(\bar{c})$ if, and only if, $f(\mathcal{I}) \models q(\bar{c})$ for all completion functions f.

So let $q(\bar{x})$ be a UCQ and $\bar{c}$ be a tuple of individuals and data values from $\mathcal{A}$.

First, assume for all models $\mathcal{I}$ of $(\mathcal{T}, \mathcal{A})$ we have $\mathcal{I} \models q(\bar{c})$. Let f be a completion function for $\mathcal{I}$. By Lemma A.2, $f(\mathcal{I})$ is a model of $(\mathcal{T}, \mathcal{A})$. Thus $f(\mathcal{I}) \models q(\bar{c})$ as desired.

Conversely, assume $f(\mathcal{I}) \models q(\bar{c})$ for all completion functions f and let $\mathcal{J}$ be a model of $(\mathcal{T}, \mathcal{A})$. We have to show that $\mathcal{J} \models q(\bar{c})$. By assumption there is a homomorphism h_0 from $\mathcal{I}$ to $\mathcal{J}$.

Construct a completion function f_0 for $\mathcal{I}$ using h_0 by setting $f_0 := h_0 \restriction_{\Delta^{\mathcal{I}}_{\mathsf{data}}}$. f_0 is a completion function since $\mathcal{D} \models \varphi(h_0(u))$ for all $\varphi(x) \in Z^{\mathcal{I}}(u)$, by definition of homomorphisms.

Now the mapping $h : f_0(\mathcal{I}) \to \mathcal{J}$ defined by setting $h(a) = h_0(a)$ for all $a \in \Delta^{\mathcal{I}}_{\mathsf{ind}}$ and $h(d) = d$ for all $d \in \mathrm{dom}(\mathcal{D})$ is a homomorphism from $f_0(\mathcal{I})$ to $\mathcal{J}$. But then $\mathcal{J} \models q(\bar{c})$ since $f_0(\mathcal{I}) \models q(\bar{c})$. □

We now introduce the chase procedure that constructs for every satisfiable KB a hom-initial pre-model of the KB (and thus also a universal pre-model of the KB).

To present the chase procedure it is convenient to regard a pre-interpretation as a set of assertions that can contain, in

addition to ABox assertions, assertions of the form $U(a, u)$ with u a data null and sets $Z(u)$ of PP formulas for u a data value or data null. We call such sets of assertions *pre-ABoxes*. As every pre-ABox can be converted into a pre-interpretation we will often not distinguish between the two. For example, we write $\exists r(a) \in S$ for a pre-ABox S if $a \in (\exists r)^{\mathcal{I}}$ for the pre-interpretation $\mathcal{I}$ corresponding to S (which is the case if there exists b with $r(a, b) \in S$). To simplify presentation, we assume that for any ABox and pre-ABox S, $p(a, b) \in S$ iff $p^-(b, a) \in S$.

Assume now a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$ is given. The rules of the chase procedure are given below and are applied exhaustively to the pre-ABox S obtained from $\mathcal{A}$ by adding $Z(d) = \{x = d\}$ for every data value d in $\mathcal{A}$. If the procedure does not terminate with output 'unsatisfiable', then denote by $\mathrm{chase}(\mathcal{T}, \mathcal{A})$ the set of assertions obtained in the limit. We construct a pre-model of $(\mathcal{T}, \mathcal{A})$, denoted $\mathrm{can}(\mathcal{T}, \mathcal{A})$, as the pre-model $(\Delta^{\mathrm{can}(\mathcal{T},\mathcal{A})}, \cdot^{\mathrm{can}(\mathcal{T},\mathcal{A})}, Z^{\mathrm{can}(\mathcal{T},\mathcal{A})})$, where:

- $\Delta^{\mathrm{can}(\mathcal{T},\mathcal{A})}_{\mathsf{ind}}$ is the set of individuals in $\mathrm{chase}(\mathcal{T}, \mathcal{A})$;
- $A^{\mathrm{can}(\mathcal{T},\mathcal{A})} = \{a \mid A(a) \in \mathrm{chase}(\mathcal{T}, \mathcal{A})\}$;
- $p^{\mathrm{can}(\mathcal{T},\mathcal{A})} = \{(a, b) \mid p(a, b) \in \mathrm{chase}(\mathcal{T}, \mathcal{A})\}$;
- $U^{\mathrm{can}(\mathcal{T},\mathcal{A})} = \{(a, v) \mid U(a, v) \in \mathrm{chase}(\mathcal{T}, \mathcal{A})\}$;
- $Z^{\mathrm{can}(\mathcal{T},\mathcal{A})}(v) = \{\varphi \mid \varphi \in Z(v)\}$ for all $v \in \Delta^{\mathrm{can}(\mathcal{T},\mathcal{A})}_{\mathsf{data}}$.

1. If $C \sqsubseteq A \in \mathcal{T}$ and $C(a) \in S$ and $A(a) \notin S$, then add $A(a)$ to S.
2. If $C \sqsubseteq \exists r \in \mathcal{T}$ and $C(a) \in S$ and there is no $r(a, c) \in S$, then add $r(a, c')$ to S for a fresh individual c'.
3. If $B_1 \sqsubseteq \neg B_2 \in \mathcal{T}$ and $B_1(a), B_2(a) \in S$, then terminate and output 'unsatisfiable'.
4. If $C \sqsubseteq \exists U \in \mathcal{T}$ and $C(a) \in S$ and there is no data value or data null v with $U(c, v) \in S$ then add $U(a, u)$ to S and set $Z(u) = \emptyset$, where u is a fresh data null.
5. If $C \sqsubseteq \exists U.\varphi \in \mathcal{T}$ and $C(a) \in S$ and there is no data value or data null v with $U(c, v) \in S$ and $\varphi \in Z(v)$, then add $U(a, u)$ to S and set $Z(u) = \{\varphi(u)\}$, where u is a fresh data null. Terminate and output 'unsatisfiable' if $\mathcal{D} \not\models \exists x \varphi(x)$.
6. If $C \sqsubseteq \forall U.\varphi \in \mathcal{T}$ and $C(a) \in S$ and $U(c, u) \in S$ for u a data value or data null such $\varphi \notin Z(u)$, then add φ to $Z(u)$. Terminate and output 'unsatisfiable' if $\mathcal{D} \not\models \exists x \bigwedge_{\varphi \in Z(u)} \varphi(x)$.
7. If $U_1 \sqsubseteq U_2 \in \mathcal{T}$ and $U_1(a, u) \in S$ and $U_2(a, u) \notin S$, then add $U_2(a, u)$ to S.
8. If $r_1 \sqsubseteq r_2 \in \mathcal{T}$ and $r_1(a, u) \in S$ and $r_2(a, u) \notin S$, then add $r_2(a, b)$ to S.

The proof of the following lemma is standard and omitted.

Lemma A.6. *Given a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$ the chase procedure outputs 'unsatisfiable' iff $(\mathcal{T}, \mathcal{A})$ is unsatisfiable. If $(\mathcal{T}, \mathcal{A})$ is satisfiable, then $\mathrm{can}(\mathcal{T}, \mathcal{A})$ is hom-initial for $(\mathcal{T}, \mathcal{A})$ and thus a universal pre-model for $(\mathcal{T}, \mathcal{A})$.*

B Proofs for Section 4

We start with the proofs of the undecidability result.

Theorem 4.1 (restated). *Let $\mathcal{D} \in \{(\mathbb{Z}, \neq), (\mathbb{Z}, <), (\mathbb{Z}, \leq), (\mathbb{Q}, \neq), (\mathbb{Q}, <)\}$. Then the query evaluation problem for OMQs with DL-Lite$^{attrib}_{\mathcal{R}}(\mathcal{D})$ TBoxes is undecidable in combined complexity.*

We first prove undecidability of query evaluation for Boolean OMQs over the datatypes $(\mathbb{Z}, \neq)$ and $(\mathbb{Q}, \neq)$. We reduce the undecidable $\mathbb{N} \times \mathbb{N}$ tiling problem which is defined as follows: given a set $\mathfrak{T} = \{T_1, \ldots, T_n\}$ of tile types defined by the colors $\mathsf{right}(T_i)$, $\mathsf{left}(T_i)$, $\mathsf{up}(T_i)$, and $\mathsf{down}(T_i)$, for $1 \leq i \leq n$, does there exist a function f: $\mathbb{N} \times \mathbb{N} \to \mathfrak{T}$ such that

- $\mathsf{right}(f(i, j)) = \mathsf{left}(f(i + 1, j))$ for all $i, j \geq 0$;
- $\mathsf{up}(f(i, j)) = \mathsf{down}(f(i, j + 1))$ for all $i, j \geq 0$.

Assume such a tiling problem $\mathfrak{T}$ is given. We define a TBox $\mathcal{T}$ and Boolean UCQ q such that for a certain ABox $\mathcal{A}$ we have $(\mathcal{T}, \mathcal{A}) \models q$ iff $\mathfrak{T}$ does not tile $\mathbb{N} \times \mathbb{N}$. Let h and v be role names, X a concept name, and U, U^1_k, and U^2_k be attribute names for $1 \leq k \leq n$. We define the TBox $\mathcal{T}$ as the set of inclusions

$$C \sqsubseteq \exists h, \quad C \sqsubseteq \exists v, \quad C \sqsubseteq \exists U, \quad C \sqsubseteq \exists U^i_k$$

where C ranges over $X, \exists v$ and $\exists h$, and $1 \leq k \leq n$, $i = 1, 2$. Consider the ABox $\mathcal{A} = \{X(a)\}$. Then the universal pre-model $\mathrm{can}(\mathcal{T}, \mathcal{A})$ of $(\mathcal{T}, \mathcal{A})$ is the infinite binary tree with edges h (for 'horizontal') and v (for 'vertical') and root a. Each node c in the tree comes with assertions $U(c, u)$ and $U^i_k(c, u^i_k)$ for data nulls u and u^i_k. We construct a Boolean UCQ q such that

$$\mathcal{T}, \mathcal{A} \models q \quad \text{iff} \quad \mathfrak{T} \text{ does not tile } \mathbb{N} \times \mathbb{N}$$

We use the abbreviations

- $(x \not\sim y) = U(x, z_x) \wedge U(y, z_y) \wedge z_x \neq z_y$
- $(x \sim y) = U(x, z_{x,y}) \wedge U(y, z_{x,y})$

$(x \sim y)$ defines a natural equivalence relation on any completion $f(\mathrm{can}(\mathcal{T}, \mathcal{A}))$ of the universal pre-model of $(\mathcal{T}, \mathcal{A})$. Denote the corresponding quotient structure by $f(\mathrm{can}(\mathcal{T}, \mathcal{A}))/\sim$. We now assemble the UCQ q as the disjunction of the following CQs. The Boolean CQ q_1 has a match in $f(\mathrm{can}(\mathcal{T}, \mathcal{A}))$ if $f(\mathrm{can}(\mathcal{T}, \mathcal{A}))/\sim$ has a grid cell that is not closed:

$$\begin{aligned} q_1 \quad \leftarrow \quad & (x_1 \sim x_2) \wedge h(x_1, y_1) \wedge (y_1 \sim y_1') \wedge v(y_1', z_1) \wedge \\ & v(x_2, y_2) \wedge (y_2 \sim y_2') \wedge h(y_2', z_2) \wedge \\ & (z_1 \not\sim z_2) \end{aligned}$$

Abbreviate for $1 \leq k \leq n$

$$T_k = U^1_k(x, v) \wedge U^2_k(x, v)$$

which intuitively stands for "the tile type T_k is true at x". The next CQs state that if two nodes represent the same grid node, then the same tile is placed on them:

$$q_{k_1,k_2} \quad \leftarrow \quad (x \sim y) \wedge T_{k_1}(x) \wedge T_{k_2}(y)$$

for $k_1 \neq k_2$. The following CQs state that the matching condition for the tiles hold:

$$q_{h,k_1,k_2} \quad \leftarrow \quad h(x,y) \wedge T_{k_1}(x) \wedge T_{k_2}(y)$$

for $\mathsf{right}(T_{k_1}) \neq \mathsf{left}(T_{k_2})$ and

$$q_{v,k_1,k_2} \quad \leftarrow \quad v(x,y) \wedge T_{k_1}(x) \wedge T_{k_2}(y)$$

for $\mathsf{up}(T_{k_1}) \neq \mathsf{down}(T_{k_2})$. Finally, we have to enforce that in each node a tile type is true. This is achieved by

$$q_0 \quad \leftarrow \quad \bigwedge_{1 \leq k \leq n} (U_k^1(x, v_{1,k}) \wedge U_k^2(x, v_{2,k}) \wedge (v_{1,k} \neq v_{2,k}))$$

It is readily checked that the resulting UCQ is as required.

As $x \neq y$ iff $x < y$ or $y < x$, undecidability of query evaluation with inequality in the datatype entails undecidability of query evaluation with $<$ in the datatype. As we have just shown that query evaluation for Boolean OMQs over the datatypes $(\mathbb{Z}, \neq)$ and $(\mathbb{Q}, \neq)$ is undecidable, we obtain that query evaluation for Boolean OMQs over the datatypes $(\mathbb{Z}, <)$ and $(\mathbb{Q}, <)$ is undecidable. It remains to consider the case of $(\mathbb{Z}, \leq)$.

Theorem B.1. *Answering OMQs over $(\mathbb{Z}, \leq)$ with DL-Lite$_{\mathcal{R}}^{attrib}(\mathbb{Z}, \leq)$ TBoxes is undecidable in combined complexity.*

Proof. We reduce the problem of answering OMQs over $(\mathbb{Z}, <)$ with *DL-Lite*$_{\mathcal{R}}^{attrib}(\mathbb{Z}, <)$ TBoxes to the problem of answering OMQs over $(\mathbb{Z}, \leq)$ with *DL-Lite*$_{\mathcal{R}}^{attrib}(\mathbb{Z}, \leq)$ TBoxes. Since the former problem was shown to be undecidable above, the theorem follows.

Consider an OMQ $Q = (\mathcal{T}, q)$ over $(\mathbb{Z}, <)$ with $\mathcal{T}$ a *DL-Lite*$_{\mathcal{R}}^{attrib}(\mathbb{Z}, <)$ TBox and a $(\mathbb{Z}, <)$-ABox $\mathcal{A}$. We reduce Q and $\mathcal{A}$ to an OMQ Q' over $(\mathbb{Z}, \leq)$ with a *DL-Lite*$_{\mathcal{R}}^{attrib}(\mathbb{Z}, \leq)$ TBox and a $(\mathbb{Z}, \leq)$-ABox $\mathcal{A}'$ such that $\mathcal{A} \models Q$ if, and only if, $\mathcal{A}' \models Q'$.

Our reduction is based on a TBox $\mathcal{T}_{\text{ord}}$ and a UCQ q_{ord} over $(\mathbb{Z}, \leq)$ that we use to define two relations $\prec$ and $\approx$ that simulate $<$ and $=$ on the attributes that occur in $\mathcal{T}$. The basic building block is a linearly ordered sequence $(I_i)_{i \in \mathbb{Z}}$ of disjoint intervals such that each value of an attribute in $\mathcal{T}$ is contained in some interval I_i. Given two values u_1 and u_2 of an attribute in $\mathcal{T}$, we will set $u_1 \approx u_2$ if u_1 and u_2 belong to the same interval, and $u_1 \prec u_2$ if the interval of u_1 precedes the interval of u_2.

Let r and V be a role name and an attribute name, respectively, that do not occur in $\mathcal{T}$, q, or $\mathcal{A}$. We define:

$$\mathcal{T}_{\text{ord}} := \left\{ \exists r^- \sqsubseteq \exists r,\ \exists r \sqsubseteq \exists r^-,\ \exists r^- \sqsubseteq \exists V \right\}.$$

We also define q_{ord} to be the UCQ consisting of the following disjuncts:

$$q_{\text{ord}}^{<} \leftarrow r(x,y) \wedge V(x,z_x) \wedge V(y,z_y) \wedge z_y \leq z_x,$$
$$q_{\text{ord}}^{U} \leftarrow r(x,y) \wedge V(y,z) \wedge U(s,z), \quad \text{for each attribute name } U \text{ in } \mathcal{T}.$$

Next, we show that $\mathcal{T}_{\text{ord}}$ and q_{ord} define the desired sequence $(I_i)_{i \in \mathbb{Z}}$ of intervals. Consider a completion $\mathcal{I}$ of $\text{can}(\mathcal{T} \cup \mathcal{T}_{\text{ord}}, \mathcal{A} \cup \{r(a_0, a_1)\})$, where a_0 and a_1 are distinct individual names. By the construction of $\mathcal{I}$, $r^{\mathcal{I}}$ is an infinite path $\ldots, a_{-2}, a_{-1}, a_0, a_1, a_2, \ldots$ without endpoints, and for each $i \in \mathbb{Z}$ there is exactly one element $v_i \in \mathbb{Z}$ such that $(a_i, v_i) \in V^{\mathcal{I}}$. For each $i \in \mathbb{Z}$, let

$$I_i := \{v \in \mathbb{Z} \mid v_i < v < v_{i+1}\}.$$

Now suppose that $\mathcal{I} \not\models q_{\text{ord}}$. Then,

- for each $i \in \mathbb{Z}$ we have $v_i < v_{i+1}$; and
- for each $(a,u) \in U^{\mathcal{I}}$ there exists a unique $i \in \mathbb{Z}$ such that $u \in I_i$.

For the second claim, observe that because the path $r^{\mathcal{I}}$ is infinite and $(\mathbb{Z}, \leq)$ is discrete, there exist $j_1 < j_2$ such that $v_{j_1} \leq u \leq v_{j_2}$. Since $\mathcal{I} \not\models q_{\text{ord}}^U$ enforces $u \neq v_i$ for all $i \in \mathbb{Z}$, this implies $u \in I_i$ for some $i \in \{j_1, j_1 + 1, \ldots, j_2 - 1\}$. Uniqueness of i follows from the disjointness of the intervals I_i. Altogether, this shows that $(I_i)_{i \in \mathbb{Z}}$ is the desired sequence of intervals.

We introduce the following abbreviations:

$$\begin{aligned} z \in (x,y) :\iff\ & r(x,y) \wedge V(x,v_x) \wedge V(y,v_y) \wedge \\ & v_x \leq z \wedge z \leq v_y, \\ (x,y) < (x',y') :\iff\ & r(x,y) \wedge r(x',y') \wedge V(y,v_y) \wedge \\ & V(x',v_{x'}) \wedge v_y \leq v_{x'}. \end{aligned}$$

In a model $\mathcal{I}$ as above, "$z \in (x,y)$" means that $(x,y) = (a_i, a_{i+1})$ for some $i \in \mathbb{Z}$ and $z \in I_i$, and "$(x,y) < (x',y')$" means that there are $i < j$ such that $(x,y) = (a_i, a_{i+1})$ and $(x',y') = (a_j, a_{j+1})$.

We use these abbreviations to define the desired orderings $\prec$ and $\approx$:

$$\begin{aligned} z_1 \approx z_2 :\iff\ & z_1 \in (x,y) \wedge z_2 \in (x,y), \\ z_1 \prec z_2 :\iff\ & z_1 \in (x_1,y_1) \wedge z_2 \in (x_2,y_2) \wedge \\ & (x_1,y_1) < (x_2,y_2), \end{aligned}$$

where "$z_1 \approx z_2$" means that z_1 and z_2 belong to the same interval I_i, and "$z_1 \prec z_2$" means that z_1 belongs to an interval that precedes the interval of z_2.

To conclude, we let $Q' := (\mathcal{T} \cup \mathcal{T}_{\text{ord}}, q')$, where q' is the UCQ given by q'', q_{ord} and q'' is obtained from q by expressing $<$ and equality on data variables in terms of $\prec$ and $\approx$. We also let $\mathcal{A}' := \mathcal{A} \cup \{r(a_0, a_1)\}$. It is straightforward to show that $\mathcal{A} \models Q$ iff $\mathcal{A}' \models Q'$. □

Safe Queries and Homogeneous Datatypes

This section provides a proof for the claim, made in Lemma 4.3, that safe OMQs over homogeneous datatypes have the BMDP. Without loss of generality, we will focus on *Boolean* safe queries.

Recall from Section 4 that a datatype $\mathcal{D}$ is *homogeneous* if every isomorphism between finite induced substructures of $\mathcal{D}$ extends to an automorphism on $\mathcal{D}$ (Chang and Keisler 1998). Examples of homogeneous datatypes are $(\mathbb{Q}, <)$ and $(\mathbb{Q}, \leq)$. Also recall that a CQ q over a datatype $\mathcal{D}$ is *safe* if any two non-data variables of q are either connected in q or are both connected to non-data answer variables, where two

variables x, y are said to be *connected* in q if x and y are connected via a path of atoms in q that does not contain data variables. Note that, since we focus on Boolean safe CQs, any two non-data variables of q are connected in q. A UCQ is safe if all its disjuncts are safe, and an OMQ is safe if its UCQ is safe.

We use some of the concepts and notations that were introduced in Section 3. In particular, we will assume familiarity with the chase procedure and the pre-model $\mathrm{can}(\mathcal{T}, \mathcal{A})$ for a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$. It will also be convenient to not distinguish between pre-interpretations and their equivalent representation as pre-ABoxes.

We heavily use the following forest representation of the pre-model $\mathrm{can}(\mathcal{T}, \mathcal{A})$.

Definition B.2. Let $(\mathcal{T}, \mathcal{A})$ be a $\mathcal{D}$-KB for a datatype $\mathcal{D}$. The *chase forest* of $\mathcal{T}$ and $\mathcal{A}$, denoted by $\mathrm{CF}(\mathcal{T}, \mathcal{A})$, is defined inductively as follows.

We start with an empty forest and add, for each assertion $\alpha \in \mathcal{A}$, a new root node v_α labelled with the set $B(v_\alpha) := \{\alpha\}$.

For the induction step, consider a node v of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ and let $B(v)$ be its label. Suppose some rule of the chase procedure can be applied to the pre-ABox $B(v)$ and generates a new atom β.[4] Let v' be the lowest ancestor of v such that $B(v')$ contains all elements that occur both in $B(v)$ and as arguments of β. If all arguments of β occur in $B(v')$, then we add β to $B(v')$. Otherwise, we create a new child v'' of v' and let $B(v'') := \{\beta\}$ be its label.

When we apply chase rules that introduce constraints for data nulls, we also add these constraints to the pre-ABox $B(v)$ of the corresponding node v.

For each node v of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$, we call $B(v)$ the *bag* of v and denote the *depth* of v by $\mathrm{depth}(v)$. Given a set V of nodes of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$, we define

$$B(V) := \bigcup_{v \in V} B(v).$$

It is easy to see that the set of all assertions that occur in the bags of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ coincides with $\mathrm{can}(\mathcal{T}, \mathcal{A})$, up to renaming of individuals and data nulls that do not occur in $\mathcal{A}$. In the following, we may therefore assume without loss of generality that the set of all assertions that occur in the bags of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ coincides *exactly* with $\mathrm{can}(\mathcal{T}, \mathcal{A})$. The following lemma is immediate from the construction of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$.

Lemma B.3. *Let $(\mathcal{T}, \mathcal{A})$ be a $\mathcal{D}$-KB. If v is a node of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$, then there are at most two elements that occur as arguments of assertions in $B(v)$, and v has at most $2 \cdot |\mathcal{T}|$ children.*

We will also require the following concepts of an *ℓ-neighborhood* and an *ℓ-type*. Informally, the ℓ-neighborhood of a node v of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ contains all assertions in bags of nodes at distance less than ℓ from v. The ℓ-type of v w.r.t. a completion function f enriches the ℓ-neighborhood $\mathcal{N}$ of v with information about the relationships between the data values assigned by f to the data nulls in $\mathcal{N}$.

[4] Here we assume that fresh constants or data nulls that are generated by chase rules of the type 2, 4, or 5 do not occur already in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$.

Definition B.4. Let $(\mathcal{T}, \mathcal{A})$ be a $\mathcal{D}$-KB, let v be a node of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$, let $\ell \geq 1$, and let f a completion function for $\mathrm{can}(\mathcal{T}, \mathcal{A})$.

- The *ℓ-neighborhood* of v in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ is defined as $\mathcal{N}_\ell(v) := B(N_\ell(v))$, where $N_\ell(v)$ is the set of all nodes at distance less than ℓ from v in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$.
- The *ℓ-type* of v in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ with respect to f, denoted by $\mathrm{tp}_\ell(v)$, is the union of $\mathcal{N}_\ell(v)$ and the set of all atoms $R(u_1, \ldots, u_r)$ over $\mathcal{D}$ with each u_i a data value or a data null in $\mathcal{N}_\ell(v)$ and $(f(u_1), \ldots, f(u_r)) \in R$.

Let $q \leftarrow \varphi$ be a Boolean CQ. The *size* of q, denoted by $|q|$, is the number of atoms that occur in φ. Given a pre-match μ of q in $\mathrm{can}(\mathcal{T}, \mathcal{A})$, we define $\mu(q)$ to be the image of q under μ, i.e., the smallest sub-interpretation $\mathcal{I}$ of $\mathrm{can}(\mathcal{T}, \mathcal{A})$ such that μ is a match of q in $\mathcal{I}$.

Lemma B.5. *Let $(\mathcal{T}, \mathcal{A})$ be a $\mathcal{D}$-KB, and let f be a completion function for $\mathrm{can}(\mathcal{T}, \mathcal{A})$.*

1. *Consider a safe CQ q and a pre-match μ of q in $\mathrm{can}(\mathcal{T}, \mathcal{A})$. Suppose that there is a node v of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ at depth at least $|q|$ whose bag contains some atom of $\mu(q)$. Then, $\mu(q)$ is contained in $\mathcal{N}_{|q|}(v) \subseteq \mathrm{tp}_{|q|}(v)$.*
2. *If the number of relations in $\mathcal{D}$ is finite, then for each $\ell \geq 1$ the number of distinct ℓ-types of nodes of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ with respect to f is at most*

$$t_\ell := 2^{m \cdot |\mathcal{T}|^{\mathcal{O}(\ell)}},$$

where m is the total number of concept names, role names, and attribute names that occur in $\mathcal{T}$.

Proof. Ad 1: Easy consequence of the definition of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ and that of safe CQs.

Ad 2: Let v be a node of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$. By Lemma B.3, the set $N_\ell(v)$ of all nodes at distance less than ℓ from v has size at most

$$\sum_{i=0}^{\ell-1} (2 \cdot |\mathcal{T}| + 1)^i \leq (2 \cdot |\mathcal{T}| + 1)^\ell.$$

Let $X := \Delta^{\mathcal{N}_\ell(v)}$. Since the bag of each node in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ contains at most two elements (Lemma B.3), we have

$$|X| \leq 2 \cdot (2 \cdot |\mathcal{T}| + 1)^\ell = \mathcal{O}(|\mathcal{T}|)^\ell.$$

Each type consists of atoms built from concept names, role names, attribute names, and relation names in $\mathcal{D}$, using the elements in X as arguments. The number of such atoms is at most $m_1 \cdot |X|^2 + m_2 \cdot |X|^r$, where m_1 is the number of concept names, role names, and attribute names in $\mathcal{T}$, m_2 is the number of relations in $\mathcal{D}$, and r is the maximum arity of a relation in $\mathcal{D}$. Consequently, the number of distinct ℓ-types of nodes of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ with respect to f is at most

$$t_\ell = 2^{m_1 \cdot |X|^2 + m_2 \cdot |X|^r} = 2^{m_1 \cdot |\mathcal{T}|^{\mathcal{O}(\ell)}}.$$

Here we use that m_2 and r are constant (because $\mathcal{D}$ is constant). □

We are now ready to show that safe OMQs over homogeneous datatypes enjoy the BMDP. The following lemma states this result for TBoxes without attribute restrictions. Remark B.7 outlines how to adapt the proof to the case of TBoxes with attribute restrictions.

Lemma B.6. *Let $Q = (\mathcal{T}, q)$ be a safe Boolean OMQ over a homogeneous datatype $\mathcal{D}$ with $\mathcal{T}$ a DL-Lite$_{\mathcal{R}}^{attrib}(\mathcal{D})$ TBox. Set $d := 2s + t_s$, where s is the maximum size of a disjunct of q and t_s is defined as in Lemma B.5.*[5]

For every $\mathcal{D}$-ABox $\mathcal{A}$ that is satisfiable relative to $\mathcal{T}$, the following are equivalent:

1. *$f(\mathrm{can}(\mathcal{T}, \mathcal{A})) \models q$ for all completion functions f.*
2. *$f(\mathrm{can}^d(\mathcal{T}, \mathcal{A})) \models q$ for all completion functions f.*

Proof. Note that since $\mathcal{T}$ and q are finite and therefore refer to only finitely many relations of $\mathcal{D}$, we can assume without loss of generality that $\mathcal{D}$ contains only finitely many relations.

Let $\mathcal{A}$ be a $\mathcal{D}$-ABox that is satisfiable relative to $\mathcal{T}$. Clearly, if q is true in all completions of $\mathrm{can}^d(\mathcal{T}, \mathcal{A})$, then q is true in all completions of $\mathrm{can}(\mathcal{T}, \mathcal{A})$. Thus, it remains to prove the other direction: if q is false in some completion of $\mathrm{can}^d(\mathcal{T}, \mathcal{A})$, then q is false in some completion of $\mathrm{can}(\mathcal{T}, \mathcal{A})$.

To this end, suppose that there is a completion function f for $\mathrm{can}^d(\mathcal{T}, \mathcal{A})$ with $f(\mathrm{can}^d(\mathcal{T}, \mathcal{A})) \not\models q$. We use f to construct a completion function $\hat{f}$ for $\mathrm{can}(\mathcal{T}, \mathcal{A})$ with $\hat{f}(\mathrm{can}(\mathcal{T}, \mathcal{A})) \not\models q$.

We construct $\hat{f}$ inductively. Let $v_1, v_2, v_3, \ldots$ be a repetition-free enumeration of all the nodes of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ at depth at least $d+1$ in a breadth-first fashion. For each $i \geq 0$ define the pre-interpretation

$$\mathcal{I}_i := \mathrm{can}^d(\mathcal{T}, \mathcal{A}) \cup B\big(\{v_j \mid 1 \leq j \leq i\}\big).$$

For each $i \geq 0$ we construct a completion function f_i for $\mathcal{I}_i$ with the following two properties:

1. $f_i(\mathcal{I}_i) \not\models q$.
2. If $i \geq 1$, then f_i coincides with f_{i-1} on all elements that occur in a bag of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ at depth less than $s + \mathrm{depth}(v_i) - d$.

Finally, we let $\hat{f}$ be the completion function of $\mathrm{can}(\mathcal{T}, \mathcal{A})$ that maps each data null u to $f_i(u)$, where i is the smallest integer with $f_i(u) = f_j(u)$ for all $j \geq i$ (this integer exists by the second property above). We then have $\hat{f}(\mathrm{can}(\mathcal{T}, \mathcal{A})) \not\models q$. It remains to construct the functions f_i.

We let $f_0 := f$. Then, f_0 is a completion function for $\mathcal{I}_0 = \mathrm{can}^d(\mathcal{T}, \mathcal{A})$ such that $f_0(\mathcal{I}_0) \not\models q$. The second property is trivially satisfied for f_0.

Next assume that f_i has been constructed. We wish to construct f_{i+1}. Let W be the set of the $t_s + 1$ deepest ancestors of v_{i+1} of depth at most $\mathrm{depth}(v_{i+1}) - s$. Then, for each node $w \in W$,

$$\mathrm{depth}(w) \leq \mathrm{depth}(v_{i+1}) - s, \tag{1}$$

$$\begin{aligned}\mathrm{depth}(w) &\geq \mathrm{depth}(v_{i+1}) - s - t_s \\ &= s + \mathrm{depth}(v_{i+1}) - d.\end{aligned} \tag{2}$$

Note that (1) implies $\mathcal{N}_s(w) \subseteq \mathcal{I}_i$, and therefore f_i is defined on all data nulls in $\mathcal{N}_s(w)$. Moreover, (2) and $\mathrm{depth}(v_{i+1}) \geq d+1$ imply that each node in W has depth at least $s+1$. Consequently, by Lemma B.5(1), each match of some disjunct of q in $\mathrm{can}(\mathcal{T}, \mathcal{A})$ that contains an atom from the bag of a node $w \in W$ is contained in $\mathcal{N}_s(w)$.

Now, since W contains $t_s + 1$ many nodes, there must be two nodes $w_1, w_2 \in W$ with $\mathrm{tp}_s(w_1) \cong \mathrm{tp}_s(w_2)$. Say, w_1 is an ancestor of w_2. Let g_0 be an isomorphism from $\mathrm{tp}_s(w_1)$ to $\mathrm{tp}_s(w_2)$. We can extend this isomorphism to an isomorphism g that includes the atoms that occur in the subtree rooted at w_1. For each data value or data null u in $\mathcal{N}_s(w_1)$, let $h_0(f_i(u)) := f_i(g(u))$. Then, h_0 is an isomorphism between finite induced substructures of $\mathcal{D}$. Since $\mathcal{D}$ is homogeneous, h_0 extends to an automorphism h of $\mathcal{D}$. Now, for each data null u in $\mathrm{can}(\mathcal{T}, \mathcal{A})$, let $f_{i+1}(u) := f_i(u)$ if u does not occur in $\mathcal{I}_{i+1}$ in the subtree below w_2, and $f_{i+1}(u) := h(f_i(g^{-1}(u)))$ otherwise.

It is easy to see that f_{i+1} coincides with f_i on all elements that occur in a bag of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ at depth less than $s + \mathrm{depth}(v_{i+1}) - d$. It remains to show that $f_{i+1}(\mathcal{I}_{i+1}) \not\models q$. Suppose that $f_{i+1}(\mathcal{I}_{i+1}) \models q$. Then there is a match μ of some disjunct q' of q in $f_{i+1}(\mathcal{I}_{i+1})$. Since μ is not a match of q' in $f_i(\mathcal{I}_i)$ (by the induction hypothesis), and f_{i+1} coincides with f_i on all atoms that occur outside the subtree of w_2 (by construction), it must be the case that $\mu(q')$ contains an atom from the subtree of w_2. This implies that $\mu(q') \subseteq \mathcal{N}_s(w_2)$ or $\mu(q')$ is contained in the subtree rooted at w_2. But then the mapping

$$\mu'(x) := \begin{cases} g^{-1}(\mu(x)), & \text{if } x \text{ is not a data variable,} \\ h^{-1}(\mu(x)), & \text{if } x \text{ is a data variable} \end{cases}$$

is a match of q' in $f_i(\mathcal{I}_i)$, a contradiction. □

Remark B.7. If $Q = (\mathcal{T}, q)$ is a safe Boolean OMQ over a homogeneous datatype $\mathcal{D}$ with $\mathcal{T}$ a *DL-Lite*$_{\mathcal{R}}^{qattrib}(\mathcal{D})$ TBox, then we augment the proof of Lemma B.6 as follows. Let C be the set of all elements of $\mathrm{dom}(\mathcal{D})$ that occur in attribute restrictions of $\mathcal{T}$. We use an extended notion of ℓ-neighborhood: the *extended ℓ-neighborhood* of a node v in $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ is defined as $\mathcal{N}_\ell(v) \cup C$. Note that this also induces an extended notion of ℓ-type, which is based on the extended ℓ-neighborhood rather than the ℓ-neighborhood. The extended notion of ℓ-type captures the relationship of data values and data nulls in the ℓ-neighborhood of a node and the elements in C. Finally, we consider the extended datatype $\hat{\mathcal{D}} = (\mathcal{D}, c_1, \ldots, c_n)$, where $c_1, \ldots, c_n$ is an enumeration of all the elements in C. This restricts isomorphisms and automorphisms to be the identity on the elements on C. Consequently, when we translate the completion function of a subtree of $\mathrm{CF}(\mathcal{T}, \mathcal{A})$ by applying an automorphism of $\hat{\mathcal{D}}$, we automatically preserve all the constraints imposed on the

[5] Here we assume that $\mathcal{D}$ contains only relations that occur in $\mathcal{T}$ or q. In particular, $\mathcal{D}$ can be assumed to be finite.

data nulls. Note that $\hat{\mathcal{D}}$ is homogeneous if $\mathcal{D}$ is homogeneous. This means that all other details of the proof can be left unchanged.

Corollary B.8. *Every safe OMQ over a homogeneous datatype enjoys the BMDP.*

Satisfiability

For the proof of Theorem 4.5 is remains to provide a polynomial reduction of satisfiability of $\mathcal{D}$-ABoxes $\mathcal{A}$ relative to TBoxes $\mathcal{T}$ with $\theta_{\mathcal{T}} = \{\varphi_1, \dots, \varphi_n\}$ to $\mathsf{CSP}_c(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n})$.

Lemma B.9. *Let $\mathcal{T}$ be a TBox with $\theta_{\mathcal{T}} = \{\varphi_1, \dots, \varphi_n\}$. Then satisfiability of $\mathcal{D}$-ABoxes $\mathcal{A}$ relative to $\mathcal{T}$ is polynomially reducible to $\mathsf{CSP}_c(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n})$.*

Proof. Consider a $\mathcal{D}$-KB $(\mathcal{T}, \mathcal{A})$. We modify the chase procedure introduced above in such a way that it always terminates after polynomially many steps and either outputs "unsatisfiable" or constructs an instance Φ_{sat} of $\mathsf{CSP}_c(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n})$ such that $(\mathcal{T}, \mathcal{A})$ is satisfiable iff $(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n}) \models \Phi_{\text{sat}}$.

The only rule that leads to non-termination of our current chase procedure is Rule 2. Now, instead of always introducing a fresh individual we only introduce a fresh individual (denoted $c_{\exists r}$) when Rule 2 is triggered by an inclusion with $\exists r$ on its right hand side for the first time. Afterwards, Rule 2 adds assertions $r(a, c_{\exists r})$ to S whenever $C \sqsubseteq \exists r \in \mathcal{T}$, $C(a) \in S$, and there is no $r(a, c) \in S$. Termination after polynomially many steps is clear.

In addition, we modify the Rules 5 and 6 of the chase procedure. In Rule 5 and 6 we do not terminate nor check any CSP. Instead, after termination of the chase without using Rule 3, we construct an instance Φ_{sat} of $\mathsf{CSP}_c(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n})$ as

$$\Phi_{\text{sat}} = \exists \bar{u} \bigwedge_{i=1}^{k} \bigwedge_{\varphi \in Z^{\mathcal{I}}(u_i)} R_\varphi(u_i)$$

where $u_1, \dots, u_k$ is a repetition-free enumeration of all data nulls that occur in the chase. It is readily checked that $(\mathcal{T}, \mathcal{A})$ is satisfiable iff $(\mathrm{dom}(\mathcal{D}), R_{\varphi_1}, \dots, R_{\varphi_n}) \models \Phi_{\text{sat}}$. □

C Proofs for Section 5

Temporal Constraint Satisfaction

Our dichotomy result of Section 5 uses the temporal constraint satisfaction framework of Bodirsky and Kára (2010a). This section reviews the relevant definitions and results of this framework. For details, we refer the interested reader to (Bodirsky and Kára 2010a).

A *temporal constraint language* is a structure of the form $\Gamma = (\mathbb{Q}, R_1, R_2, \dots)$, where each R_i is a relation on $\mathbb{Q}$ that is definable by a first-order formula $\Phi_i(x_1, \dots, x_k)$ over $(\mathbb{Q}, <)$, with k the arity of R_i:

$$R_i = \{(a_1, \dots, a_k) \in \mathbb{Q}^k \mid (\mathbb{Q}, <) \models \Phi_i(a_1, \dots, a_k)\}.$$

The formulas Φ are not allowed to contain elements of $\mathbb{Q}$ as constants.

Bodirsky and Kára prove that for every temporal constraint language Γ, $\mathsf{CSP}(\Gamma)$ is either in PTime or NP-complete. They characterize the temporal languages Γ for which $\mathsf{CSP}(\Gamma)$ is tractable in terms of preservation properties of the relations in Γ. Here, we say that a function $f\colon \mathbb{Q}^k \to \mathbb{Q}$ *preserves* a relation $R \subseteq \mathbb{Q}^n$ if for all $t_1, \dots, t_k \in R$ we have $f[t_1, \dots, t_k] \in R$, where the tuple $f[t_1, \dots, t_k]$ is obtained as follows. Given a tuple t of length n and an integer $i \in \{1, \dots, n\}$, let $t[i]$ denote the ith component of t. Then,

$$f[t_1, \dots, t_k] := \big(f(t_1[1], \dots, t_k[1]), \dots, f(t_1[n], \dots, t_k[n])\big).$$

We say that f preserves a temporal constraint language Γ if f preserves all relations in Γ.

The following functions are considered in (Bodirsky and Kára 2010a):

- $min\colon \mathbb{Q}^2 \to \mathbb{Q}$ which maps its two arguments to the minimal one;
- $mi\colon \mathbb{Q}^2 \to \mathbb{Q}$ which maps $(x, y) \in \mathbb{Q}^2$ to $\alpha(x)$ if $x = y$, to $\beta(y)$ if $x > y$, and to $\gamma(x)$ if $x < y$, where α, β, γ are any functions with $\alpha(x) < \beta(x) < \gamma(x) < \alpha(y)$ for all $x < y$;
- $mx\colon \mathbb{Q}^2 \to \mathbb{Q}$ which maps $(x, y) \in \mathbb{Q}^2$ to $\beta(x)$ if $x = y$, and to $\alpha(\min\{x, y\})$ if $x \neq y$, where α, β are any functions with $\alpha(x) < \beta(x) < \alpha(y)$ for all $x < y$;
- $ll\colon \mathbb{Q}^2 \to \mathbb{Q}$ which is any function that satisfies $ll(x, y) < ll(x', y')$ iff $x \leq 0$ and (x, y) is lexicographically smaller than (x', y'), or $x, x' > 0$ and (y, x) is lexicographically smaller than (y', x');
- constant functions $f\colon \mathbb{Q}^k \to \{a\}$, where $k \geq 0$ and $a \in \mathbb{Q}$;
- the dual of $f \in \{min, mi, mx, ll\}$, which maps $(x, y) \in \mathbb{Q}^2$ to $-f(-x, -y)$.

Let $\mathcal{F}$ be the set consisting of min, mi, mx, ll, and all constant functions. Let *dual-$\mathcal{F}$* be the set of the duals of functions in $\mathcal{F}$.

Theorem 5.2 (restated) (Bodirsky and Kára 2010a) *Let Γ be a temporal constraint language. If Γ is preserved by a function in $\mathcal{F} \cup$ dual-$\mathcal{F}$, then $\mathsf{CSP}(\Gamma)$ is in PTime. Otherwise, $\mathsf{CSP}(\Gamma)$ is NP-complete.*

We will later need the following easy result regarding preservation of the weak linear order $<$ on $\mathbb{Q}$ by functions in $\mathcal{F} \cup$ *dual-$\mathcal{F}$*.

Proposition C.1.

1. *$<$ is preserved by min, mi, mx, ll and their duals.*
2. *$<$ is not preserved by any constant function.*

Proof. We only consider preservation under min, mi, mx, ll and constant functions. The proofs for the duals of min, mi, mx, ll are similar.

Preservation under min: Let $a_1 < a_2$ and $b_1 < b_2$. We have to show that $c_1 < c_2$, where $c_i := min(a_i, b_i)$. If $c_1 = a_1$, then $c_1 = a_1 < a_2$ and $c_1 = a_1 \leq b_1 < b_2$, thus

$c_1 < c_2$. Similarly, if $c_1 = b_1$, then $c_1 = b_1 \leq a_1 < a_2$ and $c_1 = b_1 < b_2$, thus $c_1 < c_2$. This shows that $<$ is preserved under min.

Preservation under mi: Let $a_1 < a_2$ and $b_1 < b_2$. We have to show that $c_1 < c_2$, where $c_i := mi(a_i, b_i)$. Since $<$ is preserved by min, we have $\min(a_1, b_1) < \min(a_2, b_2)$. Hence, $c_1 = mi(a_1, b_1) < mi(a_2, b_2) = c_2$. Altogether, this shows that $<$ is preserved under mi.

Preservation under mx: Let $a_1 < a_2$ and $b_1 < b_2$. We have to show that $c_1 < c_2$, where $c_i := mx(a_i, b_i)$. Since $<$ is preserved by min, we have $\min(a_1, b_1) < \min(a_2, b_2)$. This implies $c_1 = mx(a_1, b_1) < mx(a_2, b_2) = c_2$. Altogether, we have shown that $<$ is preserved under mx.

Preservation under ll: Let $a_1 < a_2$ and $b_1 < b_2$. We have to show that $ll(a_1, b_1) < ll(a_2, b_2)$. If $a_1 \leq 0$, then $a_1 < a_2$ immediately yields $ll(a_1, b_1) < ll(a_2, b_2)$. Now suppose that $a_1 > 0$. Since $a_1 < a_2$, we also have $a_2 > 0$. But then, $b_1 < b_2$ immediately yields $ll(a_1, b_1) < ll(a_2, b_2)$.

Non-preservation under constant functions: For a contradiction, suppose that $<$ is preserved under a constant function $f\colon \mathbb{Q}^k \to \{c\}$. Take any $a_1 < b_1, \ldots, a_k < b_k$. Since f preserves $<$, we obtain $c = f(a_1, \ldots, a_k) < f(b_1, \ldots, b_k) = c$, which is impossible. □

A Basic Dichotomy

In this section, we combine Theorem 4.5 and Theorem 5.2 to obtain a basic dichotomy for evaluating OMQs over $(\mathbb{Q}, \leq)$ based on their datatype patterns. This is an intermediate step for the proof of Theorem 5.1.

Theorem 5.3 (restated) Let $\theta = (\theta_{\mathcal{T}}, \theta_q)$ be a datatype pattern over $(\mathbb{Q}, \leq)$, where $\theta_q = \{\psi_1, \ldots, \psi_n\}$.

1. If some function $f \in \mathcal{F} \cup \textit{dual-}\mathcal{F}$ preserves each $R_{\neg\psi_i}$, then evaluating OMQs Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ and the BMDP is in PTime.
2. Otherwise, there is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ such that evaluating Q is coNP-complete.

Proof. 1. Let Q be an OMQ over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ that enjoys the BMDP. By Theorem 4.5, evaluating Q is polynomially reducible to the complement of $\mathsf{CSP}_c(\Gamma_\theta)$. We show that $\mathsf{CSP}_c(\Gamma_\theta)$ is polynomially reducible to $\mathsf{CSP}(\Gamma)$, where

$$\Gamma = (\mathbb{Q}, <, \leq, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n}),$$

and that $\mathsf{CSP}(\Gamma)$ is in PTime. This suffices to establish the first part of the theorem.

To show that $\mathsf{CSP}_c(\Gamma_\theta)$ is polynomially reducible to $\mathsf{CSP}(\Gamma)$, consider an instance Φ of $\mathsf{CSP}_c(\Gamma_\theta)$. Replacing in Φ each atom of the form $R_\varphi(x)$, $\varphi \in \theta_{\mathcal{T}}$, by $\varphi(x)$ yields an instance Φ' of $\mathsf{CSP}_c(\mathbb{Q}, \leq, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n})$ with

$$\Gamma_\theta \models \Phi \iff (\mathbb{Q}, \leq, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n}) \models \Phi'.$$

Next, we eliminate all constants from Φ'. Let $c_1 < \cdots < c_k$ be the sequence of all elements of $\mathbb{Q}$ that occur as constants in Φ'. We simulate these constants by making each c_i an existentially quantified variable and adding constraints $c_i < c_{i+1}$, for each $i \in \{1, \ldots, k-1\}$, to ensure that any assignment preserves the relative order of these constants:

$$\Phi'' = \exists c_1 \cdots \exists c_k \left(\Phi' \wedge c_1 < c_2 \wedge \cdots \wedge c_{k-1} < c_k\right).$$

We thus obtain an instance Φ'' of $\mathsf{CSP}(\Gamma)$. We claim:

$$(\mathbb{Q}, \leq, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n}) \models \Phi' \iff \Gamma \models \Phi''.$$

The direction from left to right follows from the construction of Φ''. For the converse, assume $\Gamma \models \Phi''$. Let g be an assignment of rational numbers to the existential variables in Φ'' that satisfies the quantifier-free part of Φ'' in Γ. Pick any automorphism α of $(\mathbb{Q}, <)$ such that $\alpha(g(c_i)) = c_i$ for all $i \in \{1, \ldots, k\}$. Then $\alpha \circ g$ also satisfies the quantifier-free part of Φ'' in Γ. Since $\alpha \circ g$ interprets each c_i by itself, this implies $(\mathbb{Q}, \leq, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n}) \models \Phi'$. The sentence Φ'' can clearly be computed in polynomial time on input Φ. Altogether, we have shown that $\mathsf{CSP}_c(\Gamma_\theta)$ is polynomially reducible to $\mathsf{CSP}(\Gamma)$.

It remains to show that $\mathsf{CSP}(\Gamma)$ is in PTime. This is trivial if each of the ψ_i is empty. Otherwise, at least one of the ψ_i is non-empty, which implies that f is not a constant function. Since $<$ and $\leq$ are preserved under any non-constant function in $\mathcal{F} \cup \textit{dual-}\mathcal{F}$ (Proposition C.1), we know Γ is preserved under f and thus $\mathsf{CSP}(\Gamma)$ is in PTime (by Theorem 5.2).

2. By the theorem's hypothesis,

$$\Gamma = (\mathbb{Q}, R_{\neg\psi_1}, \ldots, R_{\neg\psi_n})$$

is not preserved by any function in $\mathcal{F} \cup \textit{dual-}\mathcal{F}$, which implies that $\mathsf{CSP}(\Gamma)$ is NP-complete (by Theorem 5.2). Since Γ is a substructure of Γ_θ, $\mathsf{CSP}_c(\Gamma_\theta)$ is also NP-complete. By Theorem 4.5, there is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = \theta$ such that the complement of $\mathsf{CSP}_c(\Gamma_\theta)$ is polynomially reducible to evaluating Q. This concludes the proof of the second part of the theorem. □

Structure of 'Tractable' Datatype Patterns

Theorem 5.3 establishes a basic P/coNP-dichotomy for evaluating OMQs over $(\mathbb{Q}, \leq)$ that enjoy the BMDP. The tractable cases of this dichotomy are characterized in terms of preservation properties of the relations $R_{\neg\psi}$, where ψ is a formula in the UCQ part θ_q of the datatype pattern. To obtain a purely syntactic characterization of these tractable cases, we here analyze the structure of formulas $\psi \in \theta_q$ such that $R_{\neg\psi}$ is preserved under one of the functions in $\mathcal{F} \cup \textit{dual-}\mathcal{F}$. This analysis is one of the main ingredients for our proof of Theorem 5.1.

We start by proving two auxilliary lemmas. The first lemma is straightforward but provides a useful tool for the proof of the second lemma, which is the core of the analysis.

Recall that for a tuple $t = (t_1, \ldots, t_n) \in \mathbb{Q}^n$ and an integer $i \in \{1, \ldots, n\}$, we have defined $t[i] = t_i$.

Lemma C.2. *Consider a function $f\colon \mathbb{Q}^2 \to \mathbb{Q}$ and elements $a_1, \ldots, a_4, b_1, \ldots, b_4 \in \mathbb{Q}$ such that*

$$f(a_1, b_1) \geq \cdots \geq f(a_4, b_4).$$

Let $1 \leq i_1 \leq i_2 \leq i_3 \leq i_4 \leq n$, and suppose that $(a_j, b_j) = (a_{j'}, b_{j'})$ if $i_j = i_{j'}$. Then, there are tuples $t_1, t_2 \in \mathbb{Q}^n$ such that $t_1[i_j] = a_j$ and $t_2[i_j] = b_j$ for all $j \in \{1, 2, 3, 4\}$, and

$$f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n]).$$

Proof. Define $t_1, t_2 \in \mathbb{Q}^n$ such that for all $p \in \{1, \ldots, n\}$, we have that

$$t_1[p] := \begin{cases} a_1, & \text{if } p \leq i_1 \\ a_2, & \text{if } i_1 < p \leq i_2 \\ a_3, & \text{if } i_2 < p \leq i_3 \\ a_4, & \text{if } i_3 < p \end{cases}$$

and

$$t_2[p] := \begin{cases} b_1, & \text{if } p \leq i_1 \\ b_2, & \text{if } i_1 < p \leq i_2 \\ b_3, & \text{if } i_2 < p \leq i_3 \\ b_4, & \text{if } i_3 < p. \end{cases}$$

Clearly, $t_1[i_j] = a_j$ and $t_2[i_j] = b_j$ for all $j \in \{1, 2, 3, 4\}$. From the construction of t_1 and t_2 and the properties of $a_1, \ldots, a_4, b_1, \ldots, b_4$, it immediately follows that $f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n])$. □

Consider a datatype pattern $\theta = (\theta_{\mathcal{T}}, \theta_q)$ over $(\mathbb{Q}, \leq)$. What is the structure of formulas $\psi \in \theta_q$ for which $R_{\neg\psi}$ is preserved by a function in $\mathcal{F} \cup \textit{dual-}\mathcal{F}$? Since ψ is acyclic (by assumption; see Section 5), the negation of ψ that defines $R_{\neg\psi}$ is equivalent to a disjunction Ψ of atomic formulas $x < y$, with x and y variables, such that the directed graph with the variables of Ψ as its vertices, and edges (y, x) for each atomic formula $x < y$ of Ψ is acyclic. We call such formulas Ψ *acyclic disjunctive* formulas.

Lemma C.3. *Let $R \subseteq \mathbb{Q}^n$ be defined by an acyclic disjunctive formula Ψ over $(\mathbb{Q}, <)$. Let $f \in \{min, mi, mx\}$.*

1. *If R is preserved under f, then for every two disjuncts $x_i < x_j$ and $x_{i'} < x_{j'}$ of Ψ we have $j = j'$.*
2. *If R is preserved under dual-f, then for every two disjuncts $x_i < x_j$ and $x_{i'} < x_{j'}$ of Ψ we have $i = i'$.*

Proof. We will only consider the case that R is preserved under f. The dual of f can be dealt with similarly.

Let R be preserved under f, and let $x_i < x_j$ and $x_{i'} < x_{j'}$ be disjuncts of Ψ. For the sake of contradiction, assume $j \neq j'$. Without loss of generality, we assume that $j < j'$. We are going to construct tuples $t_1, t_2 \in R$ such that $t_3 = f(t_1, t_2) \notin R$.

Since Ψ is acyclic, we can assume that the variables $x_1, \ldots, x_n$ are topologically sorted, i.e., if $x_p < x_q$ is an atom of Ψ, then $p < q$. In particular, $i < j$ and $i' < j'$. By the topological ordering, any tuple $t \in \mathbb{Q}^n$ with $t[i] < t[j]$ or $t[i'] < t[j']$ belongs to R, whereas no tuple $t \in \mathbb{Q}^n$ with $t[1] \geq \cdots \geq t[n]$ can belong to R. We will use these properties to obtain the desired tuples t_1 and t_2.

We distinguish the following three cases:

CASE 1 ($i' \leq i$): In this case, we have $i' \leq i < j < j'$. Let $a_i, a_{i'}, a_j, a_{j'} \in \mathbb{Q}$ and $b_i, b_{i'}, b_j, b_{j'} \in \mathbb{Q}$ be defined by $a_i = a_{i'} = b_i = b_{i'} = 2$, $b_j = 1$, $a_{j'} = 0$, and $a_j = b_{j'} = 3$; see Figure 2 for an illustration. We then have $a_i < a_j$ and $b_{i'} < b_{j'}$. It is also straightforward to verify that $f(a_i, b_i) = f(a_{i'}, b_{i'}) > f(a_j, b_j) > f(a_{j'}, b_{j'})$. Indeed, $min(a_i, b_i) = min(a_{i'}, b_{i'}) = 2$, $min(a_j, b_j) = 1$, and $min(a_{j'}, b_{j'}) = 0$, so the claim is true for $f = min$. For mi and mx, the claim is true, since $min(x, y) > min(x', y')$ implies $mi(x, y) > mi(x', y')$ and $mx(x, y) > mx(x', y')$. Now, Lemma C.2 implies that there are tuples $t_1, t_2 \in \mathbb{Q}^n$ such that $t_1[i] < t_1[j]$, $t_2[i'] < t_2[j']$, and $f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n])$. Hence, $t_1, t_2 \in R$ and $t_3 = f(t_1, t_2) \notin R$.

a_i	$a_{i'}$	a_j	$a_{j'}$
2	2	3	0
2	2	1	3
b_i	$b_{i'}$	b_j	$b_{j'}$

Figure 2: Choice of $a_i, a_{i'}, a_j, a_{j'} \in \mathbb{Q}$ and $b_i, b_{i'}, b_j, b_{j'} \in \mathbb{Q}$ in Case 1.

CASE 2 ($i < i' < j$): In this case, we have $i < i' < j < j'$. Let $a_i, a_{i'}, a_j, a_{j'} \in \mathbb{Q}$ and $b_i, b_{i'}, b_j, b_{j'} \in \mathbb{Q}$ be defined by $b_i = 3$, $a_{i'} = 2$, $b_j = 1$, $a_{j'} = 0$, $a_i = b_{i'} = 4$, and $a_j = b_{j'} = 5$; see Figure 3 for an illustration. We then have $a_i < a_j$ and $b_{i'} < b_{j'}$. It is straightforward to verify that $f(a_i, b_i) > f(a_{i'}, b_{i'}) > f(a_j, b_j) > f(a_{j'}, b_{j'})$. Indeed, $min(a_i, b_i) = 3$, $min(a_{i'}, b_{i'}) = 2$, $min(a_j, b_j) = 1$, and $min(a_{j'}, b_{j'}) = 0$, so the claim is true for $f = min$. For mi and mx, the claim is true, since $min(x, y) > min(x', y')$ implies $mi(x, y) > mi(x', y')$ and $mx(x, y) > mx(x', y')$. Now, Lemma C.2 implies that there are tuples $t_1, t_2 \in \mathbb{Q}^n$ such that $t_1[i] < t_1[j]$, $t_2[i'] < t_2[j']$, and $f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n])$. Hence, $t_1, t_2 \in R$ and $t_3 = f(t_1, t_2) \notin R$.

a_i	$a_{i'}$	a_j	$a_{j'}$
4	2	5	0
3	4	1	5
b_i	$b_{i'}$	b_j	$b_{j'}$

Figure 3: Choice of $a_i, a_{i'}, a_j, a_{j'} \in \mathbb{Q}$ and $b_i, b_{i'}, b_j, b_{j'} \in \mathbb{Q}$ in Case 2.

CASE 3 ($j \leq i'$): In this case, we have $i < j \leq i' < j'$. Let $a_i, a_j, a_{i'}, a_{j'} \in \mathbb{Q}$ and $b_i, b_j, b_{i'}, b_{j'} \in \mathbb{Q}$ be defined by $a_i = 2$, $b_j = b_{i'} = 1$, $a_{j'} = 0$, and $b_i = a_j = a_{i'} = b_{j'} = 3$; see Figure 4 for an illustration. We then have $a_i < a_j$ and $b_{i'} < b_{j'}$. It is also straightforward to verify that $f(a_i, b_i) > f(a_j, b_j) > f(a_{i'}, b_{i'}) > f(a_{j'}, b_{j'})$. Indeed, $min(a_i, b_i) = 2$, $min(a_j, b_j) = 1$, and $min(a_{i'}, b_{i'}) = min(a_{j'}, b_{j'}) = 0$, so the claim is true for $f = min$. For mi and mx, the claim is true, since $min(x, y) > min(x', y')$ implies $mi(x, y) > mi(x', y')$ and $mx(x, y) > mx(x', y')$. Now, Lemma C.2 implies that there are tuples $t_1, t_2 \in \mathbb{Q}^n$ such that $t_1[i] < t_1[j]$, $t_2[i'] < t_2[j']$, and $f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n])$. Hence, $t_1, t_2 \in R$ and $t_3 = f(t_1, t_2) \notin R$.

a_i	a_j	$a_{i'}$	$a_{j'}$
2	3	3	0
3	1	1	3
b_i	b_j	$b_{i'}$	$b_{j'}$

Figure 4: Choice of $a_i, a_j, a_{i'}, a_{j'} \in \mathbb{Q}$ and $b_i, b_j, b_{i'}, b_{j'} \in \mathbb{Q}$ in Case 3.

and $b_{i'} < b_{j'}$. It is also straightforward to verify that $f(a_i, b_i) > f(a_j, b_j) = f(a_{i'}, b_{i'}) > f(a_{j'}, b_{j'})$. Indeed, $min(a_i, b_i) = 2$, $min(a_j, b_j) = min(a_{i'}, b_{i'}) = 1$, and $min(a_{j'}, b_{j'}) = 0$, so the claim is true for $f = min$. For mi and mx, the claim is true, since $min(x, y) > min(x', y')$ implies $mi(x, y) > mi(x', y')$ and $mx(x, y) > mx(x', y')$. Now, Lemma C.2 implies that there are tuples $t_1, t_2 \in \mathbb{Q}^n$ such that $t_1[i] < t_1[j]$, $t_2[i'] < t_2[j']$, and $f(t_1[1], t_2[1]) \geq \cdots \geq f(t_1[n], t_2[n])$. Hence, $t_1, t_2 \in R$ and $t_3 = f(t_1, t_2) \notin R$.

Altogether, this concludes the proof. □

The following is the main lemma of this section.

Lemma 5.4 (restated) Let $R \subseteq \mathbb{Q}^n$ be defined by an acyclic disjunctive formula Ψ over $(\mathbb{Q}, <)$. If R is preserved by a function in $\mathcal{F} \cup$ *dual-*$\mathcal{F}$, then Ψ has the form $\bigvee_{i=1}^{k} x_i < x_0$ if $f \in \mathcal{F}$, and $\bigvee_{i=1}^{k} x_0 < x_i$ if $f \in$ *dual-*$\mathcal{F}$.

Proof. Let $\Psi = \bigvee_{1 \leq i \leq k} y_{s_i} < y_{t_i}$. Without loss of generality, we assume that $s_i \neq t_i$ for all $i \in \{1, \ldots, k\}$, and that any two pairs $(s_p, t_p), (s_q, t_q)$ with $p \neq q$ are distinct. If $k \leq 1$, then Ψ already has the required form. It remains to consider the case that $k \geq 2$. Note that in this case f cannot be a constant function. Furthermore, if f is *ll* or *dual-ll*, then the lemma follows from (Bodirsky and Kára 2010b). In what follows, we therefore assume that $k \geq 2$ and that f is one of min, mi, mx, and their duals.

We distinguish the following two cases:

Case 1: f is min, mi, or mx. In this case, Lemma C.3 implies that for each $j \in \{2, \ldots, k\}$ we have $t_1 = t_j$. In particular, Ψ has the form $\bigvee_{i=1}^{k} x_i < x_0$.

Case 2: f is the dual of min, mi, or mx. In this case, Lemma C.3 implies that for each $j \in \{2, \ldots, k\}$ we have $s_1 = s_j$. This implies that Ψ has the form $\bigvee_{i=1}^{k} x_0 < x_i$.

Altogether, this concludes the proof of the lemma. □

PTime Completeness

Theorem 5.6 (restated). *There is a rooted OMQ Q over $(\mathbb{Q}, \leq)$ with $\mathsf{dtype}(Q) = (\emptyset, \{x \leq y \wedge x \leq z\})$ such that evaluating Q is PTime-complete.*

Proof. By Part 2 of Theorem 4.5 it suffices to show that $\mathsf{CSP}_c(\mathbb{Q}, R_{\neg\psi})$ is PTime-hard, where $\psi = x \leq y \wedge x \leq z$. To this end we show that the alternating reachability problem (Tutte 1982; Immerman 1999) is polynomially reducible to $\mathsf{CSP}_c(\mathbb{Q}, R_{\neg\psi})$. An *alternating graph* is a directed graph $G = (V, E)$ where V is the disjoint union of the set $V_\exists$ of *existential* vertices and the set $V_\forall$ of *universal* vertices. An *alternating path* from vertex x to vertex y in G exists, in short, $\text{apath}_G(x, y)$ holds, if

1. $x = y$, or
2. $x \in V_\exists$ and there is a $z \in V$ with $(x, z) \in E$ and $\text{apath}_G(z, y)$ holds, or
3. $x \in V_\forall$ and for all $z \in V$ with $(x, z) \in E$, $\text{apath}_G(z, y)$ holds.

Alternating reachability is the problem to decide, given an alternating graph G and designated vertices s, t, whether $\text{apath}_G(s, t)$ holds. Alternating reachability is still PTime-hard if we assume that G is acyclic, that all vertices have out-degree either 0 or 2, that no universal vertex has outdegree 0, that s is existential and has no incoming edge, and that t is universal and has no outgoing edge. Assume G and vertices s, t with these properties are given. We regard the set V of vertices of G as variables, take the constants $0, 1 \in \mathbb{Q}$ and construct a PP sentence $\varphi_{G,s,t}$ with constants over $(\mathbb{Q}, R_{\neg\psi})$ as the conjunction of the following formulas:

- all $R_{\neg\psi}(v, w, w')$ such that $v \in V_\exists$ and w, w' are both successors of v;
- all $R_{\neg\psi}(v, w, w)$ such that $v \in V_\forall$ and w is a successor of v,
- all $R_{\neg\psi}(v, 0, 0)$ such that $v \neq t$ has outdegree 0, and
- $R_{\neg\psi}(0, s, s)$.

Recall that $R_{\neg\psi} = \{(a, b, c) \in \mathbb{Q}^3 \mid a < b \vee a < c\}$. Then one can easily show that $\text{apath}_G(s, t)$ holds iff $(\mathbb{Q}, R_{\neg\psi}) \models \varphi_{G,s,t}$, as required. □