Flag & Check: Data Access with Monadically Defined Queries (Extended Technical Report)

Technical Report 3030, Institute AIFB, Karlsruhe Institute of Technology

Sebastian Rudolph
Institute AIFB
Karlsruhe Institute of Technology, DE
sebastian.rudolph@kit.edu

Markus Krötzsch
Department of Computer Science
University of Oxford, UK
markus.kroetzsch@cs.ox.ac.uk

ABSTRACT

We introduce *monadically defined queries* (MODEQs) and *nested monadically defined queries* (NEMODEQs), two querying formalisms that extend conjunctive queries, conjunctive two-way regular path queries, and monadic Datalog queries. Both can be expressed as Datalog queries and in monadic second-order logic, yet they have a decidable query containment problem and favorable query answering complexities: a data complexity of P, and a combined complexity of NP (MODEQs) and PSpace (NEMODEQs).

Moreover, (NE)MODEQ answering remains decidable in the presence of a generic class of tuple-generating dependencies. In addition, techniques to rewrite queries under dependencies into (NE)MODEQs are introduced. Rewriting can be applied partially, and (NE)MODEQ answering is still decidable if the non-rewritable part of the TGDs permits decidable (NE)MODEQ answering on other grounds.

1. INTRODUCTION

Query languages are fundamental to the design of database systems. A good query language should be able to express a wide range of common information needs, and allow queries to be answered efficiently with limited computational resources. Moreover, databases are often considered in combination with dependencies, e.g., in the form of *tuple-generating dependencies* (TGDs), which are also playing an important role in data exchange, information integration, and database integrity checking [1]. While query answering under dependencies is undecidable in general, there are many decidable cases, and a query language should be robustly applicable to such extensions [39]. Another important task in database management and optimization is to check *query containment*, that is, to determine whether the answers of one query are contained in the answers of another query over arbitrary databases, possibly under the additional assumption that certain constraints are satisfied. A query language should therefore allow for such checks.

Unfortunately, these basic requirements are in conflict. Very simple query languages like *conjunctive queries* (CQs, [17]) allow for efficient query answering (NP combined/AC_0 data[1]) and containment checking, but have very limited expressivity. First-order logic (FOL) queries extend expressivity, but are still restricted to "local" queries, excluding, e.g., the transitive closure of a relation. Query containment is undecidable for FOL, and query answering becomes PSpace-complete for combined complexity [43], but remains AC_0 for data [31]. Another extension of CQs is *Datalog*, which introduces rule-based recursion. The price are higher complexities (ExpTime combined/P data), and undecidability of query containment [23, 42]. FOL and Datalog are incomparable; both are subsumed by second-order logic (SO), which is more expressive but also more complex (ExpSpace combined/PH data) [31, 44].

To find more tractable query languages, various smaller fragments of Datalog have been considered. *Linear Datalog* allows only one inferred predicate per rule body, which significantly reduces query complexity (PSpace combined/NLogSpace data) [29]. However, query containment remains undecidable. Two query languages for which containment is decidable are *monadic Datalog* and *conjunctive 2-way regular path queries* (C2RPQs) [27, 15]. The query complexity of C2RPQs (NP combined/NLogSpace data) is slightly lower than that of monadic Datalog (NP combined/P data), but the expressivity of the languages is incomparable. In particular, monadic Datalog cannot express transitive closure. Two well-known query languages subsuming monadic Datalog and C2RPQs are Datalog and *monadic second-order logic* (MSO) [36]. Query containment is decidable for neither of these. Both languages are incomparable, even regarding query complexities (MSO has PSpace combined/PH data [43, 45, 36]), their common upper bound being SO.

This reveals a glaring gap in the landscape of known query languages: no formalism that captures monadic Datalog and C2RPQs ensures tractable data complexity and decidable query containment. To address this, we propose *monadically defined queries* (MODEQs) and *nested monadically defined queries* NEMODEQs as novel query formalisms that combine these desirable properties. Their relationship to the aforementioned languages is also illustrated in Fig. 1.

The contribution of this paper can be split in two parts. In the first part, we introduce the new querying formalism

[1]As usual, query complexities refer to the problem of deciding whether a query has a particular certain answer. *Combined complexity* is the complexity for arbitrary queries and databases; *data complexity* is the complexity if the query is fixed or bounded.

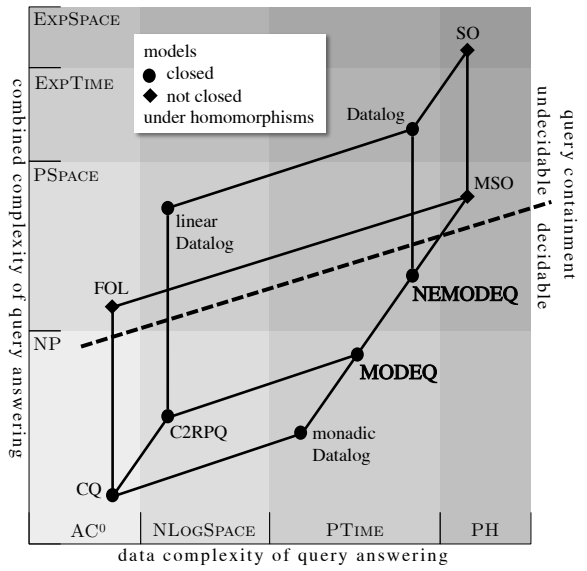

Figure 1: Overview of complexities and relations of monadically defined queries to other query formalisms as established in this paper; information indicated for conjunctive queries (CQ) and conjunctive 2-way regular path queries (C2RPQ) also hold when allowing unions (UCQ and UC2RPQ); all other formalisms are closed under unions

and clarify its relations to established query notions and the complexity for query answering. More precisely:

- We define MODEQs, discuss the underlying intuition, and provide examples demonstrating their expressivity.
- We show that MODEQs capture (unions of) CQs as well as C2RPQs and monadic Datalog queries.
- We prove MODEQ answering to be NP-complete for combined complexity (i.e., on par with CQs) and P-complete for data complexity, ensuring data-tractability as one of the central desiderata for querying large data sets.

We then extend MODEQs with nested subqueries, leading to a broader class of NEMODEQs that generalize MODEQs.

- We show that NEMODEQs (thus MODEQs) are expressible by both Datalog queries and MSO formulae.
- We prove NEMODEQ answering to be PSpace-complete for combined complexity (on par with FOL-based query languages like SQL) and P-complete for data complexity.
- We show that, unlike for Datalog and MSO queries, query containment for (NE)MODEQs is decidable.

In the second part of the paper, we study (NE)MODEQs in the context of dependencies and ontology-based data access. To this end, an important tool are (finite or infinite) *universal models*, which represent solutions to data exchange and constraint repair problems in the presence of TGDs [24]. As models of (NE)MODEQ are closed under homomorphisms, we find that universal models can be used to answer such queries, making them very robust to a wide class of TGDs.

- We immediately obtain decidability of query answering under all TGDs that admit a finite universal model. This property of TGDs is undecidable in general, but can be approximated by various notions of *acyclicity* [25, 26, 24, 38, 30, 34].
- More generally, we show that MODEQ answering is decidable in the presence of rules giving rise to (possibly infinite) universal models of *bounded treewidth*. This applies to many lightweight ontology languages as well as guarded TGDs and generalizations thereof [9, 4, 34, 5].
- In analogy to the popular notion of first-order rewritability, we introduce (NE)MODEQ rewritability, and we identify basic criteria for rewriting Datalog rules.
- Finally, we show that query answering is decidable under any set of TGDs that can be decomposed into one that is (NE)MODEQ-rewritable and one with the bounded-treewidth-model property.

Proofs omitted in the main paper are given in the Appendix.

2. PRELIMINARIES

We consider a standard language of first-order predicate logic, based on a finite set $\mathbf{C}$ of *constant symbols*, a finite set $\mathbf{P}$ of *predicate symbols*, and an infinite set $\mathbf{V}$ of first-order *variables*. Each predicate $p \in \mathbf{P}$ is associated with a natural number $\mathsf{ar}(p)$ called the *arity* of p. The list of predicates and constants forms the language's *signature* (or *schema*) $\mathscr{S} = \langle \mathbf{P}, \mathbf{C} \rangle$. We generally assume $\mathscr{S} = \langle \mathbf{P}, \mathbf{C} \rangle$ to be fixed, and only refer to it explicitly if needed.

Databases, Rules, and Queries. A *term* is a variable $x \in \mathbf{V}$ or a constant $c \in \mathbf{C}$. We use symbols s, t to denote terms, x, y, z, v, w to denote variables, a, b, c to denote constants. Expressions like $\boldsymbol{t}, \boldsymbol{x}, \boldsymbol{c}$ denote finite lists of such entities. We use the standard predicate logic definitions of *atom* and *formula*, using symbols φ, ψ for the latter. A formula is *ground* if it contains no variables. A *database*, usually denoted by D, is a finite set of ground atoms. We write $\varphi[\boldsymbol{x}]$ to emphasize that a formula φ has free variables $\boldsymbol{x}$; we write $\varphi[\boldsymbol{c}/\boldsymbol{x}]$ for the formula obtained from φ by replacing each variable in $\boldsymbol{x}$ by the respective constant in $\boldsymbol{c}$ (both lists must have the same length). A formula without free variables is a *sentence*.

A *conjunctive query* (CQ) is a formula $Q[\boldsymbol{x}] = \exists \boldsymbol{y}.\psi[\boldsymbol{x}, \boldsymbol{y}]$ where $\psi[\boldsymbol{x}, \boldsymbol{y}]$ is a conjunction of atoms. A *tuple-generating dependency* (TGD) is a formula of the form $\forall \boldsymbol{x}, \boldsymbol{y}.\varphi[\boldsymbol{x}, \boldsymbol{y}] \rightarrow \exists \boldsymbol{z}.\psi[\boldsymbol{x}, \boldsymbol{z}]$ where φ and ψ are conjunctions of atoms, called the *body* and *head* of the TGD, respectively. TGDs never have free variables, so we usually omit the universal quantifier when writing them. We use the symbol Σ, possibly with subscripts, to denote sets of TGDs. A *Datalog rule* is a TGD without existentially quantified variables; sets of Datalog rules will be denoted by symbols $\mathbb{P}, \mathbb{R}, \mathbb{S}$.

We use the standard semantics of first-order logic (FOL). An *interpretation* $\mathcal{I}$ consists of a (possibly infinite) set $\Delta^{\mathcal{I}}$ called *domain* and a function $\cdot^{\mathcal{I}}$ that maps constants c to domain elements $c^{\mathcal{I}} \in \Delta^{\mathcal{I}}$ and predicate symbols p to relations $p^{\mathcal{I}} \subseteq (\Delta^{\mathcal{I}})^{\mathsf{ar}(p)}$, thereby $p^{\mathcal{I}}$ is called the *extension* of p. A *variable assignment* for $\mathcal{I}$ is a function $\mathcal{Z} : \mathbf{V} \rightarrow \Delta^{\mathcal{I}}$. Conditions for $\mathcal{I}$ and $\mathcal{Z}$ to satisfy a FOL formula φ (i.e., to be a *model* of φ, written $\mathcal{I}, \mathcal{Z} \models \varphi$) are defined as usual. If φ is a sentence, then $\mathcal{Z}$ is irrelevant for satisfaction and can be

omitted. An *answer* to a CQ $Q[\boldsymbol{x}]$ over a database D and set Σ of TGDs is a list of constants $\boldsymbol{c}$ for which $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$.

Given an interpretation $\mathcal{I}$ and a formula $\varphi[\boldsymbol{x}]$ with free variables $\boldsymbol{x} = \langle x_1, \ldots, x_m \rangle$, the *extension* of $\varphi[\boldsymbol{x}]$ is the subset of $(\Delta^{\mathcal{I}})^m$ containing all those tuples $\langle \delta_1, \ldots, \delta_m \rangle$ for which $\mathcal{I}, \{x_i \mapsto \delta_i \mid 1 \leq i \leq m\} \models \varphi[\boldsymbol{x}]$. Two formulae $\varphi[\boldsymbol{x}]$ and $\psi[\boldsymbol{x}]$ with the same free variables $\boldsymbol{x}$ are called *equivalent* if for every $\mathcal{I}$ their extensions coincide.

Homomorphisms and Universal Models. Given interpretations $\mathcal{I}, \mathcal{J}$, a *homomorphism* π from $\mathcal{I}$ to $\mathcal{J}$ is a function $\pi : \Delta^{\mathcal{I}} \to \Delta^{\mathcal{J}}$ such that: (i) for all constants c, we have $\pi(c^{\mathcal{I}}) = c^{\mathcal{J}}$, and (ii) for all predicate symbols p and list of domain elements $\boldsymbol{\delta}$, we have $\boldsymbol{\delta} \in p^{\mathcal{I}}$ implies $\pi(\boldsymbol{\delta}) \in p^{\mathcal{J}}$.

Finding query answers is facilitated in practice since one may focus on universal models. A *universal model* of a set of sentences Ψ is an interpretation $\mathcal{I}$ such that (i) $\mathcal{I} \models \Psi$, and (ii) for every interpretation $\mathcal{J}$ with $\mathcal{J} \models \Psi$, there is a homomorphism from $\mathcal{I}$ to $\mathcal{J}$. For TGDs (and for plain databases), there is always a universal model if there is any model at all. It can be defined by a (possibly infinite) construction process called the *chase* [24]. In particular, we let $\mathcal{I}(D \cup \Sigma)$ denote the universal model, for which every homomorphism into any other model of $D \cup \Sigma$ is injective. $\mathcal{I}(D \cup \Sigma)$ always exists for satisfiable $D \cup \Sigma$, and it is unique up to isomorphism.

For a wide range of queries, entailment of query answers can be reduced to model checking in the universal model:

Fact 1 (Entailment via model checking). *If $Q[\boldsymbol{x}]$ is a query for which the set of models of $\exists \boldsymbol{x}.Q[\boldsymbol{x}]$ is closed under homomorphisms, then, for every database D and set Σ of TGDs, $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$ if and only if either $D \cup \Sigma$ is inconsistent or $\mathcal{I}(D \cup \Sigma) \models Q[\boldsymbol{c}/\boldsymbol{x}]$.*

This applies to CQs and to all other query languages studied herein. The case $\Sigma = \emptyset$ shows that one can equivalently represent databases using models $\mathcal{I}(D)$ instead of sets of facts D. Our perspective is more natural when using TGDs.

3. MONADICALLY DEFINED QUERIES

We now introduce a new query formalism, called *monadically defined queries (MODEQs)*, and state complexity results on query answering in this language. To deepen our understanding for the expressivity of MODEQs, we show that they strictly generalize the well-known query formalisms of *conjunctive 2-way regular path queries* (Section 3.1) and *monadic Datalog queries* (Section 3.2).

The heart of our query formalism is a mechanism for defining new predicates based on existing ones. This mechanism – which we refer to as *"flag & check"* – specifies new predicates by providing a procedure for testing if a particular tuple $\boldsymbol{\delta}=\langle \delta_1, \ldots, \delta_m \rangle \in (\Delta^{\mathcal{I}})^m$ is in the predicate's extension or not. To this end, the candidate tuple is first "flagged" by associating each δ_i with an auxiliary constant name λ_i that represents this element. The "check" is performed by running a Datalog program with this fixed interpretation of the constants λ_i. The check succeeds if a special fact `hit` is derived.

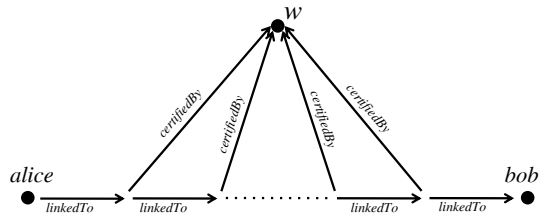

Figure 2: Structure recognized by MODEQ $\mathfrak{Q}_2$ in Example 2

Example 1. *To illustrate the idea, we consider a typical transitive closure query. Suppose that the binary predicate certifiedBy represents the direct certification of one entity by another, e.g., in a security application. We are interested in certification chains, which could be expressed in Datalog as follows:*

$$certifiedBy(x, y) \rightarrow certChain(x, y) \quad (1)$$
$$certChain(x, y) \wedge certifiedBy(y, z) \rightarrow certChain(x, z) \quad (2)$$

Corresponding Datalog rules $\mathbb{P}_1$ for "flag & check" are:

$$certifiedBy(\lambda_1, y) \rightarrow \mathtt{U}_1(y) \quad (3)$$
$$\mathtt{U}_1(y) \wedge certifiedBy(y, z) \rightarrow \mathtt{U}_1(z) \quad (4)$$
$$\mathtt{U}_1(\lambda_2) \rightarrow \mathtt{hit} \quad (5)$$

We define certChain to contain all pairs $\langle \delta_1, \delta_2 \rangle$ for which $\mathbb{P}_1$ entails `hit` *when interpreting λ_1 as δ_1 and λ_2 as δ_2.*

As in Example 1, the Datalog rules that we consider for the checking phase only use `hit` or new, unary predicates $\mathtt{U}_i$ in rule heads. Such unary predicates can be imagined as "colors" that are assigned to elements of the domain, and the check thus is a deterministic, recursive procedure of coloring the domain, starting from the flagged candidate elements. This idea is defined formally as follows.

Definition 1 (MODEQ syntax and semantics). *Given a signature $\mathscr{S}$, a* monadically defined predicate (MODEP) *of arity m is based on a signature $\mathscr{S}'$ that extends $\mathscr{S}$ with m fresh constant symbols $\lambda_1, \ldots, \lambda_m$, a fresh nullary predicate* `hit`*, and $k \geq 0$ fresh unary predicates $\mathtt{U}_1, \ldots, \mathtt{U}_k$. A MODEP is a set $\mathbb{P}$ of Datalog rules over $\mathscr{S}'$ where only $\mathtt{U}_1, \ldots, \mathtt{U}_k$, and* `hit` *occur in rule heads.*

Let $\mathcal{I}$ be an interpretation over $\mathscr{S}$. The extension $\mathbb{P}^{\mathcal{I}}$ of $\mathbb{P}$ is the set of all tuples $\langle \delta_1, \ldots, \delta_m \rangle \in (\Delta^{\mathcal{I}})^m$ for which $\mathcal{I}' \models \mathbb{P}$ implies $\mathcal{I}' \models$ `hit`*, for all interpretations $\mathcal{I}'$ that extend $\mathcal{I}$ to the symbols in $\mathscr{S}'$ such that $\langle \lambda_1^{\mathcal{I}'}, \ldots, \lambda_m^{\mathcal{I}'} \rangle = \langle \delta_1, \ldots, \delta_m \rangle$.*

A monadically defined query (MODEQ) *is a conjunctive query that uses both normal predicates and monadically defined predicates in its atoms. The semantics of MODEQs is defined in the obvious way via the semantics of MODEPs.*

Example 2. *The rules* (3)–(5) *define a binary MODEP $\mathbb{P}_1$, so the query of Example 1 can be written as a MODEQ $\mathfrak{Q}_1[v, w] = \mathbb{P}_1(v, w)$. For another example, assume there are entities certifying the security of message handling in certain nodes of a network. We are interested in entities y that can (directly) certify secure treatment of the message at all nodes*

on some path from Alice to Bob, as illustrated in Fig. 2. This is expressed by the MODEQ $\mathfrak{Q}_2[w] = \mathbb{P}_2(w)$, where $\mathbb{P}_2$ consists of the following rules:

$$\mathit{certifiedBy}(x, \lambda_1) \to \mathtt{U}_3(x) \quad (6)$$
$$\mathit{linkedTo}(\mathit{alice}, x) \to \mathtt{U}_2(x) \quad (7)$$
$$\mathtt{U}_2(x) \wedge \mathtt{U}_3(x) \wedge \mathit{linkedTo}(x, x') \to \mathtt{U}_2(x') \quad (8)$$
$$\mathtt{U}_2(\mathit{bob}) \to \mathtt{hit} \quad (9)$$

This new query notion is rather powerful. It is easy to see that it subsumes conjunctive queries (CQs) as well as unions of CQs. Indeed, given k CQs $\exists \boldsymbol{y}_i.Q_i[\boldsymbol{x}, \boldsymbol{y}_i]$ with $i \in \{1, \dots, k\}$, their union is expressed as the MODEQ $\{Q_i[\boldsymbol{\lambda}, \boldsymbol{y}_i] \to \mathtt{hit} \mid 1 \leq i \leq k\}(\boldsymbol{x})$. Before showing that MODEQs also capture more powerful query languages, we observe closure under homomorphism (see Fact 1) and state a basic complexity result.

Theorem 2. *For every MODEQ $\mathfrak{Q}[\boldsymbol{x}]$, the set of models of $\exists \boldsymbol{x}.\mathfrak{Q}[\boldsymbol{x}]$ is closed under homomorphism.*

Theorem 3 (MODEQ answering complexity). *Testing if $\boldsymbol{c}$ is an answer to a MODEQ $\mathfrak{Q}[\boldsymbol{x}]$ over a database D is* P*-complete in the size of D, and* NP*-complete in the combined size of D and $\mathfrak{Q}$.*

Hardness follows from the fact that MODEQs subsume monadic Datalog, shown in Section 3.2 below. P membership for data complexity is a consequence of the fact that Datalog subsumes MODEQs, demonstrated in Section 4.1. Membership in NP for combined complexity is established directly by showing that every query match is witnessed by a proof of polynomial size that can be guessed and verified in polynomial time.

3.1 MODEQs Capture Regular Path Queries

We now show that MODEQs subsume *conjunctive two-way regular path queries (C2RPQs)*, which generalize CQs by regular expressions over binary predicates [27, 15]. Variants of this type of queries are used, e.g., by the XPath query language for querying semi-structured XML data. Recent versions of the SPARQL 1.1 query language for RDF also support some of regular expressions that can be evaluated under a similar semantics.

C2RPQs are defined like MODEQs, but with MODEPs replaced by another form of defined predicates based on regular expressions over binary predicates and their inverses:

Definition 2 (C2RPQ syntax and semantics). *A* conjunctive two-way regular path predicate (C2RPP) *is a regular expression over the alphabet $\Gamma = \{p, p^- \mid \mathsf{ar}(p) = 2\}$ of normal and inverse binary predicate symbols. All C2RPPs are of arity 2. Consider an interpretation $\mathcal{I}$. For inverse predicates p^-, we define $(p^-)^{\mathcal{I}} := \{\langle \delta_2, \delta_1 \rangle \mid \langle \delta_1, \delta_2 \rangle \in p^{\mathcal{I}}\}$. For a C2RPP P, we set $\langle \delta, \delta' \rangle \in P^{\mathcal{I}}$ if there is a word $\gamma_1 \dots \gamma_n$ matching the regular expression P, and a sequence $\delta_0 \dots \delta_n$ of domain elements such that $\delta_0 = \delta$, $\delta_n = \delta'$, and $\langle \delta_i, \delta_{i+1} \rangle \in \gamma_i^{\mathcal{I}}$ for every $i \in \{0, \dots, n-1\}$.*

A conjunctive two-way regular path query (C2RPQ) *is a conjunctive query that uses both normal predicates and C2RPPs in its atoms. The semantics of C2RPQs is defined in the obvious way based on the semantics of C2RPPs.*

Example 3. *The query of Example 1 is expressed by the C2RPQ $\mathit{certifiedBy}^*(x, y)$. Another C2RPQ with inverses is*

$$\mathit{mountain}(x) \wedge \mathit{continent}(y) \wedge (\mathit{locatedIn} | \mathit{hasPart}^-)^*(x, y). \quad (10)$$

Query answering for C2RPQs is NP-complete regarding the size of the database and query, which is the same as for CQs. In terms of data complexity, C2RPQs are NLogSpace-complete, and thus harder than CQs (AC_0). One can show hardness via graph reachability, and membership via a translation to linear Datalog [13].

Definition 3 (C2RPQ to MODEQ translation). *Consider a C2RPP P and a finite automaton $\mathcal{A}_P = \langle \Gamma, S, I, F, T \rangle$ that recognizes P. The binary MODEP* $\mathsf{modep}(P)$ *consists of the rules*

$$\to \mathtt{U}_s(\lambda_1) \quad \text{for every initial state } s \in I,$$
$$\mathtt{U}_s(z) \wedge p(z, z') \to \mathtt{U}_{s'}(z') \quad \text{for every transition } \langle s, p, s' \rangle \in T,$$
$$\mathtt{U}_s(z) \wedge p(z', z) \to \mathtt{U}_{s'}(z') \quad \text{for every transition } \langle s, p^-, s' \rangle \in T,$$
$$\mathtt{U}_s(\lambda_2) \to \mathtt{hit} \quad \text{for every final state } s \in F.$$

Given a C2RPQ Q, a MODEQ $\mathsf{modeq}(Q)$ *is obtained by replacing every C2RPP P in Q by* $\mathsf{modep}(P)$.

The intuition behind the translation of C2RPQs to MODEQs is to find bindings for x and y in $P(x, y)$ by simulating all possible runs of the automaton corresponding to a C2RPQ. Colors $\mathtt{U}_s$ are associated to states s of the automaton to record which domain elements can be reached in which states when starting in an initial state at x. The success criterion is that y is colored by a final state. One can thus show the following:

Theorem 4 (MODEQs capture C2RPQs). *For every C2RPQ Q, the MODEQ* $\mathsf{modeq}(Q)$ *can be constructed in linear time, and is equivalent to Q. In particular, the answers for Q and* $\mathsf{modeq}(Q)$ *coincide.*

Example 4. *Let Q be the regular path query* (10)*. The query* $\mathsf{modeq}(Q)$ *is $\mathit{mountain}(x) \wedge \mathit{continent}(y) \wedge \mathbb{P}(x, y)$ where $\mathbb{P} = \mathsf{modep}((\mathit{locatedIn} | \mathit{hasPart}^-)^*)$ consists of the rules:*

$$\to \mathtt{U}(\lambda_1) \quad (11)$$
$$\mathtt{U}(z) \wedge \mathit{locatedIn}(z, z') \to \mathtt{U}(z') \quad (12)$$
$$\mathtt{U}(z) \wedge \mathit{hasPart}(z', z) \to \mathtt{U}(z') \quad (13)$$
$$\mathtt{U}(\lambda_2) \to \mathtt{hit}. \quad (14)$$

Here we only need one "color" $\mathtt{U}$ that is propagated over locatedIn and inversely over hasPart. The pairs in $\mathbb{P}$ are those for which this process, started at the first element, will eventually color the second argument.

However, the expressivity of MODEQs goes well beyond that of C2RPQs, even when considering only binary predicates. This follows from the easy observation that for every C2RPQ Q there is an integer n, such that whenever Q matches into a graph G, it also matches into a graph G' where all vertices have degree $\leq n$ and from which there is a homomorphism into G. It is easy to see that the MODEQ $\mathfrak{Q}_2$ from Example 2 does not have this property.

3.2 MODEQs Capture Monadic Datalog

Monadic Datalog queries are another type of query language that enjoys favorable computational properties. They are used, e.g., for information extraction from the Web [28]. We now show that MODEQs can express monadic Datalog queries, which is another way to see that they are strictly more general than C2RPQs.

DEFINITION 4 (MONADIC DATALOG QUERY). *Given a signature $\mathscr{S}$, a* Datalog query *is based on a signature $\mathscr{S}'$ that extends $\mathscr{S}$ with additional predicates, called* intensional database (IDB) predicates*. A Datalog query is a pair $\langle \mathtt{goal}, \mathbb{S} \rangle$, where* goal *is an IDB predicate, and $\mathbb{S}$ is a set of Datalog rules over $\mathscr{S}'$ where only IDB predicates occur in rule heads and* goal *does not occur in rule bodies. A* monadic Datalog query *is a query where all IDB predicates other than* goal *have arity* 1.

Given an interpretation $\mathcal{I}$, the extension *of $\langle \mathtt{goal}, \mathbb{S} \rangle$ is the set of tuples $\boldsymbol{\delta}$ over $\Delta^{\mathcal{I}}$ for which every extension $\mathcal{I}'$ of $\mathcal{I}$ to $\mathscr{S}'$ which satisfies $\mathbb{S}$ must also satisfy $\boldsymbol{\delta} \in \mathtt{goal}^{\mathcal{I}'}$.*

Note that sometimes in the literature, the arity of goal is restricted to 1. Allowing it to be arbitrary does not affect the complexity of the formalism.

DEFINITION 5 (MONADIC DATALOG TO MODEQ). *Given a monadic Datalog query $Q = \langle \mathtt{goal}, \mathbb{S} \rangle$, we let* modeq$(Q)$ *denote the MODEQ $\mathbb{P}(\boldsymbol{x})$ where $\mathbb{P}$ is obtained from $\mathbb{S}$ by*

- *replacing each rule $\varphi[\boldsymbol{x}, \boldsymbol{y}] \rightarrow \mathtt{goal}(\boldsymbol{x})$ by $\varphi[\boldsymbol{\lambda}, \boldsymbol{y}] \rightarrow \mathtt{hit}$,*
- *replacing each IDB predicate, uniformly and injectively, by a predicate $\mathtt{U}_j$.*

THEOREM 5 (MODEQS CAPTURE MONADIC DATALOG). *For every monadic Datalog query Q, the MODEQ* modeq(Q) *can be constructed in linear time, and is equivalent to Q. In particular, the answers for Q and* modeq(Q) *coincide.*

From the correspondence thus established it follows that the lower complexity bounds of monadic Datalog carry over and MODEQ answering on databases must thus be P-hard for data complexity and NP-hard for combined complexity [28], showing one direction of Theorem 3. MODEQs are strictly more expressive than monadic Datalog queries, shown by the fact that even a simple connectedness query like the one in Example 1 cannot be expressed in monadic Datalog.

We would like to specifically note that, although the rules used in the definitions of MODEPs are in fact monadic Datalog rules, the query evaluation schemes underlying monadic Datalog and MODEQs are fundamentally different: while in the case of monadic Datalog, all elements of the extension can be obtained at once by a forward chaining saturation process on the given interpretation (or database), the *flag & check* strategy that underlies the semantics definition of MODEPs crucially hinges on each potential extension element being verified in a *separate* saturation process. In Section 4.1, we will see that this idea can be captured by Datalog queries, but not by monadic ones.

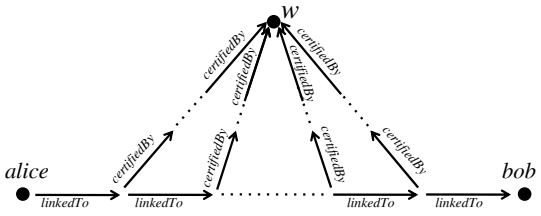

Figure 3: **Structure recognized by $\mathfrak{Q}_3$ in Example 5**

4. NESTED MODEQS

Query nesting is the process of using an n-ary subquery instead of an n-ary predicate symbol within a query, with the obvious semantics. In this section, we use this mechanism to extend MODEQs, leading to the more general language of *nested monadically defined queries (NEMODEQs)*. We then show that queries of this type can be expressed in Datalog (Section 4.1) and monadic second-order logic (Section 4.2). These results extend to MODEQs as a special case of NEMODEPs, and help to establish some additional upper bounds for complexity.

It is interesting to ask if nesting of queries actually leads to a new query language or not. A query language is *closed under nesting* if every query with nested subqueries can be expressed by some non-nested query. Many query languages are trivially closed under nestings as they allow nesting as part of their syntax (e.g., CQs, FOL, MSO, and SO), other languages require more or less complex reformulations to eliminate nested queries (e.g., UCQs, Datalog, and monadic Datalog), others are not closed under nestings (e.g., linear Datalog). Example 6 below shows that MODEQs are not closed under nestings, motivating the following definition.

DEFINITION 6 (NEMODEQ SYNTAX & SEMANTICS). *Let $\mathscr{S}$ be the underlying signature. A* nested monadically defined predicate (NEMODEP) *of degree* 1 *over $\mathscr{S}$ is a MODEP over $\mathscr{S}$. Consider a finite set $\mathbb{P}_i$ ($i = 1, \ldots, k$) of NEMODEPs of degree $\leq d$ over $\mathscr{S}$. A NEMODEP of degree $d+1$ is a MODEP over a signature $\mathscr{S}'$ that extends $\mathscr{S}$ with additional predicate names $\mathbb{P}_i$ that have the same arity as the respective queries.*

The semantics of NEMODEPs of degree $d > 1$ is defined as for MODEPs, based on the (recursively defined) semantics of NEMODEPs of degree $< d$. A nested monadically defined query (NEMODEQ) *is a conjunctive query that uses both normal predicates and nested monadically defined predicates in its atoms. The semantics of NEMODEQs is defined in the obvious way.*

Note that the auxiliary symbols λ_i, $\mathtt{U}_j$, and hit do not need to be distinct in different subqueries, or in queries and their subqueries. This does not cause semantic interactions.

EXAMPLE 5. *Consider again Example 2, and assume that we are now also interested in entities that can certify the security of a communication indirectly, i.e., through a chain of certifications, as shown in Fig. 3. This can be expressed by nesting MODEP $\mathbb{P}_1$ in $\mathfrak{Q}_2$, leading to a NEMODEQ of*

degree 2 $\mathfrak{Q}_3 = \mathbb{P}_3(w)$*, where* $\mathbb{P}_3$ *coincides with* $\mathbb{P}_2$ *except that rule* (6) *is replaced by* $\mathbb{P}_1(x, \lambda_1) \to \mathtt{U}_3(x)$.

However, even if a query is more easily expressed as a NEMODEQ, it might still be expressible as a MODEQ. The next example shows that this is the case for $\mathfrak{Q}_3$ above, and presents a query that cannot be expressed by any MODEQ.

Example 6. $\mathfrak{Q}_3$ *of Example 5 can equivalently be expressed by the MODEQ* $\mathfrak{Q}_4[w] = \mathbb{P}_4(w)$*, where* $\mathbb{P}_4$ *consists of the following rules:*

$$\mathit{certifiedBy}(x, \lambda_1) \to \mathtt{U}_3(x) \quad (15)$$
$$\mathtt{U}_3(y) \wedge \mathit{certifiedBy}(x, y) \to \mathtt{U}_3(x) \quad (16)$$
$$\mathit{linkedTo}(\mathit{alice}, x) \to \mathtt{U}_2(x) \quad (17)$$
$$\mathtt{U}_2(x) \wedge \mathtt{U}_3(x) \wedge \mathit{linkedTo}(x, x') \to \mathtt{U}_2(x') \quad (18)$$
$$\mathtt{U}_2(\mathit{bob}) \to \mathtt{hit} \quad (19)$$

To define a NEMODEQ that cannot be expressed as a MODEQ, we first modify this example to ask for all pairs of persons who are connected by a communication chain that is certified by a single entity, i.e., we query for the possible pairs of "alice" and "bob." Let $\mathbb{P}_5$ *be the ternary MODEP that consists of the following rules:*

$$\mathit{certifiedBy}(x, \lambda_1) \to \mathtt{U}_3(x) \quad (20)$$
$$\mathtt{U}_3(y) \wedge \mathit{certifiedBy}(x, y) \to \mathtt{U}_3(x) \quad (21)$$
$$\mathit{linkedTo}(\lambda_2, x) \to \mathtt{U}_2(x) \quad (22)$$
$$\mathtt{U}_2(x) \wedge \mathtt{U}_3(x) \wedge \mathit{linkedTo}(x, x') \to \mathtt{U}_2(x') \quad (23)$$
$$\mathtt{U}_2(\lambda_3) \to \mathtt{hit} \quad (24)$$

We now form a NEMODEQ that asks for all pairs of persons who can communicate through a chain of such secure channels that goes via multiple people. Moreover, we require that all of these people are "friends," that is, trustworthy in the context of the application. Let $\mathbb{P}_6$ *be the binary NEMODEP with the following rules:*

$$\to \mathtt{U}_1(\lambda_1) \quad (25)$$
$$\mathtt{U}_1(y) \wedge \mathbb{P}_5(x, y, z) \wedge \mathit{friend}(z) \to \mathtt{U}_1(z) \quad (26)$$
$$\mathtt{U}_1(y) \wedge \mathbb{P}_5(x, y, \lambda_2) \to \mathtt{hit} \quad (27)$$

The NEMODEQ $\mathfrak{Q}_4 = \mathbb{P}_6(v, w)$ *(see Fig. 4) cannot be expressed as a MODEQ. To show this, one assumes the existence of such a MODEQ and constructs a database where it must accept a match that is not accepted by* $\mathfrak{Q}_4$*. A formal proof is given in Proposition 22 in the Appendix.*

4.1 Expressing NEMODEQs in Datalog

We now show that NEMODEQs of arbitrary degree can be expressed as Datalog queries. To this end, the auxiliary predicates have to be "contextualized," which increases their arity. Hence the translation usually does not lead to monadic Datalog queries.

Definition 7 (NEMODEQ to Datalog). *Given a MODEP* $\mathbb{P}$ *of arity m, the set* $\mathsf{datalog}(\mathbb{P})$ *of Datalog rules over an extended signature contains, for each rule in* $\mathbb{P}$*, a new rule obtained by replacing*

- *each constant* λ_i *with a variable* x_{λ_i}*,*
- *each atom* $\mathtt{U}_i(z)$ *with the atom* $\hat{U}_i(z, x_{\lambda_1}, \ldots, x_{\lambda_m})$ *where* $\hat{U}_i$ *is a fresh predicate of arity* $m + 1$*,*
- *each atom* $\mathtt{hit}$ *with the atom* $p_{\mathbb{P}}(x_{\lambda_1}, \ldots, x_{\lambda_m})$ *where* $p_{\mathbb{P}}$ *is a fresh predicate symbol of arity m.*

For a NEMODEP $\mathbb{P}$ *of degree* $d > 1$*, let* $\mathbb{P}'$ *be the MODEP obtained by replacing each direct sub-NEMODEP* $\mathbb{Q}$ *of* $\mathbb{P}$ *with the predicate* $p_{\mathbb{Q}}$*. The Datalog translation of* $\mathbb{P}$ *is recursively defined as:*

$$\mathsf{datalog}(\mathbb{P}) := \mathsf{datalog}(\mathbb{P}') \cup \bigcup_{\mathbb{Q} \text{ a direct sub-NEMODEP of } \mathbb{P}} \mathsf{datalog}(\mathbb{Q}).$$

Given a NEMODEQ $\mathfrak{Q}[\boldsymbol{x}] = \exists \boldsymbol{y}.\varphi[\boldsymbol{x}, \boldsymbol{y}]$*, we define its translation* $\mathsf{datalog}(\mathfrak{Q})$ *as* $\mathsf{datalog}(\{\varphi[\boldsymbol{\lambda}, \boldsymbol{y}] \to \mathtt{hit}\})$*, where the predicate used to replace* $\mathtt{hit}$ *will be denoted by* $p_{\mathfrak{Q}}$*.*

Note that the predicates $\hat{U}_i$ must be globally fresh, even if multiple subqueries use the same $\mathtt{U}_i$. The rules in $\mathsf{datalog}(\mathfrak{Q})$ might be unsafe, i.e., they may contain universally quantified variables in the head that do not occur in the body. This is no problem with the logical semantics we consider.

Theorem 6 (Datalog expressibility of NEMODEQs). *For any NEMODEQ* $\mathfrak{Q}$*,* $\mathsf{datalog}(\mathfrak{Q})$ *can be constructed in linear time. Moreover, the queries* $\mathfrak{Q}[\boldsymbol{x}]$ *and* $\langle p_{\mathfrak{Q}}, \mathsf{datalog}(\mathfrak{Q})\rangle$ *are equivalent, i.e., their answers coincide.*

Example 7. *The Datalog translation for the MODEQ* $\mathfrak{Q}_3$ *from Example 5 is as follows:*

$\mathsf{datalog}(\mathfrak{Q}_3)$

$\mathsf{datalog}(\mathbb{P}_3)$

$\mathsf{datalog}(\mathbb{P}_1)$

$$\mathit{certifiedBy}(x_{\lambda_1}, y) \to \hat{U}_1(y, x_{\lambda_1}, x_{\lambda_2})$$
$$\hat{U}_1(y, x_{\lambda_1}, x_{\lambda_2}) \wedge \mathit{certifiedBy}(y, z) \to \hat{U}_1(z, x_{\lambda_1}, x_{\lambda_2})$$
$$\hat{U}_1(x_{\lambda_2}, x_{\lambda_1}, x_{\lambda_2}) \to p_{\mathbb{P}_1}(x_{\lambda_1}, x_{\lambda_2})$$

$$p_{\mathbb{P}_1}(x, x_{\lambda_1}) \to \hat{U}_3(x, x_{\lambda_1})$$
$$\mathit{linkedTo}(\mathit{alice}, x) \to \hat{U}_2(x, x_{\lambda_1})$$
$$\hat{U}_2(x, x_{\lambda_1}) \wedge \hat{U}_3(x, x_{\lambda_1}) \wedge \mathit{linkedTo}(x, x') \to \hat{U}_2(x', x_{\lambda_1})$$
$$\hat{U}_2(\mathit{bob}, x_{\lambda_1}) \to p_{\mathbb{P}_3}(x_{\lambda_1})$$

$$p_{\mathbb{P}_3}(x_{\lambda_1}) \to p_{\mathfrak{Q}_3}(x_{\lambda_1})$$

Using backward-chaining, the goal $p_{\mathfrak{Q}}(\boldsymbol{x})$ can be expanded under the rules $\mathsf{datalog}(\mathfrak{Q})$ to obtain a (possibly infinite) set of CQs that do not contain auxiliary predicates $p_{\mathfrak{Q}'}$. Thus $\mathfrak{Q}$ can be considered as a union of (possibly infinitely many) conjunctive queries.

The linear translation of NEMODEQs (and thus also MODEQs) to Datalog leads to various results. First, NEMODEQs inherit Datalog's PTime upper bound for data complexity of query answering [23]. The results of Section 3.2 show that this bound is tight. Second, we find that the models of NEMODEQs are closed under homomorphisms, since Datalog has this property. Again, this shows that query entailment coincides with model checking (Fact 1).

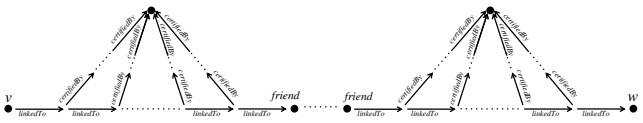

Figure 4: Structure recognized by $\mathfrak{Q}_4$ in Example 6

THEOREM 7. *For any NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$, the set of models of $\exists \boldsymbol{x}.\mathfrak{Q}$ is closed under homomorphisms.*

4.2 Expressing NEMODEQs in MSO Logic

In this section, we show that NEMODEQs can also be expressed in *monadic second-order logic (MSO)*, the extension of first-order logic with *set variables*, used like predicates of arity 1. To distinguish them from object variables x, y, z, we denote set variables by the uppercase letter U, possibly with subscripts, hinting at their close relation to the unary coloring predicates $\mathtt{U}$. We adhere to the standard semantics of MSO that we will not repeat here.

To simplify the presentation of the next definition, we henceforth assume that every variable x, constant λ_i, or monadic predicate $\mathtt{U}_j$ is used in at most one (sub-)predicate $\mathbb{P}$ of any NEMODEQ or NEMODEP we consider. This can always be achieved by renaming variables and predicates.

DEFINITION 8 (NEMODEQ TO MSO). *For a MODEP $\mathbb{P}$ of arity m with auxiliary unary predicates $\mathtt{U}_1, \ldots, \mathtt{U}_k$, and a list of terms $\boldsymbol{t} = \langle t_1, \ldots, t_m \rangle$, we define an MSO formula*

$$\mathsf{mso}(\mathbb{P}(\boldsymbol{t})) \coloneqq \forall U_1, \ldots, U_k. \neg \bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho, \boldsymbol{t})$$

where $\mathsf{mso}(\rho, \boldsymbol{t})$ is the rule obtained from rule ρ by replacing each occurrence of a constant λ_i by t_i, each occurrence of `hit` *by $\bot$ (the falsity atom), and each occurrence of a unary predicate $\mathtt{U}_i$ by a set variable U_i. We extend* mso *to NEMODEPs of higher degree by applying it recursively to NEMODEP atoms. For a NEMODEQ $\mathfrak{Q}$, we obtain $\mathsf{mso}(\mathfrak{Q})$ by replacing every NEMODEP atom $\mathbb{P}(\boldsymbol{t})$ in $\mathfrak{Q}$ by $\mathsf{mso}(\mathbb{P}(\boldsymbol{t}))$.*

By replacing `hit` with $\bot$, the derivation of a query match becomes the derivation of an inconsistency. The formula $\mathsf{mso}(\mathbb{P}(\boldsymbol{t}))$ evaluates to true if this occurs for all possible interpretations of the predicates $\mathtt{U}_j$, expressed here by universal quantification over the set variables U_j. The interpretation of $\boldsymbol{t}$ corresponds to the flagged tuple $\boldsymbol{\lambda}$ that is to be checked. It is thus easy to see that the translation captures the semantic conditions of Definitions 1 and 6.

THEOREM 8 (MSO EXPRESSIBILITY OF NEMODEQS). *For every NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$, $\mathsf{mso}(\mathfrak{Q}[\boldsymbol{x}])$ can be constructed in linear time and is equivalent to $\mathfrak{Q}[\boldsymbol{x}]$.*

EXAMPLE 8. *Consider NEMODEQ $\mathfrak{Q}_3$ from Example 5. Then $\mathsf{mso}(\mathfrak{Q}_3)$ is the MSO formula*

$$\forall U_2, U_3. \neg \begin{pmatrix} \forall v.\Big(\boxed{\forall U_1. \neg \begin{pmatrix} \forall y.(certBy(v,y) \rightarrow U_1(y)) \\ \wedge \forall y,z.(U_1(y) \wedge certBy(y,z) \rightarrow U_1(z)) \\ \wedge \quad (U_1(w) \rightarrow \bot) \end{pmatrix}} \rightarrow U_3(v) \Big) \\ \wedge \quad \forall x.(linkedTo(alice, x) \rightarrow U_2(x)) \\ \wedge \quad \forall x,x'.(U_2(x) \wedge U_3(x) \wedge linkedTo(x,x') \rightarrow U_2(x')) \\ \wedge \quad (U_2(bob) \rightarrow \bot) \end{pmatrix}$$

with free variable w (using certBy to abbreviate certifiedBy). The framed subformula is $\mathsf{mso}(\mathbb{P}_1(v, w))$.

Expressibility of NEMODEQs (and thus also MODEQs) in MSO is a useful feature, which we will further exploit below. For the moment, we just note the direct consequence that the PSPACE combined complexity of model checking in MSO directly gives us PSPACE-membership of query answering for NEMODEQs and MODEQs.

The following theorem closes the gap w.r.t. the combined complexity of query answering by showing PSPACE hardness for NEMODEQs by a reduction from the validity problem of quantified Boolean formulae.

THEOREM 9 (COMPLEXITY OF NEMODEQ ANSWERING). *Checking if $\boldsymbol{c}$ is an answer to a NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$ over a database D is* P*-complete in the size of D, and* PSPACE*-complete in the size of D and $\mathfrak{Q}$.*

Figure 1 gives an overview of the relationships established so far, regarding both expressivity and complexity. MODEQs feature the same complexities as monadic Datalog, while providing a significant extension of expressivity. The step to NEMODEQs leads to increased combined complexity. Nevertheless, combined complexity is still lower than for Datalog and data complexity is still lower than for MSO. In addition, in the next sections, we will show that NEMODEQs (and MODEQs) are also more well-behaved than these two when it comes to checking containment or interaction with rule sets that give rise to infinite structures.

5. DECIDING QUERY CONTAINMENT

Checking query containment is an essential task in database management, facilitating query optimization, information integration and exchange, and database integrity checking. The *containment* or *subsumption problem* of two queries $\mathfrak{P}$ and $\mathfrak{Q}$ is the question whether the answers of $\mathfrak{Q}$ are contained in the answers of $\mathfrak{P}$ over any database. In this section, we show that this problem is decidable for NEMODEQs. At its core, this result is based on previous work by Courcelle [20], from which we can derive the following general theorem, which is interesting in its own right:

THEOREM 10 (DECIDING DATALOG CONTAINMENT IN MSO). *Consider a Datalog query $\langle \mathtt{goal}, \mathbb{P} \rangle$ and an MSO query φ. If the models of φ are closed under homomorphism, then it is decidable if the query $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ.*

The underling result in [20] is formulated for queries without free variables and without constant symbols, using a variant of multi-sorted monadic second-order logic and a notion of graph grammar derived from Datalog queries. The appendix recalls the relevant notions and relates them to our setting to prove Theorem 10.

By Theorems 6, 7, and 8, NEMODEQs can be considered as Datalog queries and as MSO queries whose models are closed under homomorphisms. This shows the following:

THEOREM 11 (DECIDING NEMODEQ CONTAINMENT). *The query containment problem for NEMODEQs is decidable.*

The complexity of NEMODEQ containment remains to be determined. A lower bound is the 2EXPTIME-hardness of monadic Datalog containment shown recently [7].

6. QUERYING UNDER DEPENDENCIES

Dependencies play an important role in many database applications, be it to formulate constraints, to specify views for data integration, or to define relationships in data exchange. Dependencies can be viewed as logical implications, like the TGDs introduced in Section 2, and one is generally interested in answering queries w.r.t. to their logical entailments. Universal models (Section 2) can be viewed as solutions to data exchange problems or as minimal ways of repairing constraint violations over a database. Querying under dependencies thus corresponds to finding *certain answers*, as is common, e.g., in data integration scenarios.

EXAMPLE 9. *Consider the following set of TGDs:*

$$\begin{aligned} \mathit{hasAuthor}(x,y) &\rightarrow \mathit{publication}(x) \\ \mathit{cites}(x,y) &\rightarrow \mathit{publication}(x) \wedge \mathit{publication}(y) \\ \mathit{publication}(x) &\rightarrow \exists y.\mathit{hasAuthor}(x,y) \end{aligned}$$

For a database $\{\mathit{hasAuthor}(a,c), \mathit{cites}(a,b)\}$*, the conjunctive query* $\exists z.(\mathit{hasAuthor}(x,z))$ *has* $x \mapsto a$ *as its only answer. When taking the above dependencies into account, the certain answers additionally contain* $x \mapsto b$*.*

The extension of TGDs with equality is known as *embedded dependencies*, a special case of which are *equality-generating dependencies* [1]. We will not focus on equality here, and state most of our results for TGDs.

The problem of computing CQ answers under TGDs is undecidable in general [16, 6]. A practical approach for computing certain answers is to compute a finite universal model, if possible. All combinations of TGDs and databases have universal models, but it is undecidable whether any such model is finite; the *core chase* is a known semi-decision procedure that computes a finite universal model whenever it exists [24]. Many other variants of the chase have been proposed to compute (finite) universal models in certain cases.

One can compute certain answers under TGDs even in cases where no finite universal model exists, as long as there is a universal model that is sufficiently "regular" to allow for a finite representation. One of the most general criteria is based on the well-known notion of *treewidth* (see, e.g., [4] for a formal definition). A rule set Σ is called *bounded treewidth set (bts)* if for every database D, there is a universal model $\mathcal{I}(D \cup \Sigma)$ of bounded treewidth. The treewidth bound in each case can depend on Σ and D. Recognizing whether a rule set has this property is undecidable in general [4], but many sufficient conditions have been identified. This includes the case of rule sets with finite models as a special case of models with bounded treewidth. TGDs without existential quantifiers, called *full dependencies* or Datalog rules [1], trivially have finite models. More elaborate are various notions of *acyclicity* based on analyzing the interaction of TGDs [25, 26, 37, 38, 4, 34].

Cases where $\mathcal{I}(D \cup \Sigma)$ may be infinite but treewidth-bounded are also manifold. A basic case are *guarded TGDs*, inspired by the guarded fragment of first-order logic [2]. These have been generalized to *weakly guarded TGDs* [9] and *frontier-guarded rules* [4], both of which are subsumed by *weakly frontier-guarded TGDs* [4]. The most expressive currently known bts fragments are *greedy bts TGDs* [5] and *glut-guarded TGDs* [34].

Another type of dependencies comes from the area of *Description Logics* (DLs). Originally conceived as ontology languages, DLs have also been applied to express database constraints [14]. DLs use a different syntax, but share a similar first-order semantics that makes them compatible with TGDs. Most DLs considered in database applications are Horn logics (that allow for universal models), and can thus be presented as rules [10].

EXAMPLE 10. *The set of TGDs in Example 9 can equivalently be expressed by the DL-Lite ontology*

$$\begin{aligned} \exists \mathit{hasAuthor} &\sqsubseteq \mathit{publication} & \exists \mathit{cites} &\sqsubseteq \mathit{publication} \\ \mathit{publication} &\sqsubseteq \exists \mathit{hasAuthor} & \exists \mathit{cites}^- &\sqsubseteq \mathit{publication}. \end{aligned}$$

Many DLs enjoy tree-model properties, but some expressive DLs are not bounded treewidth. This mainly applies to DLs that support transitivity or its generalizations [35, 40].

As mentioned before, the entailment problem $D, \Sigma \models Q$ is known to be decidable if Σ is bts and Q is a conjunctive query. Our main result of this section is that this extends to NEMODEQs provided that the following holds.[2]

CONJECTURE 12. *Satisfiability of monadic second-order logic on countable interpretations of bounded treewidth is decidable.*

THEOREM 13 (NEMODEQ ANSWERING UNDER BTS). *Let D be a database and let Σ be a set of rules for which the treewidth of $\mathcal{I}(D \cup \Sigma)$ is bounded. Let $\mathfrak{Q}[\boldsymbol{x}]$ be a NEMODEQ.*

1. *$D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ if and only if $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ has no countable model with bounded treewidth.*
2. *If Conjecture 12 holds, then $D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ is decidable.*

The proof of this theorem exploits universality of $\mathcal{I}(D \cup \Sigma)$ and preservation of NEMODEQ matches under homomorphisms (Theorem 7). The second claim follows from the first by applying Conjecture 12 and the observation that the MSO theory $D \cup \Sigma \cup \{\neg\mathsf{mso}(\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}])\}$ is equivalent to $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ by Theorem 8.

7. MODEQ REWRITABILITY

Query rewriting is an important technique for answering queries under dependencies, and the main alternative to the chase. The general idea is to find "substitute queries" that can be evaluated directly over the database, without taking TGDs into account, and yet deliver the same answer. We now extend this approach to MODEQs and NEMODEQs, and we establish basic cases where Datalog queries can be rewritten to MODEQs, which we extend further in Section 8.

[2]This statement is often taken for granted and proof sketches have been communicated. A similar result was shown in [19] for a different notion of *width*. A modern account of the relevant proof techniques that uses our notion of treewidth is given in [22] for finite graphs. Formulating the proof of [19] in these terms, one could show Conjecture 12 [21]. We cautiously characterize this statement as a conjecture since no full proof has been published.

The most common notion is *first-order rewritability*, where a conjunctive query and a set of TGDs is rewritten into a first-order query – typically a union of conjunctive queries [1]. Importantly, the rewriting does not depend on the underlying database but only on the inital query and TGDs.

Example 11. *Given the rule set from Example 9 and the conjunctive query* $\exists v, w.(\mathit{hasAuthor}(v, w))$*, an appropriate first-order rewriting is*

$$\exists v, w.(\mathit{hasAuthor}(v, w)) \vee \exists v, w.(\mathit{cites}(v, w)) \vee \exists v.(\mathit{publication}(v)).$$

First-order rewritability is a desirable property as there are efficient implementations for evaluating first-order queries (that is, SQL). First-order rewritable sets of TGDs are also called *finite unification sets* [4]. It is undecidable whether a set of TGDs belongs to this class, but an iterative backward chaining algorithm can be defined that terminates on FO-rewritable rule sets and provides the rewritten FO formula [4]. Known sufficient conditions for FO-rewritability led to the definition of *atomic-hypothesis rules* and *domain restricted rules* [4], *linear Datalog+/–* [10], as well as *sticky sets of TGDs* and *sticky-join sets of TGDs* [11, 12]. These criteria were recently found to be subsumed by an efficiently checkable condition that gives rise to the class of *weakly recursive TGDs* [18]. Important FO-rewritable description logics include the DL-Lite family of logics [14]. However, many useful TGDs can not be expressed as first-order queries.

Example 12. *First-order logic cannot express transitive closure, so there is no first-order rewriting for the CQ* $s(v) \wedge r(v, w) \wedge s(w)$ *under the TGD* $r(x, y) \wedge r(y, z) \rightarrow r(x, z)$.

This motivates the consideration of more expressive query languages in query rewriting.

Example 13. *The TGD and query of Example 12 can be rewritten as a Datalog query*

$$r(x, y) \rightarrow r_{\text{IDB}}(x, y) \tag{28}$$
$$r_{\text{IDB}}(x, y) \wedge r_{\text{IDB}}(y, z) \rightarrow r_{\text{IDB}}(x, z) \tag{29}$$
$$s(v) \wedge r_{\text{IDB}}(v, w) \wedge s(w) \rightarrow \texttt{goal}, \tag{30}$$

but also as a conjunctive regular path query

$$\exists v, w.s(v) \wedge r^*(v, w) \wedge s(w).$$

Rewriting CQs under TGDs into Datalog queries is interesting, but the evaluation of Datalog queries remains complex. The undecidability of Datalog query containment also makes it intrinsically difficult to determine if such a query captures a set of TGDs. The use of C2RPQs is more interesting. Yet, to the best of our knowledge, rewriting of conjunctive queries into C2RPQs has been addressed only very implicitly by now [40]. Moreover, as discussed in Section 3.1, C2RPQs are still rather constrained: besides further structural restrictions, they only allow for recursion over binary predicates. We therefore consider query rewritability under MODEQs and NEMODEQs, defined as follows:

Definition 9 (NEMODEQ rewritability). *Let* Σ *be a set of TGDs and let* $Q[\boldsymbol{x}]$ *be a CQ. A (NE)MODEQ* $\mathfrak{Q}_{Q,\Sigma}$ *is a* rewriting *of* Q *under* Σ *if, for all databases* D *and potential query answers* $\boldsymbol{c}$*, we have* $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$ *iff* $D \models \mathfrak{Q}_{Q,\Sigma}[\boldsymbol{c}/\boldsymbol{x}]$. Σ *is called* (NE)MODEQ-rewritable *if every conjunctive query* Q *has a (NE)MODEQ rewriting for* Σ *and* Q.

Rewritability of conjunctive queries entails rewritability of MODEQs, i.e., the conditions of Definition 9 hold even when considering MODEQs instead of CQs. Indeed, CQs that occur in rule bodies in a MODEQ can generally be replaced using a MODEQ for the respective CQ, provided that the existentially quantified variables in the CQ are not used anywhere else in the rule body:

Lemma 14 (Replacement Lemma). *Consider a set* Σ *of TGDs, a conjunctive query* $Q = \exists \boldsymbol{y}.\psi[\boldsymbol{x}, \boldsymbol{y}]$*, and a NEMODEQ* $\mathfrak{Q}[\boldsymbol{x}]$ *that is a rewriting for* Σ *and* Q*. Then* Q *and* $\mathfrak{Q}$ *are equivalent in all models of* Σ*, i.e.,* $\Sigma \models \forall \boldsymbol{x}.Q[\boldsymbol{x}] \leftrightarrow \mathfrak{Q}[\boldsymbol{x}]$.

Let $\psi[\boldsymbol{t}/\boldsymbol{x}, \boldsymbol{y}'/\boldsymbol{y}]$ *be the conjunction of* Q *with variables* $\boldsymbol{x}$ *replaced by terms* $\boldsymbol{t}$ *and variables* $\boldsymbol{y}$ *replaced by variables* $\boldsymbol{y}'$*. We say that* $\psi[\boldsymbol{t}/\boldsymbol{x}, \boldsymbol{y}'/\boldsymbol{y}]$ *is a* match *in a Datalog rule* ρ *if* ρ *is of the form* $\psi[\boldsymbol{t}/\boldsymbol{x}, \boldsymbol{y}'/\boldsymbol{y}] \wedge \varphi \rightarrow \chi$ *where the* $\boldsymbol{y}'$ *occur neither in* φ *nor in* χ.

Given some NEMODEQ $\mathfrak{P}[z]$ *over* Σ*, let* $\mathfrak{P}'[z]$ *denote a NEMODEQ obtained by replacing a match* $\psi[\boldsymbol{t}/\boldsymbol{x}, \boldsymbol{y}'/\boldsymbol{y}]$ *of* Q *in some rule of* $\mathfrak{P}$ *by* $\mathfrak{Q}[\boldsymbol{t}/\boldsymbol{x}]$*, where the bound variables in* $\mathfrak{Q}$ *do not occur in* $\mathfrak{P}$*. Then* $\mathfrak{P}$ *and* $\mathfrak{P}'$ *are equivalent in all models of* Σ*, i.e.,* $\Sigma \models \forall z.\mathfrak{P}[z] \leftrightarrow \mathfrak{P}'[z]$.

Since every CQ can be expressed as MODEQ, all FO-rewritable rule sets are also MODEQ-rewritable. This covers all FO-rewritable classes mentioned above. Moreover, Section 3.2 implies that every set of monadic Datalog rules is MODEQ-rewritable. Since MODEQs are strictly more expressive than monadic Datalog queries, one would expect to find larger classes of MODEQ-rewritable TGDs. An appropriate generalization of monadic Datalog is as follows:

Definition 10 (j-oriented rule set). *We call a set of Datalog rules* Σ j-oriented *for the integer* j *if all head predicates have the same arity* n*, and* $1 \leq j \leq n$*, and we have: if a rule's body contains an atom* $p(\boldsymbol{t})$ *for some head predicate* p *and the rule's head contains an atom* $q(\boldsymbol{t}')$*, then* $\boldsymbol{t}$ *and* $\boldsymbol{t}'$ *agree on all positions other than possibly* j.

Intuitively speaking, recursive derivations in j-oriented rule sets can only modify the content of a single position j while keeping all other arguments fixed in all derived facts.

Example 14. *The following rule set* Σ_{family} *is* 3*-oriented. We use atoms* $\mathit{parentsSon}(x, y, z)$ *and* $\mathit{parentsDghtr}(x, y, z)$ *to denote that* z *is the son and daughter of* x *and* y*, respectively.*

$$\mathit{parentsSon}(x, y, z) \wedge \mathit{hasBrother}(z, z') \rightarrow \mathit{parentsSon}(x, y, z')$$
$$\mathit{parentsSon}(x, y, z) \wedge \mathit{hasSister}(z, z') \rightarrow \mathit{parentsDghtr}(x, y, z')$$
$$\mathit{parentsDghtr}(x, y, z) \wedge \mathit{hasBrother}(z, z') \rightarrow \mathit{parentsSon}(x, y, z')$$
$$\mathit{parentsDghtr}(x, y, z) \wedge \mathit{hasSister}(z, z') \rightarrow \mathit{parentsDghtr}(x, y, z')$$

Definition 11 (Single predicate rewriting). *For a* j*-oriented set* Σ *of Datalog rules and a head predicate* p *of* Σ*, a MODEP* $\mathbb{P}_{p,\Sigma}$ *is defined as follows. Let* $\mathtt{U}_q$ *be an auxiliary unary predicate for each head predicate* q *in* Σ*, let* $\mathtt{V}_i$ *be*

a auxiliary unary predicate for each $i \in \{1, \ldots, \mathsf{ar}(p)\}$ with $i \neq j$, and let $\tilde{z}_j$ be an additional variable not occurring in Σ. Then $\mathbb{P}_{p,\Sigma}$ contains the following rules:

- *a rule $\mathtt{U}_p(\lambda_j) \rightarrow \mathtt{hit}$;*
- *for each set variable $\mathtt{V}_i$, a rule $\rightarrow \mathtt{V}_i(\lambda_i)$ with empty body;*
- *for each $\psi \rightarrow q(t_1, \ldots, t_n) \in \Sigma$, a rule $\psi' \rightarrow \mathtt{U}_q(t_j)$ where ψ' is obtained from ψ by replacing each atom of the form $q'(t_1, \ldots, t'_j, \ldots, t_n)$ for a head predicate q' by $\mathtt{U}_{q'}(t'_j)$, and by adding for each term t_i with $i \neq j$ a new body atom $\mathtt{V}_i(t_i)$;*
- *a rule $q'(\lambda_1, \ldots, \tilde{z}_j, \ldots, \lambda_n) \rightarrow \mathtt{U}_{q'}(\tilde{z}_j)$ for each head predicate q'.*

For a list $\boldsymbol{z}$ of $\mathsf{ar}(p)$ variables, the MODEQ $\mathfrak{Q}_{p,\Sigma}[\boldsymbol{z}]$ is defined as $\mathbb{P}_{p,\Sigma}(\boldsymbol{z})$.

This operation allows us to express the extension of a predicate p by means of a MODEQ.

Theorem 15 (Single predicate rewriting correctness). *If Σ is j-oriented and p is a head predicate, then $\mathfrak{Q}_{p,\Sigma}[\boldsymbol{z}]$ is a rewriting for Σ and $p(\boldsymbol{z})$.*

Example 15. *For the rule set in Example 14, we obtain $\mathfrak{Q}_{parentsSon,\Sigma}[z_1, z_2, z_3] = \mathbb{P}_{parentsSon,\Sigma}(z_1, z_2, z_3)$ with rules*

$$\mathtt{U}_{parentsSon}(\lambda_3) \rightarrow \mathtt{hit} \qquad \rightarrow \mathtt{V}_1(\lambda_1) \qquad \rightarrow \mathtt{V}_2(\lambda_2)$$

$$\mathtt{V}_1(x) \wedge \mathtt{V}_2(y) \wedge \mathtt{U}_{parentsSon}(z) \wedge hasBrother(z, z') \rightarrow \mathtt{U}_{parentsSon}(z')$$

$$\mathtt{V}_1(x) \wedge \mathtt{V}_2(y) \wedge \mathtt{U}_{parentsSon}(z) \wedge hasSister(z, z') \rightarrow \mathtt{U}_{parentsDghtr}(z')$$

$$\mathtt{V}_1(x) \wedge \mathtt{V}_2(y) \wedge \mathtt{U}_{parentsDghtr}(z) \wedge hasBrother(z, z') \rightarrow \mathtt{U}_{parentsSon}(z')$$

$$\mathtt{V}_1(x) \wedge \mathtt{V}_2(y) \wedge \mathtt{U}_{parentsDghtr}(z) \wedge hasSister(z, z') \rightarrow \mathtt{U}_{parentsDghtr}(z')$$

$$parentsSon(\lambda_1, \lambda_2, \tilde{z}_3) \rightarrow \mathtt{U}_{parentsSon}(\tilde{z}_3)$$

$$parentsDghtr(\lambda_1, \lambda_2, \tilde{z}_3) \rightarrow \mathtt{U}_{parentsDghtr}(\tilde{z}_3).$$

The $\mathtt{V}$ predicates are not really needed here, since rule bodies do not impose any conditions on the respective variables. If no constants from $\mathbf{C}$ occur, one could always replace $\mathtt{V}$ with the respective λs, but expressions like $\mathtt{V}_1(c)$ would require an equality predicate to state $\lambda_1 \approx c$. For the semantics of NEMODEQs to be meaningful, constants λ_i must always be allowed to be equal to other constants, even if a unique name assumption is adopted for constants in $\mathbf{C}$.

Using the Replacement Lemma 14, we can extend Theorem 15 to arbitrary conjunctive queries:

Theorem 16 (j-orientedness implies rewritability). *Every j-oriented rule set is MODEQ-rewritable.*

8. REWRITING LAYERS OF TGDS

In the previous section, we have identified a first criterion for MODEQ-rewritability, and thus decidability of query entailment. However, there are many cases where only some of the given TGDs are rewritable. On the other hand, our results from Section 6 guarantee that NEMODEQ answering is still decidable in the presence of TGDs that are in bts, based on techniques that do not require rewriting. We now show how to combine both results by applying query rewriting to a subset of TGDs that is suitably "layered above" the remaining TGDs. This allows us to define a class of *fully oriented rule sets* that generalizes j-oriented rule sets to cover the full expressiveness of NEMODEQs. More generally, the combination of bts and NEMODEQ-rewritability captures some of the most expressive ontology languages for which query answering is known to be decidable.

Given a set of TGDs, we first clarify which subsets of TGDs can be rewritten into queries that can be evaluated over the remaining TGDs and databases without loosing results. To this end, we consider a notion of *rule dependency*. Related notions were first described in [3], and independently in [24]. Our presentation is closely related to [4].

Definition 12 (rule dependency, cut). *Let $\rho_1 = B_1 \rightarrow H_1$ and $\rho_2 = B_2 \rightarrow H_2$ be two TGDs. We say that ρ_2* depends on *ρ_1, written $\rho_1 \prec \rho_2$, if there is*

- *a database D,*
- *a substitution θ of all variables in B_1 with terms in D such that $\theta(B_1) \subseteq D$, and*
- *a substitution θ' of all variables in B_2 with terms in $D \cup \theta(H_1)$ such that $\theta'(B_2) \subseteq D \cup \theta(H_1)$ but $\theta'(B_2) \not\subseteq D$.*

We say that ρ_2 strongly depends on *ρ_1, written $\rho_1 \prec\!\!\prec \rho_2$, if H_1 contains a predicate that occurs in B_2.*

A (strong) cut *of a set of rules Σ is a partition $\Sigma_1 \cup \Sigma_2$ of Σ such that no rule in Σ_1 (strongly) depends on a rule in Σ_2. It is denoted $\Sigma_1 \triangleright \Sigma_2$ ($\Sigma_1 \triangleright\!\!\triangleright \Sigma_2$).*

The notion of rule dependencies encodes which rule can possibly trigger which other rule. Checking if a rule depends on another is an NP-complete task [4]. We thus introduce the simpler notion of strong dependency that can be checked in polynomial time. Clearly, dependency implies strong dependency, but the converse might not be true.

Example 16. *Consider the following Datalog rules:*

$$A(x) \wedge B(x) \rightarrow C(x) \tag{31}$$

$$C(x) \rightarrow \exists v.p(x, v), A(v) \tag{32}$$

Rule (32) *strongly depends on* (31) *and vice versa. Moreover,* (32) *depends on* (31)*, where the database of Definition 12 could be $D = \{A(c), B(c)\}$ using $\theta = \theta' = \{x \mapsto c\}$. However,* (31) *does not depend on* (32)*: a substitution θ' can map x to v (introduced as a new term when applying* (32)*), but the required fact $B(v)$ is not derived by* (32) *and cannot be in any initial database D (since it is not ground).*

Intuitively speaking, we can evaluate a TGD set Σ of the form $\Sigma_1 \triangleright \Sigma_2$ by first applying the rules in Σ_1, and then applying the rules of Σ_2. This is the essence of the following theorem, shown in [4].

Theorem 17 (Baget et al.). *Let Σ be a set of rules admitting a cut $\Sigma_1 \triangleright \Sigma_2$. Then, for every database D and every conjunctive query $Q[\boldsymbol{x}]$ we have that $D \cup \Sigma \models Q[\boldsymbol{x}/\boldsymbol{c}]$ exactly if there is a Boolean conjunctive query Q' such that $D \cup \Sigma_1 \models Q'$ and $Q', \Sigma_2 \models Q[\boldsymbol{x}/\boldsymbol{c}]$.*

We can thus rewrite queries in "layers" based on cuts:

LEMMA 18 (QUERY REWRITING WITH CUTS). *Let D be a database and let $\Sigma_1 \triangleright \Sigma_2$ be two sets of TGDs.*

1. *If $\mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{x}]$ is a NEMODEQ-rewriting of a conjunctive query $Q[\boldsymbol{x}]$, then:*

   $D \cup \Sigma_1 \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$ if and only if $D \cup \Sigma_1 \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$.

2. *If Σ_1 and Σ_2 are NEMODEQ-rewritable, then so is $\Sigma_1 \cup \Sigma_2$.*

This observation has two useful implications. The second item outlines an approach of extending and combining rewriting procedures that we will elaborate on in the remainder of this section. The first item hints at a very general approach for constructing TGD languages for which query answering is decidable, as expressed in the next theorem.

THEOREM 19 (QUERY ANSWERING WITH CUTS). *Consider rule sets $\Sigma_1 \triangleright \Sigma_2$, such that NEMODEQ answering is decidable under Σ_1, and NEMODEQ rewriting is decidable under Σ_2. Then NEMODEQ answering is decidable under $\Sigma_1 \cup \Sigma_2$.*

Using Theorem 13, this result specifically applies in cases where Σ_1 is a bounded treewidth set. We have noted that there are a number of effectively checkable criteria for this general class of TGDs. Much less is known about NEMODEQ rewritability beyond rewritability to unions of CQs. For a more general criterion, we can extend j-oriented rules along the lines of Lemma 18 (2).

DEFINITION 13 (FULLY ORIENTED RULE SET). *Let $\approx_{\ll}$ be the reflexive symmetric transitive closure of $\ll$. The set Σ is* fully oriented *if, for every $\rho \in \Sigma$, the equivalence class $[\rho]_{\ll} = \{\rho' \in \Sigma \mid \rho \approx_{\ll} \rho'\}$ is j-oriented (not necessarily for the same j and predicate arity).*

Given a fully oriented rule set, we can construct NEMODEQ rewritings for individual classes $[\rho]_{\ll}$ as in Definition 11, and combine these rewritings using Lemma 14:

THEOREM 20 (FULLY ORIENTED RULE SETS ARE REWRITABLE). *For a rule set Σ, it can be detected in polynomial time if Σ is fully oriented. Every fully oriented set Σ is MODEQ-rewritable.*

The use of $\ll$ instead of $\prec$ is relevant for deciding full orientedness in polynomial time. Even with this restriction, fully oriented Datalog queries have the same expressivity as NEMODEQs. Moreover, every NEMODEQ-rewritable TGD set can be expressed as a set of rules that can be transformed into a MODEQ using Theorem 20.

THEOREM 21 (NEMODEQS = FULLY ORIENTED DATALOG). *For every NEMODEQ $\mathfrak{Q}$, the rule set* datalog($\mathfrak{Q}$) *of Definition 7 is fully oriented. Moreover, for every NEMODEQ-rewritable set Σ of TGDs, there is a fully oriented set of Datalog rules Σ' such that:*

- *every predicate p in Σ has a corresponding head predicate q_p in Σ' that does not occur in Σ,*
- *for every database D and conjunctive query $Q[\boldsymbol{x}]$ that do not contain predicates of the form q_p, and for every list of constants $\boldsymbol{c}$, $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$ iff $D \cup \Sigma' \models Q'[\boldsymbol{c}/\boldsymbol{x}]$ where Q' is obtained from Q by replacing all predicates p by q_p.*

Another interesting criterion for NEMODEQ rewritability has been studied for Description Logics (DLs), where all predicates are of arity one or two. Expressive DL ontologies consist of two kinds of terminological axioms: concept inclusion axioms and role inclusion axioms. A role inclusion axiom is a Datalog rule of the form $R_1(x_0, x_1) \wedge \ldots \wedge R_n(x_{n-1}, x_n) \rightarrow R(x_0, x_n)$, which can be viewed as a generalized transitivity statement. Even in relatively inexpressive DLs, role inclusion axioms lead to undecidability of CQ answering [35]; in more expressive DLs they even lead to undecidability of simpler reasoning problems [**?**]. To overcome this, syntactic restrictions are imposed on role inclusions, to ensure that all role inclusions can be captured using finite automata [**?**, 33]. This is equivalent to rewriting role inclusions to regular path queries, and has indeed been exploited to decide CQ answering over expressive DLs [40]. This can be viewed as an implicit application of Theorem 19.[3] Many reasoning procedures for DLs are based on tree-like models, so various approaches to DL CQ answering can indeed be viewed as special combinations of bounded treewidth sets and NEMODEQ rewritable sets of TGDs [35, 40]. This supports the relevance of the general relationships observed here, and it motivates the further study of criteria for NEMODEQ rewritability. The rewriting methods used in DLs are based on regular languages, so it seems promising to consider their generalizations to graph grammars when dealing with arbitrary TGDs [22].

9. CONCLUSION

Monadically defined queries and their nested extension achieve a balance between expressivity and computability. They capture and significantly extend the query capabilities of (unions of) conjunctive queries as well as (unions of) conjunctive two-way regular path queries and monadic Datalog queries, the prevailing querying paradigms for structured and semi-structured databases. At the same time, they are conveniently expressible both in Datalog and monadic second-order logic. Yet, as opposed to these two, they ensure decidability of query containment and of query answering in the presence of depdendencies that allow for universal models with bounded treewidth – a property shared by many of the known decidable TGD classes.

The novel notions of MODEQ-rewritability and NEMODEQ-rewritability significantly extend first-order rewritability, which has been proven useful for theoretical considerations and practical realization of query answering alike. This extension allows for capturing much larger classes of TGD sets covering features like transitivity, which are considered difficult to handle within the known decision frameworks. Moreover, (NE)MODEQ-rewritable TGD sets can smoothly be integrated with bounded treewidth TGD sets as long as certain dependency constraints are obeyed. This provides a valuable perspective on rule-based data access as a task that can be solved by combining bottom-up techniques like the *chase* with top-down techniques like query rewriting.

[3] DL concept and role inclusion axioms are not always separated by a cut. There are well known rewriting methods to achieve this [32].

Our work raises a number of interesting questions for future research: How general are (NE)MODEQs? Are there larger, more expressive fragments which jointly satisfy all the established properties? Is every rewriting for a TGD set and a given CQ that is expressible in MSO logic equivalent to a NEMODEQ? What is the precise complexity of deciding query containment? Which more general syntactic criteria ensure NEMODEQ-rewritability? Can all fragments of TGDs (including those considered in description logics) for which conjunctive query answering is known to be decidable be captured as a combination of bts and NEMODEQ-rewriting? Answering these questions will not only contribute to our understanding of NEMODEQs, but also provide a more unified view on query answering under dependencies in general.

Acknowledgments. We gratefully acknowledge the contributions of various people: Bruno Courcelle and Detlef Seese gave very helpful comments on their work on MSO and structures of bounded treewidth; Diego Calvanese helped to clarify complexity issues of C2RPQs; Pierre Bourhis provided important insights that have led to our formulation of Theorem 10; various anonymous reviewers gave valuable input on a preliminary version of this work.

10. REFERENCES

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison Wesley, 1994.

[2] H. Andréka, I. Németi, and J. van Benthem. Modal languages and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27(3):217–274, 1998.

[3] J.-F. Baget. Improving the forward chaining algorithm for conceptual graphs rules. In D. Dubois, C. A. Welty, and M.-A. Williams, editors, *KR*, pages 407–414. AAAI Press, 2004.

[4] J.-F. Baget, M. Leclère, M.-L. Mugnier, and E. Salvat. On rules with existential variables: Walking the decidability line. *Artificial Intelligence*, 175(9–10):1620–1654, 2011.

[5] J.-F. Baget, M.-L. Mugnier, S. Rudolph, and M. Thomazo. Walking the complexity lines for generalized guarded existential rules. In Walsh [46], pages 712–717.

[6] C. Beeri and M. Y. Vardi. The implication problem for data dependencies. In *Proceedings of the 8th Colloquium on Automata, Languages and Programming*, pages 73–85. Springer, 1981.

[7] M. Benedikt, P. Bourhis, and P. Senellart. Monadic datalog containment. In A. Czumaj, K. Mehlhorn, A. M. Pitts, and R. Wattenhofer, editors, *ICALP (2)*, volume 7392 of *LNCS*, pages 79–91. Springer, 2012.

[8] G. Brewka and J. Lang, editors. *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR'08)*. AAAI Press, 2008.

[9] A. Calì, G. Gottlob, and M. Kifer. Taming the infinite chase: Query answering under expressive relational constraints. In Brewka and Lang [8], pages 70–80.

[10] A. Calì, G. Gottlob, and T. Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. In Paredaens and Su [41], pages 77–86.

[11] A. Calì, G. Gottlob, and A. Pieris. Advanced processing for ontological queries. *Proceedings of VLDB 2010*, 3(1):554–565, 2010.

[12] A. Calì, G. Gottlob, and A. Pieris. Query answering under non-guarded rules in Datalog+/-. In P. Hitzler and T. Lukasiewicz, editors, *Web Reasoning and Rule Systems*, volume 6333 of *LNCS*, pages 1–17. Springer, 2010.

[13] D. Calvanese. Personal communication, September 2011.

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

[15] D. Calvanese, G. D. Giacomo, M. Lenzerini, and M. Y. Vardi. Reasoning on regular path queries. *SIGMOD Record*, 32(4):83–92, 2003.

[16] A. K. Chandra, H. R. Lewis, and J. A. Makowsky. Embedded implicational dependencies and their inference problem. In *Conference Proceedings of the 13th Annual ACM Symposium on Theory of Computation (STOC'81)*, pages 342–354. ACM, 1981.

[17] A. K. Chandra and P. M. Merlin. Optimal implementation of conjunctive queries in relational data bases. In J. E. Hopcroft, E. P. Friedman, and M. A. Harrison, editors, *Proceedings of the 9th Annual ACM Symposium on Theory of Computing (STOC'77)*, pages 77–90. ACM, 1977.

[18] C. Civili and R. Rosati. A broad class of first-order rewritable tuple-generating dependencies. In P. Barceló and R. Pichler, editors, *Proceedings of the 2nd Workshop on the Resurgence of Datalog in Academia and Industry (Datalog 2.0, 2012)*, volume 7494 of *LNCS*. Springer, 2012.

[19] B. Courcelle. The monadic second-order logic of graphs, ii: Infinite graphs of bounded width. *Mathematical Systems Theory*, 21(4):187–221, 1989.

[20] B. Courcelle. Recursive queries and context-free graph grammars. *Theoretical Computer Science*, 78(1):217–244, 1991.

[21] B. Courcelle. Personal communication, August 2011.

[22] B. Courcelle and J. Engelfriet. Graph structure and monadic second-order logic, a language theoretic approach. manuscript, to be published at Cambridge University Press; available at http://www.labri.fr/perso/courcell/Book/TheBook.pdf, April 2011.

[23] E. Dantsin, T. Eiter, G. Gottlob, and A. Voronkov. Complexity and expressive power of logic programming. *ACM Computing Surveys*, 33(3):374–425, 2001.

[24] A. Deutsch, A. Nash, and J. B. Remmel. The chase revisited. In M. Lenzerini and D. Lembo, editors, *Proc. 27th Symposium on Principles of Database Systems (PODS'08)*, pages 149–158. ACM, 2008.

[25] A. Deutsch and V. Tannen. Reformulation of XML queries and constraints. In D. Calvanese, M. Lenzerini, and R. Motwani, editors, *Proceedings of the 9th International Conference on Database Theory (ICDT 2003)*, volume 2572 of *LNCS*, pages 225–241. Springer, 2003.

[26] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.

[27] D. Florescu, A. Levy, and D. Suciu. Query containment for conjunctive queries with regular expressions. In *Proceedings of the seventeenth ACM symposium on Principles of database systems*, PODS '98, pages 139–148. ACM, 1998.

[28] G. Gottlob and C. Koch. Monadic datalog and the expressive power of languages for web information extraction. *J. ACM*, 51(1):74–113, 2004.

[29] G. Gottlob and C. H. Papadimitriou. On the complexity of single-rule datalog queries. *Inf. Comput.*, 183(1):104–122, 2003.

[30] S. Greco and F. Spezzano. Chase termination: A constraints rewriting approach. *Proceedings of VLDB 2010*, 3(1):93–104, 2010.

[31] N. Immerman. Languages that capture complexity classes. *SIAM J. Comput.*, 16(4):760–778, 1987.

[32] Y. Kazakov. $\mathcal{RIQ}$ and $\mathcal{SROIQ}$ are harder than $\mathcal{SHOIQ}$. In Brewka and Lang [8], pages 274–284.

[33] Y. Kazakov. An extension of complex role inclusion axioms in the description logic $\mathcal{SROIQ}$. In *Proceedings of the 5th International Joint Conference on Automated Reasoning (IJCAR 2010)*, LNCS. Springer, 2010.

[34] M. Krötzsch and S. Rudolph. Extending decidable existential rules by joining acyclicity and guardedness. In Walsh [46], pages 963–968.

[35] M. Krötzsch, S. Rudolph, and P. Hitzler. Conjunctive queries for a tractable fragment of OWL 1.1. In K. Aberer et al., editor, *Proceedings of the 6th International Semantic Web Conference (ISWC'07)*, volume 4825 of *LNCS*, pages 310–323. Springer, 2007.

[36] L. Libkin. *Elements of Finite Model Theory*. Springer, 2004.

[37] B. Marnette. Generalized schema-mappings: from termination to tractability. In Paredaens and Su [41], pages 13–22.

[38] M. Meier, M. Schmidt, and G. Lausen. On chase termination beyond stratification. *Proceedings of VLDB 2009*, 2(1):970–981, 2009.

[39] M.-L. Mugnier. Ontological query answering with existential rules. In S. Rudolph and C. Gutierrez, editors, *Web Reasoning and Rule Systems (RR 2011)*, volume 6902 of *LNCS*, pages 2–23. Springer, 2011.

[40] M. Ortiz, S. Rudolph, and M. Simkus. Query answering in the Horn fragments of the description logics $\mathcal{SHOIQ}$ and $\mathcal{SROIQ}$. In Walsh [46], pages 1039–1044.
[41] J. Paredaens and J. Su, editors. *Proc. 28th Symposium on Principles of Database Systems (PODS'09)*. ACM, 2009.
[42] O. Shmueli. Equivalence of DATALOG queries is undecidable. *J. Log. Program.*, 15(3):231–241, 1993.
[43] L. J. Stockmeyer. *The Complexity of Decision Problems in Automata Theory and Logic*. PhD thesis, Massachusetts Institute of Technology, 1974.
[44] L. J. Stockmeyer. The polynomial-time hierarchy. *Theor. Comput. Sci.*, 3(1):1–22, 1976.
[45] M. Y. Vardi. The complexity of relational query languages. In H. R. Lewis, B. B. Simons, W. A. Burkhard, and L. H. Landweber, editors, *STOC*, pages 137–146. ACM, 1982.
[46] T. Walsh, editor. *Proc. 22nd Int. Joint Conf. on Artificial Intelligence (IJCAI'11)*. AAAI Press/IJCAI, 2011.

APPENDIX

This appendix provides proofs that have been omitted in the main paper for reasons of space.

Proofs for Section 3

THEOREM 2. *For every MODEQ $\mathfrak{Q}[\boldsymbol{x}]$, the set of models of $\exists \boldsymbol{x}.\mathfrak{Q}[\boldsymbol{x}]$ is closed under homomorphism.*

PROOF. As MODEQs are a special case of NEMODEQs, the result follows from Theorem 7. □

THEOREM 3 (MODEQ ANSWERING COMPLEXITY). *Testing if $\boldsymbol{c}$ is an answer to a MODEQ $\mathfrak{Q}[\boldsymbol{x}]$ over a database D is* P*-complete in the size of D, and* NP*-complete in the combined size of D and $\mathfrak{Q}$.*

PROOF. P membership for data complexity is a direct consequence of the fact that MODEQs are subsumed by NEMODEQs which in turn can be polynomially translated to Datalog queries for which the problem is known to be in P [23, Theorem 4.4].

While P hardness is a consequence of the fact that MODEQs subsume monadic Datalog (shown later in Theorem 5), we provide a direct proof by reducing entailment in propositional Horn logic to MODEQ answering. Given a set $\mathcal{H}$ of propositional Horn clauses, we introduce for every propositional atom a occurring therein a constant c_a. We also introduce one additional constant *nil*. Moreover, for every Horn clause $C \in \mathcal{H}$ with $C = a_1 \wedge \ldots \wedge a_n \rightarrow a$, we introduce constants $b_{C,1}, \ldots, b_{C,n}$ and ground atoms $\mathit{entails}(b_{C,1}, c_a)$, $\mathit{first}(b_{C,i}, c_{a_i})$ for all $i \in \{1, \ldots, n\}$, as well as $\mathit{rest}(b_{C,n}, \mathit{nil})$, and $\mathit{rest}(b_{C,i}, b_{C,i+1})$ for all $i \in \{1, \ldots, n-1\}$. Then the a propositional atom is entailed by $\mathcal{H}$ exactly if it is an answer for the MODEQ $\mathfrak{Q}_{\mathit{true}}[x] = \mathbb{P}(x)$ with $\mathbb{P}$ consisting of the rules

$$\begin{aligned} & \rightarrow \quad \mathtt{U}_1(\mathit{nil}) \\ \mathit{first}(y,z) \wedge \mathtt{U}_1(z) \wedge \mathit{rest}(y,z') \wedge \mathtt{U}_1(z') & \rightarrow \quad \mathtt{U}_1(y) \\ \mathtt{U}_1(y) \wedge \mathit{entails}(y,z) & \rightarrow \quad \mathtt{U}_1(z) \\ \mathtt{U}_1(\lambda_1) & \rightarrow \quad \mathtt{hit}. \end{aligned}$$

NP hardness for combined complexity is a straightforward consequence of the fact that MODEQs capture CQs for which the problem is known to be NP-hard.

We prove NP membership by showing that each query match gives rise to a witness which can be verified in polynomial time. For a database D and a MODEQ $\mathfrak{Q}[\boldsymbol{x}] = \exists \boldsymbol{y}.\psi[\boldsymbol{x}, \boldsymbol{y}]$ (with ψ being a conjunction containing normal atoms and MODEP atoms), $\boldsymbol{c}$ is an answer iff there exists a tuple $\boldsymbol{d}$ of constants, such that $D \models \psi[\boldsymbol{c}, \boldsymbol{d}]$. As $\psi[\boldsymbol{c}, \boldsymbol{d}]$ is a conjunction of (normal and MODEP) ground atoms, we can check this for every conjunct separately. For normal ground atoms $p(\boldsymbol{e})$, we have $D \models p(\boldsymbol{e})$ iff $p(\boldsymbol{e}) \in D$, which can be checked in polynomial time. For MODEP ground atoms $\mathbb{P}(\boldsymbol{e})$, we find that $D \models \mathbb{P}(\boldsymbol{e})$ iff there is a Datalog derivation of `hit` from D via applying rules from $\mathbb{P}[\boldsymbol{e}/\boldsymbol{\lambda}]$. Due to the arity restriction of head predicates in $\mathbb{P}$ there are only polynomially many atoms that can be derived via $\mathbb{P}[\boldsymbol{e}/\boldsymbol{\lambda}]$, therefore there must be a proof for $\mathbb{P}(\boldsymbol{e})$ of polynomial size.

Hence, we define the witness for $\boldsymbol{c}$ being an answer for D and $\mathfrak{Q}[\boldsymbol{x}] = \exists \boldsymbol{y}.\psi[\boldsymbol{x}, \boldsymbol{y}]$ to contain the tuple $\boldsymbol{d}$ providing us with bindings to the existentially quantified variables and a polynomial-size proof for every MODEP atom $\mathbb{P}(\boldsymbol{e})$ contained in $\psi[\boldsymbol{c}, \boldsymbol{d}]$. Clearly this witness can be verified in polynomial time. □

Theorem 4 (MODEQs capture C2RPQs). *For every C2RPQ Q, the MODEQ* $\mathsf{modeq}(Q)$ *can be constructed in linear time, and is equivalent to Q. In particular, the answers for Q and* $\mathsf{modeq}(Q)$ *coincide.*

Proof. The linear construction time follows directly from the definition (taking into account that for every regular expression, the corresponding automaton can be obtained in linear time). The rest of the claim is an immediate consequence of the fact that P and $\mathsf{modep}(P)$ are equivalent for every conjunctive 2-way regular path predicate P. This can be seen as follows (letting $\mathcal{I}$ be an arbitrary interpretation):

To see that $P^{\mathcal{I}} \subseteq \mathsf{modep}(P)^{\mathcal{I}}$ we recap that $P^{\mathcal{I}}$ contains those pairs $\langle \delta, \delta' \rangle$ where δ' can be reached from δ over a path $\delta = \delta_0 \xrightarrow{\gamma_1} \cdots \xrightarrow{\gamma_n} \delta_n = \delta'$ such that $\gamma_1 \dots \gamma_n$ satisfies the regular expression P. The latter is the case exactly if the corresponding automaton $\mathcal{A}_P$ accepts $\gamma_1 \dots \gamma_n$, i.e., there is a sequence $\langle s_0, \dots, s_n \rangle$ of states with s_0 being an initial and s_n being a final state such that $\langle s_{i-1}, \gamma_i, s_i \rangle \in T$ for all $i \in \{1, \dots, n\}$. We can now show that $\langle \delta, \delta' \rangle$ must be in $\mathsf{modep}(P)^{\mathcal{I}}$ by noting that in every extension $\mathcal{I}'$ of $\mathcal{I}$ satisfying the rules from $\mathsf{modep}(P)$ as well as $\delta = \lambda_1^{\mathcal{I}'}$ and $\delta' = \lambda_2^{\mathcal{I}'}$ (1) we have $\delta \in \mathtt{U}_{s_0}$ due to the according initial-state-rule, (2) we can thus infer by induction that $\delta_i \in \mathtt{U}_{s_i}$ by the according transition rules (3) we hence arrive at $\delta' \in \mathtt{U}_{s_n}$ and can apply the final-state-rule to prove that `hit` must hold.

The proof of $\mathsf{modep}(P)^{\mathcal{I}} \subseteq P^{\mathcal{I}}$ is very similar: starting from a derivation sequence for `hit`, we can easily construct an according path from δ to δ'. □

Theorem 5 (MODEQs capture monadic Datalog). *For every monadic Datalog query Q, the MODEQ* $\mathsf{modeq}(Q)$ *can be constructed in linear time, and is equivalent to Q. In particular, the answers for Q and* $\mathsf{modeq}(Q)$ *coincide.*

Proof. The linear construction time follows directly from the definition. For showing the equivalence claim between the monadic Datalog query $Q = (\mathtt{goal}, \mathbb{S})$ and $\mathsf{modeq}(Q)$, we assume w.l.o.g. that all unary IDB predicates in Q are already of the shape $\mathtt{U}_j$ and therefore the rule sets of $\mathbb{S}$ and $\mathsf{modeq}(Q)$ coincide on all rules not referring to `goal` or `hit`. Given an interpretation $\mathcal{I}$, let $\mathcal{I}''$ denote the extended interpretation obtained by saturating $\mathcal{I}$ under these rules. Then, a tuple $\boldsymbol{\delta}$ from $\mathcal{I}$ is in the extension of Q if $\mathcal{I}'' \models B[\boldsymbol{\delta}/\boldsymbol{x}]$ for the body of some rule $B \to \mathtt{goal}[x]$ from $\mathbb{S}$. Yet, since $\mathsf{modeq}(Q)$ contains a rule $B[\boldsymbol{\lambda}/\boldsymbol{x}] \to \mathtt{hit}$ this is exactly the case if $\boldsymbol{\delta}$ is in the extension of $\mathsf{modeq}(Q)$. □

Proofs for Section 4

Proposition 22. *The NEMODEQ $\mathfrak{Q}_4$ of Example 6 cannot be expressed as a MODEQ.*

Proof. Suppose for a contradiction that there is a MODEQ $\mathfrak{Q}$ that expresses the query $\mathfrak{Q}_4$. Without loss of generality, assume that no MODEP occurs more than once in $\mathfrak{Q}$ (if it does, we can replace it by a copy that is defined by the same rules). Moreover, we assume that the MODEPs in $\mathfrak{Q}$ use distinct names for their auxiliary unary predicates `U` and for their constants λ, again without loss of generality. Let ℓ be the maximal number of variables in any of the rules used (in MODEPs) in $\mathfrak{Q}$, and let m be the total number of constants λ (in MODEPs) in $\mathfrak{Q}$ (i.e., the sum of the arities of all MODEPs in $\mathfrak{Q}$).

We first define a structure S of the shape illustrated in Fig. 3 to contain exactly the following relations:

- S contains a chain of relations $a \xrightarrow{\mathit{linkedTo}} e_1 \xrightarrow{\mathit{linkedTo}} \dots \xrightarrow{\mathit{linkedTo}} e_{2\ell} \xrightarrow{\mathit{linkedTo}} b$;
- for each e_i ($i = 1, \dots, 2\ell$), S contains a chain $e_i \xrightarrow{\mathit{certifiedBy}} f_{i1} \xrightarrow{\mathit{certifiedBy}} \dots \xrightarrow{\mathit{certifiedBy}} f_{i\ell} \xrightarrow{\mathit{certifiedBy}} c$;
- the elements a and b are in the unary relation *friend*.

Thus S has $2\ell(\ell + 1) + 3$ elements. We now use multiple copies of S to define a structure T of the shape that is illustrated in Fig. 4. The structure T consists of $m + 1$ copies of S, connected in a chain. Formally, T is defined as the disjoint union of $m+1$ copies $S_0, \dots, S_m$ of S, factorized by the equivalence relation $\{b_i \approx a_{i+1} \mid 0 \leq i < m\}$, where we use a_i and b_i to denote the elements a and b, respectively, from S_i.

Clearly, $\{v \mapsto a_0, w \mapsto b_m\}$ is an answer to $\mathfrak{Q}_4$ over T. Thus, by assumption, $\{v \mapsto a_0, w \mapsto b_m\}$ is also an answer to $\mathfrak{Q}$ over T. This match is therefore entailed in every model of T. By Theorem 2 and Fact 1, we only need to consider the universal model $\mathcal{I}(T)$, the elements of which are in one-to-one correspondence with the elements of T. Thus, the validity of the answer $\{v \mapsto a_0, w \mapsto b_m\}$ is witnessed by a query match on $\mathcal{I}(T)$, which is based on a particular match for each MODEP in $\mathfrak{Q}$. We can consider these matches as a list of m elements of $\mathcal{I}(T)$ (and thus also of T) which were used to interpret the constants λ in $\mathfrak{Q}$. We call these the *matched elements* of T. Since there are only m constants λ in $\mathfrak{Q}$, there are at most m matched elements. Hence, there is some S_o in T which does not contain any matched elements, other than possibly a_o and b_o (since they also belong to adjacent substructures).

We construct a new structure T' from T as follows:

- For all $(e_i)_o$ in the substructure S_o with $i = 2, \dots, 2\ell$, add a chain $(e_i)_o \xrightarrow{\mathit{certifiedBy}} (f'_{i1})_o \xrightarrow{\mathit{certifiedBy}} \dots \xrightarrow{\mathit{certifiedBy}} (f'_{i\ell})_o \xrightarrow{\mathit{certifiedBy}} c'_o$, where c'_o and the $(f'_{ij})_o$ are fresh elements distinct from any element in T.
- Delete all elements $(f_{(2\ell)1})_o, \dots, (f_{(2\ell)\ell})_o$.

T' no longer contains a chain of links from a_o to b_o that is certified by a single authority: only the first $2\ell - 1$ elements are certified by c_o, and only the last $2\ell - 1$ elements are certified by c'_o. Likewise, it is not possible to split the chain between a_o and b_o into multiple chains of links, since none of the intermediate elements belongs to the class *friend*. Thus $\{v \mapsto a_0, w \mapsto b_m\}$ is not an answer to $\mathfrak{Q}_4$.

Nevertheless, we can show that $\{v \mapsto a_0, w \mapsto b_m\}$ is an answer to $\mathfrak{Q}$, which yields the required contradiction. To

do this, it suffices to show that the matches of MODEPs in $\mathfrak{Q}$ that witness the answer $\{v \mapsto a_0, w \mapsto b_m\}$ of $\mathfrak{Q}$ over T are still valid. Thus consider some MODEP $\mathbb{P}$ in $\mathfrak{Q}$ with auxiliary constants $\lambda_1, \dots, \lambda_{\mathsf{ar}(\mathbb{P})}$ that have been matched to elements $\boldsymbol{\delta} = \delta_1, \dots, \delta_{\mathsf{ar}(\mathbb{P})}$ of T (or, equivalently, $\mathcal{I}(T)$). By construction of T', the elements $\boldsymbol{\delta}$ are also in T' and $\mathcal{I}(T')$. We show that $\boldsymbol{\delta} \in \mathbb{P}^{\mathcal{I}(T')}$.

By our assumption, every extension $\mathcal{I}$ of $\mathcal{I}(T)$ with $\lambda_i^{\mathcal{I}} = \delta_i$ and $\mathcal{I} \models \mathbb{P}$ satisfies `hit`. Hence there is a sequence of applications of rules from $\mathbb{P}$ by which `hit` can be derived (using the pre-defined interpretation of λ_i). The same sequence of rules can be used to derive `hit` for extensions $\mathcal{I}'$ of $\mathcal{I}(T')$ with $\lambda_i^{\mathcal{I}'} = \delta_i$. Indeed, rules of $\mathbb{P}$ have at most ℓ variables, hence their applicability does only depend on a substructure of size at most ℓ. It is easy to see that $\mathcal{I}(T)$ and $\mathcal{I}(T')$ have the same substructures of size ℓ, up to renaming of elements that are not referred to by any constants in $\mathbb{P}$. It can be shown by induction that this property is preserved when considering unary predicates derived by any number of rule applications. This shows that $\mathcal{I}' \models \mathbb{P}$ implies $\mathcal{I}' \models$ `hit`. □

Theorem 6 (Datalog expressibility of NEMODEQs). *For any NEMODEQ $\mathfrak{Q}$, $\mathsf{datalog}(\mathfrak{Q})$ can be constructed in linear time. Moreover, the queries $\mathfrak{Q}[\boldsymbol{x}]$ and $\langle p_{\mathfrak{Q}}, \mathsf{datalog}(\mathfrak{Q})\rangle$ are equivalent. In particular, their answers coincide.*

Proof. The claimed linear time bound is immediate from the definition.

The claimed equivalence is a straightforward consequence of the following claim, which we will prove in the sequel: $\forall \boldsymbol{x}.\mathbb{P}(\boldsymbol{x}) \leftrightarrow (p_{\mathbb{P}}, \mathsf{datalog}(\mathbb{P}))[\boldsymbol{x}]$ for all NEMODEPs $\mathbb{P}$ of degree ≥ 1. We show this by induction on the degree of $\mathbb{P}$.

We first show $\forall \boldsymbol{x}.\mathbb{P}(\boldsymbol{x}) \rightarrow (p_{\mathbb{P}}, \mathsf{datalog}(\mathbb{P}))[\boldsymbol{x}]$ for all NEMODEPs $\mathbb{P}$. Consider an interpretation $\mathcal{I}$ for which $\mathcal{I} \models \mathsf{datalog}(\mathbb{P})$ and a variable assignment $\mathcal{Z}$ such that $\mathcal{I}, \mathcal{Z} \models \mathbb{P}(x_1, \dots, x_m)$. By definition, the latter means that all $\mathcal{I}'$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ and $\mathcal{I}' \models \mathbb{P}$ must satisfy `hit`. $(*)$

We define such an $\mathcal{I}'$ as follows:

- $\mathcal{I}'$ has the same domain as $\mathcal{I}$ and the two interpretations coincide on $\mathcal{I}$'s vocabulary,
- $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ for all $i \in \{1, \dots, m\}$
- $\mathtt{U}_i^{\mathcal{I}'} = \{\delta \mid \langle \delta, \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle \in \hat{U}_i^{\mathcal{I}}\}$
- $\mathtt{hit}^{\mathcal{I}'} = \{\langle\rangle\}$ if $\langle \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle \in p_{\mathbb{P}}^{\mathcal{I}}$ and $\mathtt{hit}^{\mathcal{I}'} = \emptyset$ otherwise.

We now show that indeed $\mathcal{I}' \models \rho$ for all $\rho \in \mathbb{P}$. Let $\rho = \forall \boldsymbol{y}.B_1 \wedge \dots \wedge B_\ell \rightarrow H$ and assume for some $\mathcal{Z}'$ that $\mathcal{I}', \mathcal{Z}' \models B$ for all $B \in \{B_1, \dots, B_\ell\}$. Let now $\rho' = \forall \boldsymbol{y}'.B_1' \wedge \dots \wedge B_\ell' \rightarrow H'$ be the rule from $\mathsf{datalog}(\mathbb{P})$ corresponding to ρ and let $\mathcal{Z}'' = \{x_{\lambda_i} \mapsto \lambda_i^{\mathcal{I}'} \mid 1 \leq i \leq m\}$. Then we find that $\mathcal{I}, \mathcal{Z}'' \cup \mathcal{Z}' \models B'$ for all $B' \in \{B_1', \dots, B_\ell'\}$ because

- for B a normal atom, we have $B' = B[\lambda_1/x_1 \dots \lambda_m/x_m]$ as well as $\lambda_i^{\mathcal{I}', \mathcal{Z}'} = x_i^{\mathcal{I}, \mathcal{Z}'' \cup \mathcal{Z}'}$ for all $i \in \{1, \dots, m\}$ and the extensions of the corresponding predicate in $\mathcal{I}'$ and $\mathcal{I}$ coincide by construction,
- for B an atom of degree ≥ 1, the coincidence follows from the induction hypothesis,
- for $B = \mathtt{U}_j(t)$, we have $B' = \hat{U}_j(t, x_{\lambda_1}, \dots, x_{\lambda_m})$ and satisfaction coincides by construction

Now, since $\mathcal{I} \models \mathsf{datalog}(\mathbb{P})$, we obtain $\mathcal{I}, \mathcal{Z} \cup \mathcal{Z}' \models H'$. Therefrom follows $\mathcal{I}', \mathcal{Z}' \models H$ since

- for $H' = \hat{U}_j(t, x_{\lambda_1}, \dots, x_{\lambda_m})$ we have $H = \mathtt{U}_j(t)$ and satisfaction coincides by construction,
- for $H' = p_{\mathbb{P}}(x_{\lambda_1}, \dots, x_{\lambda_m})$ we have $H =$ `hit` and satisfaction coincides by construction.

Now, having constructed an $\mathcal{I}'$ with $\mathcal{I}'(\lambda_i) = \mathcal{Z}(x_i)$ and $\mathcal{I}' \models \mathbb{P}$ we can infer $\mathcal{I}' \models$ `hit` via $(*)$ and therefore $\mathcal{I}, \mathcal{Z} \models p_{\mathbb{P}}(x_1, \dots, x_m)$, which finishes the first direction.

We proceed by showing $\models \forall \boldsymbol{x}.(p_{\mathbb{P}}, \mathsf{datalog}(\mathbb{P}))[\boldsymbol{x}] \rightarrow \mathbb{P}(\boldsymbol{x})$. To this end, consider an interpretation $\mathcal{I}$ and a variable assignment $\mathcal{Z}$ such that $\mathcal{I}, \mathcal{Z} \models (p_{\mathbb{P}}, \mathsf{datalog}(\mathbb{P}))[\boldsymbol{x}]$. The latter means that in every extension $\mathcal{J}$ of $\mathcal{I}$ that satisfies all rules from $\mathsf{datalog}(\mathbb{P})$ must also hold $\mathcal{J}, \mathcal{Z} \models p_{\mathbb{P}}(\boldsymbol{x})$. $(**)$

Let now $\mathcal{I}'$ be an arbitrary extension of $\mathcal{I}$ satisfying $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ for all $i \in \{1, \dots, m\}$ and $\mathcal{I}' \models \rho$ for all $\rho \in \mathbb{P}$.

We now construct an extension $\mathcal{J}'$ of $\mathcal{I}$, as follows

- $\mathcal{J}'$ has the same domain as $\mathcal{I}$ and the two interpretations coincide on the vocabulary of the latter,
- $\hat{U}_i^{\mathcal{J}'} = \{\langle \delta, \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle \mid \delta \in \mathtt{U}_i^{\mathcal{I}'}\} \cup \{\langle \delta_0, \delta_1, \dots, \delta_n\rangle \mid \delta_0 \in \Delta^{\mathcal{I}}, \langle \delta_1, \dots, \delta_m\rangle \neq \langle \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle\}$
- $p_{\mathbb{P}}^{\mathcal{J}'} = \Delta^{\mathcal{I}}$ if $\mathtt{hit}^{\mathcal{I}'} = \{\langle\rangle\}$ (i.e. if $\mathcal{I}'$ satisfies `hit`), otherwise $p_{\mathbb{P}}^{\mathcal{J}'} = \Delta^{\mathcal{I}} \setminus \{\langle \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle\}$.

By construction, $\mathcal{J}'$ satisfies $\mathsf{datalog}(\mathbb{P})$ and therefore, by $(**)$, follows $\mathcal{J}', \mathcal{Z} \models p_{\mathbb{P}}(\boldsymbol{x})$, i.e., $\langle \mathcal{Z}(x_1), \dots, \mathcal{Z}(x_m)\rangle \in p_{\mathbb{P}}^{\mathcal{J}'}$. Again by construction of $\mathcal{J}$, this is the case only if $\mathtt{hit}^{\mathcal{I}'} = \{\langle\rangle\}$, whence we have shown $\mathcal{I}' \models$ `hit`, which finishes the second direction. □

Theorem 7. *For any NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$, the set of models of $\exists \boldsymbol{x}.\mathfrak{Q}$ is closed under homomorphisms.*

Proof. As established in Theorem 6, NEMODEQs are expressible as Datalog queries. However, for the latter, preservation under homomorphisms is well known and easy to show. □

Theorem 8 (MSO expressibility of NEMODEQs). *For every NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$, $\mathsf{mso}(\mathfrak{Q}[\boldsymbol{x}])$ can be constructed in linear time and is equivalent to $\mathfrak{Q}[\boldsymbol{x}]$.*

Proof. We show the equivalence for NEMODEPs by an induction on their degree. From this, the general claim is a straightforward consequence.

Let $\mathcal{I}$ be an arbitrary interpretation and $\mathcal{Z}$ a variable assignment for $\boldsymbol{x}$. Then we can establish the following sequence of equivalent propositions:

- $\mathcal{I}, \mathcal{Z} \models \mathbb{P}(\boldsymbol{x})$
- every extension $\mathcal{I}'$ of $\mathcal{I}$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ and $\mathcal{I}' \models \mathbb{P}$ must satisfy `hit`
- every extension $\mathcal{I}'$ of $\mathcal{I}$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ and $\mathcal{I}' \models \mathbb{P}'$ with $\mathbb{P}' = \mathbb{P}[\bot/\mathtt{hit}]$ must satisfy $\bot$; in other words, no such extension exists
- (via induction hypothesis) no extension $\mathcal{I}'$ of $\mathcal{I}$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ simultaneously satisfies $\mathsf{mso}(\rho)$ for all $\rho \in \mathbb{P}$
- no extension $\mathcal{I}'$ of $\mathcal{I}$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ satisfies $\bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho)$

- all extensions $\mathcal{I}'$ of $\mathcal{I}$ with $\lambda_i^{\mathcal{I}'} = \mathcal{Z}(x_i)$ satisfy $\neg \bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho)$
- letting $\mathcal{I}_\lambda$ denote $\mathcal{I}$ extended by $\{\lambda_i \mapsto \mathcal{Z}(x_i) \mid 1 \leq i \leq m\}$, $\mathcal{I}_\lambda, \Xi \models \neg \bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho)$ holds for every set-variable assignment $\Xi : \{U_1, \dots, U_k\} \to 2^{\Delta^{\mathcal{I}}}$
- $\mathcal{I}_\lambda \models \forall U_1, \dots, U_k. \neg \bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho)$
- $\mathcal{I}, \mathcal{Z} \models (\forall U_1, \dots, U_k. \neg \bigwedge_{\rho \in \mathbb{P}} \mathsf{mso}(\rho))[\boldsymbol{x}/\boldsymbol{\lambda}]$
- $\mathcal{I}, \mathcal{Z} \models \mathsf{mso}(\mathbb{P})[\boldsymbol{x}]$

Thus we have shown the claimed correspondence. □

Theorem 9 (complexity of NEMODEQ answering). *Checking if $\boldsymbol{c}$ is an answer to a NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$ over a database D is* P*-complete in the size of D, and* PSpace*-complete in the size of D and $\mathfrak{Q}$.*

Proof. PSpace membership for combined complexity is a direct consequence of PSpace completeness of model checking in monadic second-order logic [43, 45], keeping in mind that due to Fact 1, entailment coincides with model checking if the premise is a set of ground facts.

We show PSpace hardness by providing a reduction from the validity problem of quantified Boolean formulae (QBFs). We recap that for any QBF, it is possible to construct in polynomial time an equivalent QBF that has the specific shape $Q_1 x_1 Q_2 x_2 \dots Q_n x_n \bigvee_{L \in \mathcal{L}} \bigwedge_{\ell \in L} \ell$, with $Q_1, \dots Q_n \in \{\exists, \forall\}$ and $\mathcal{L}$ being a set of sets of literals over the propositional variables $x_1, \dots, x_n$. In words, we assume our QBF to be in prenex form with the propositional part of the formula in disjunctive normal form. For every literal set $L = \{x_{k_1}, \dots, x_{k_i}, \neg x_{k_{i+1}}, \dots, \neg x_{k_j}\}$, we now define the n-ary MODEP $p_L = \{t(\lambda_{k_1}) \wedge \dots \wedge t(\lambda_{k_i}) \wedge f(\lambda_{k_{i+1}}) \wedge \dots \wedge f(\lambda_{k_j}) \to \mathtt{hit}\}$. Moreover, we define the n-ary MODEP $p_{\mathcal{L}} = \{p_L(\lambda_1, \dots, \lambda_n) \to \mathtt{hit} \mid L \in \mathcal{L}\}$. Letting $p_{\mathcal{L}} = p_n$ we now define MODEPs $p_{n-1} \dots p_0$ in descending order. If $Q_i = \exists$, then the $i-1$-ary MODEP p_{i-1} is defined as the singleton rule set $\{p_{i-1}(\lambda_1, \dots, \lambda_{i-1}, y) \to \mathtt{hit}\}$. In case $Q_i = \exists$, we let $p_{i-1} = \{p_{i-1}(\lambda_1, \dots, \lambda_{i-1}, y) \to \mathtt{U}(y) \quad \mathtt{U}(y) \wedge f(y) \wedge \mathtt{U}(z) \wedge t(z) \to \mathit{hit}\}$. Now, let D be the database containing the two individuals 0 and 1 and the facts $f(0)$ and $t(1)$. We now show that the considered QBF is true exactly if $D \models p_0()$. To this end, we first note that by construction the extension of p_L contains exactly those n-tuples $\langle \delta_1, \dots, \delta_n \rangle$ for which the corresponding truth value assignment *val*, sending x_i to **true** iff $\delta_i = 1$, makes the formula $\bigwedge_{\ell \in L} \ell$ true. In the same way, the extension of $p_{\mathcal{L}}$ represents the set of truth value assignments satisfying $\bigvee_{L \in \mathcal{L}} \bigwedge_{\ell \in L} \ell$. Then, by descending induction, we can show that the extensions of p_i encode the assignments to free propositional variables of the subformula $Q_{i+1} x_{i+1} \dots Q_n x_n \bigvee_{L \in \mathcal{L}} \bigwedge_{\ell \in L} \ell$ that make this formula true. Consequently, p_0 has a nonempty extension if the entire considered QBF is true.

For data complexity, P membership follows from expressibility in Datalog shown in Theorem 6, for which the problem is known to be in P [23]. P hardness follows from the respective result for MODEQs established in Theorem 3. □

Proofs for Section 5

In this section, we establish the proof of Theorem 10, which we repeat here for convenience.

Theorem 10 (Deciding Datalog Containment in MSO). *Consider a Datalog query $\langle \mathtt{goal}, \mathbb{P} \rangle$ and an MSO query φ. If the models of φ are closed under homomorphism, then it is decidable if the query $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ.*

We establish this by transferring earlier results of Courcelle [20]. To do this in a self-contained way, we recall the main notions used in said work, and relate them to our present formalism. Further details and examples can be found in [20]. The results in [20] only apply to Datalog queries and MSO formulae without free variables and without constant symbols. The next two results show how one can lift this to the general case.

Lemma 23. *For every Datalog query $\langle \mathtt{goal}, \mathbb{P} \rangle$ and MSO query φ, one can construct in linear time a Boolean Datalog query $\langle \mathtt{goal}', \mathbb{P}' \rangle$ and Boolean MSO query φ' such that $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ iff $\langle \mathtt{goal}', \mathbb{P}' \rangle$ is contained in φ'.*

Proof. Let $\langle \mathtt{goal}, \mathbb{P} \rangle$ and φ be as in the claim. Assume that φ has $\mathsf{ar}(\mathtt{goal})$ free first-order variables and no free second-order variables – if this is not the case, then query containment is trivially impossible and one can easily find queries that satisfy the claim.

We reduce the containment problem to a containment problem of Boolean queries. To this end, let $\mathtt{result}$ be a fresh predicate symbol with $\mathsf{ar}(\mathtt{result}) = \mathsf{ar}(\mathtt{goal})$. We define a new Boolean Datalog query $\langle \mathtt{goal}', \mathbb{P}' \rangle$ by introducing a new nullary IDB predicate $\mathtt{goal}'$ and setting

$$\mathbb{P}' \coloneqq \mathbb{P} \cup \{\mathtt{goal}(\boldsymbol{y}) \wedge \mathtt{result}(\boldsymbol{y}) \to \mathtt{goal}'\}.$$

We define a new Boolean MSO query φ' by setting $\varphi' \coloneqq \exists \boldsymbol{x}. \varphi[\boldsymbol{x}] \wedge \mathtt{result}(\boldsymbol{x})$, where $\boldsymbol{x}$ is the list of free variables of φ. It is easy to see that $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ (over all databases with signature $\mathbf{P}$) if and only if $\langle \mathtt{goal}', \mathbb{P}' \rangle$ is contained in φ' (over all databases with signature $\mathbf{P} \cup \{\mathtt{result}\}$). □

Lemma 24. *For every Datalog query $\langle \mathtt{goal}, \mathbb{P} \rangle$ and MSO query φ, one can construct in linear time a Datalog query $\langle \mathtt{goal}, \mathbb{P}' \rangle$ and MSO query φ' that do not contain constant symbols such that $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ iff $\langle \mathtt{goal}, \mathbb{P}' \rangle$ is contained in φ'.*

Proof. Let $\langle \mathtt{goal}, \mathbb{P} \rangle$ and φ be as in the claim. For every constant $c \in \mathbf{C}$, let o_c denote a fresh unary predicate symbol, and let x_c be a fresh variable.

The set $\mathbb{P}'$ is defined to contain, for every rule $\psi_b \to \psi_h \in \mathbb{P}$, the rule $\psi_b' \wedge \bigwedge_{c \in \mathbf{C}} o_c(x_c) \to \psi_h' \in \mathbb{P}$, where ψ_b' and ψ_h' are obtained from ψ_b and ψ_h, respectively, by replacing each $c \in \mathbf{C}$ by x_c.

To define φ', let $\hat{\varphi}$ denote the formula obtained by replacing every $c \in \mathbf{C}$ by x_c. Let ψ be the conjunction of all formulae of form $o_c(x) \wedge o_c(x') \wedge p(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_{\mathsf{ar}(p)}) \to p(x_1, \dots, x_{i-1}, x', x_{i+1}, \dots, x_{\mathsf{ar}(p)})$ where $c \in \mathbf{C}$, $p \in \mathbf{P}$, and

$i \in \{1, \dots, \mathsf{ar}(p)\}$. We define φ' to be the formula

$$\forall \boldsymbol{x_c}. \left(\left(\bigwedge_{c \in \mathbf{C}} (\exists x. o_c(x) \wedge o_c(x_c)) \wedge \forall \boldsymbol{x}. \psi \right) \rightarrow \hat{\varphi} \right),$$

where $\boldsymbol{x_c}$ is the list of all variables x_c with $c \in \mathbf{C}$, and $\boldsymbol{x}$ is the list of all variables in ψ.

By this definition, the premise of φ' requires that each o_c is non-empty, and that all elements in o_c (if there is more than one) are indistinguishable by predicates in $\mathbf{P}$ ($*$). For databases that do not satisfy ($*$), the premise is false, and φ' entails all possible answers, and thus certainly contains $\langle \mathtt{goal}', \mathbb{P}' \rangle$.

Every interpretation $\mathcal{I}'$ over the new signature (with predicates o_c) that satisfies ($*$) induces an interpretation $\mathcal{I}$ of the original signature (with constants $\mathbf{C}$), obtained by identifying all domain elements in $o_c^{\mathcal{I}'}$ with $c^{\mathcal{I}}$. Note that it is possible for two constants to have the same interpretation. Conversely, every interpretation $\mathcal{I}$ over the original signature induces an interpretation $\mathcal{I}'$ over the new signature with $o_c^{\mathcal{I}'} = \{c^{\mathcal{I}}\}$. It is easy to see that the answers of $\langle \mathtt{goal}, \mathbb{P} \rangle$ over $\mathcal{I}$ are contained in the answers of φ over $\mathcal{I}$ if and only if the answers of $\langle \mathtt{goal}, \mathbb{P}' \rangle$ over $\mathcal{I}'$ are contained in the answers of φ' over $\mathcal{I}'$, as required. □

In the remainder of this section, we thus consider only Boolean queries that do not contain constant symbols.

Essential to the discussion in [20] are specific notions of (hyper)graphs and a monadic second-order logic that we will introduce first. These definitions depend on a set $\mathbf{P}$ of predicates; constant symbols, considered in our earlier signatures, do not feature here.

Definition 14. *A* concrete graph over $\mathbf{P}$ *is a tuple of the form* $\langle V, E, \mathsf{lab}, \mathsf{vert} \rangle$ *where V is a set of vertices, E is a set of edges,* $\mathsf{lab} : E \to \mathbf{P}$ *is an edge labelling function, and* vert *is a total function from edges $e \in E$ to tuples* $\mathsf{vert}(e)$ *of length* $\mathsf{ar}(\mathsf{lab}(e))$.

Given graphs G_1 and G_2, a homomorphism $h : G_1 \to G_2$ *is a pair of mappings $h_{\mathbf{v}} : V_1 \to V_2$ and $h_{\mathbf{e}} : E_1 \to E_2$ such that* $\mathsf{lab}_2 = h_{\mathbf{e}} \circ \mathsf{lab}_1$ *and* $\mathsf{vert}_2 = h_{\mathbf{v}} \circ \mathsf{vert}_1$ *(where we apply $h_{\mathbf{v}}$ to a tuple of vertices by applying it to each component). An* isomorphism *is a bijective homomorphism.*

Two concrete graphs are isomorphic if there is an isomorphism between them, and this defines an equivalence relation on concrete graphs. A graph *is an equivalence class of this relation, i.e., a maximal set of isomorphic concrete graphs over $\mathbf{P}$ (and some base set of vertex names). A homomorphism between graphs G_1 and G_2 is a homomorphism from a concrete graph in G_1 to a concrete graph in G_2.*

Abstracting from concrete graphs is relevant when generating expansion graphs from Datalog queries below, where the naming of vertices should not have any impact. For most practical purposes, one can still work with some concrete graph that represents a graph. Courcelle further considers graphs that have distinguished nodes, called *sources*, which we do not require here (i.e., we consider only graphs with an empty list of sources).

To state properties of graphs logically, Courcelle uses a specific notion of two-sorted monadic second-order logic.

Definition 15. *The multi-sorted monadic second-order logic* MSO_G *is defined over two sorts: the* vertex sort $\mathbf{v}$ *and the* edge sort $\mathbf{e}$. *Given a set of predicate symbols $\mathbf{P}$, the set $\mathbf{P}_G$ of two-sorted predicates is defined as* $\mathbf{P}_G := \{\mathsf{edg}_p \mid p \in \mathbf{P}\}$ *where* $\mathsf{ar}(\mathsf{edg}_p) = \langle \mathbf{e}, \mathbf{v}, \dots, \mathbf{v} \rangle$ *with* $\mathsf{ar}(p)$ *occurrences of* $\mathbf{v}$ *(recall that arities in multi-sorted logic are tuples of sorts).*

An MSO_G formula *(*MSO_G interpretation*) over a signature of (unsorted) predicate symbols $\mathbf{P}$ is a formula (interpretation) of two-sorted monadic second-order logic over the signature of predicate symbols $\mathbf{P}_G$. The term* unsorted formula *(*unsorted interpretation*) is used to emphasize that a formula (interpretation) is not MSO_G.*

In the following, we consider the predicate signature $\mathbf{P}$ to be fixed and do not mention it explicitly. MSO_G interpretations $\mathcal{J}$ can be considered as graphs G in an obvious way, with each tuple $\langle \epsilon, \boldsymbol{\delta} \rangle \in \mathsf{edg}_p^{\mathcal{J}}$ corresponding to a p-labelled edge between the vertices $\boldsymbol{\delta}$ in (any concrete graph representing) G. We will freely switch between these perspectives. Both can capture structures with multiple edges that have the same label and the same list of vertices. By forgetting the multiplicity of edges, we can transform graphs into unsorted interpretations:

Definition 16. *For an MSO_G interpretation $\mathcal{J}$, the unsorted interpretation $F(\mathcal{J})$ is defined as follows:*

- $\Delta^{F(\mathcal{J})} := \Delta_{\mathbf{v}}^{\mathcal{J}}$,
- *for all $p \in \mathbf{P}$ and $\boldsymbol{\delta} \in (\Delta_{\mathbf{v}}^{\mathcal{J}})^{\mathsf{ar}(p)}$, set $\boldsymbol{\delta} \in p^{F(\mathcal{J})}$ iff $\langle \epsilon, \boldsymbol{\delta} \rangle \in \mathsf{edg}_p^{\mathcal{J}}$ for some $\epsilon \in \Delta_{\mathbf{e}}^{\mathcal{J}}$.*

Given a homomorphism of graphs $h : \mathcal{J}_1 \to \mathcal{J}_2 = \langle h_{\mathbf{v}}, h_{\mathbf{e}} \rangle$, the homomorphism of interpretations $F(h) : F(\mathcal{J}_1) \to F(\mathcal{J}_2)$ is defined as $F(h) := h_{\mathbf{v}}$.

Conversely, for an unsorted interpretation $\mathcal{I}$, the MSO_G interpretation $G(\mathcal{I})$ is defined as follows:

- $\Delta_{\mathbf{v}}^{G(\mathcal{I})} := \Delta^{\mathcal{I}}$,
- $\Delta_{\mathbf{e}}^{G(\mathcal{I})} := \{\langle p, \boldsymbol{\delta} \rangle \mid p \in \mathbf{P}, \boldsymbol{\delta} \in p^{\mathcal{I}}\}$,
- *for all $p \in \mathbf{P}$, set* $\mathsf{edg}_p^{G(\mathcal{I})} := \{\langle \epsilon, \boldsymbol{\delta} \rangle \mid \epsilon = \langle p, \boldsymbol{\delta} \rangle \in \Delta_{\mathbf{e}}^{G(\mathcal{I})}\}$.

Given a homomorphism of interpretations $h : \mathcal{I}_1 \to \mathcal{I}_2$, the homomorphism of graphs $G(h) : G(\mathcal{I}_1) \to G(\mathcal{I}_2)$ consists of the mappings $G(h)_{\mathbf{v}} := h$ and $G(h)_{\mathbf{e}}$, defined as $\langle p, \boldsymbol{\delta} \rangle \mapsto \langle p, h(\boldsymbol{\delta}) \rangle$ (applying h to the tuple $\boldsymbol{\delta}$ component-wise).

It is easy to verify that F and G are well-defined, in particular that $F(h)$ and $G(h)$ are indeed homomorphisms. Algebraically speaking, F and G thus define *functors* between the categories of unsorted interpretations (with their homomorphisms) and graphs (with their homomorphisms). G has also been defined in [20, Definition 3.2]. The following is immediate from the definition:

Lemma 25. *$\mathcal{I} = F(G(\mathcal{I}))$ for all interpretations $\mathcal{I}$.*

An MSO_G formula φ is *insensitive to multiple edges* if $\mathcal{J} \models \varphi$ if and only if $G(F(\mathcal{J})) \models \varphi$. We can also transform MSO formulae to MSO_G formulae:

Definition 17. *Given an MSO formula φ, the MSO_G formula $g(\varphi)$ is obtained by the following replacements:*

- *replace each unsorted (first-order or second-order) variable υ in φ by a fresh (first-order or second-order) variable υ' of sort* **v***;*
- *replace each atomic subformula $p(\boldsymbol{x})$ in φ with $p \in \mathbf{P}$ by the formula $\exists z.\mathsf{edg}_p(z, \boldsymbol{x})$ where z is a fresh first-order variable of sort* **e***.*

Lemma 26. *For all MSO_G interpretations $\mathcal{J}$ and MSO formulae φ, we have $\mathcal{J} \models g(\varphi)$ iff $F(\mathcal{J}) \models \varphi$.*

Proof. Every unsorted variable assignment $\mathcal{Z}$ corresponds to a sorted variable assignment $g(\mathcal{Z})$ defined over variables of sort **v** in an obvious way. The following is an easy consequence of the definitions: for every unsorted atom $\varphi = p(\boldsymbol{x})$ and variable assignment $\mathcal{Z}$, we have $\mathcal{J}, g(\mathcal{Z}) \models g(\varphi)$ iff $F(\mathcal{J}), \mathcal{Z} \models \varphi$ $(*)$. The equivalence $(*)$ can be extended to arbitrary unsorted formulae φ with a straightforward induction. From this, the claim follows. Note that all free variables in $g(\varphi)$ are of sort **v** by definition, so it suffices to restrict attention to variable assignments of the form $g(\mathcal{Z})$ here. □

Lemma 27. *Every formula of the form $g(\varphi)$ is insensitive to multiple edges.*

Proof. Given any MSO_G interpretation $\mathcal{J}$, by Lemma 26, $\mathcal{J} \models g(\varphi)$ iff $F(\mathcal{J}) \models \varphi$. By Lemma 25, the latter is equivalent to $F(G(F(\mathcal{J}))) \models \varphi$. This is equivalent to $G(F(\mathcal{J})) \models g(\varphi)$, again by Lemma 26. Hence $\mathcal{J} \models g(\varphi)$ iff $G(F(\mathcal{J})) \models g(\varphi)$. □

Lemma 28. *If the models of φ are closed under homomorphisms, then so are the models of $g(\varphi)$.*

Proof. Consider a homomorphism $h : \mathcal{J}_1 \to \mathcal{J}_2$ between MSO_G interpretations, and assume that $\mathcal{J}_1 \models g(\varphi)$. By Lemma 26, $\mathcal{J}_1 \models g(\varphi)$ implies $F(\mathcal{J}_1) \models \varphi$. By Definition 16, $F(h) : F(\mathcal{J}_1) \to F(\mathcal{J}_2)$ is a homomorphism of interpretations. Since the models of φ are closed under homomorphisms, $F(\mathcal{J}_1) \models \varphi$ implies $F(\mathcal{J}_2) \models \varphi$. By Lemma 26, $\mathcal{J}_2 \models g(\varphi)$, as claimed. □

This completes our analysis of the relationship between our notion of MSO and Courcelle's MSO_G defined on graphs in the sense of Definition 14. We can relate Datalog queries to graphs as follows:

Definition 18. *Let $\langle \mathtt{goal}, \mathbb{P} \rangle$ be a Boolean Datalog query. A* partial expansion *is a conjunction of first-order atoms, with the set of partial expansions of $\langle \mathtt{goal}, \mathbb{P} \rangle$ defined inductively as follows:*

- *The atom propositional* goal() *is a partial expansion.*
- *If there is a partial expansion of form $p(\boldsymbol{y}) \wedge \varphi$ and a rule $\psi \to p(\boldsymbol{x}) \in \mathbb{P}$, then $\psi\sigma\theta \wedge \varphi\theta$ is a partial expansion, where σ is a renaming that bijectively maps variables in $\psi \to p(\boldsymbol{x})$ to fresh variable names not used in $p(\boldsymbol{y}) \wedge \varphi$ yet, and θ is a most-general unifier of $p(\boldsymbol{x})\sigma$ and $p(\boldsymbol{y})$.*

A (complete) expansion *of $\langle \mathtt{goal}, \mathbb{P} \rangle$ is a partial expansion of $\langle \mathtt{goal}, \mathbb{P} \rangle$ that contains no IDB predicates.*

Every expansion corresponds to a concrete graph that has variables for its vertices and atoms labelled by predicates for its edges. This induces a set *$L(\mathtt{goal}, \mathbb{P})$* of expansion graphs *of $\langle \mathtt{goal}, \mathbb{P} \rangle$.*

Expansion graphs are called *computation graphs* in [20]. It is well-known that every Datalog query is equivalent to the (infinite) disjunction of its expansions, where all variables are existentially quantified. Restating this in terms of expansion graphs and universal models, we obtain:

Lemma 29. *A database D entails a Boolean Datalog query $\langle \mathtt{goal}, \mathbb{P} \rangle$ iff there is some $G \in L(\mathtt{goal}, \mathbb{P})$ for which there is a homomorphism $G \to G(\mathcal{I}(D))$.*

Using the above terminology, we can reformulate part of Theorem 5.5 of [20] as follows:

Theorem 30 (Courcelle). *If φ is an MSO_G formula that is insensitive to multiple edges and whose models are closed under homomorphisms, then it is decidable whether $G \models \varphi$ for all $G \in L(\mathtt{goal}, \mathbb{P})$.*

Note that we identify graphs and MSO_G interpretations when writing $G \models \varphi$.

Lemma 31. *Theorem 10 holds for Boolean queries without constant symbols.*

Proof. Let $\langle \mathtt{goal}, \mathbb{P} \rangle$ and φ be as in the claim. We claim that $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ if and only if $G \models g(\varphi)$ for all $G \in L(\mathtt{goal}, \mathbb{P})$ $(*)$. By Lemma 28, the models of $g(\varphi)$ are closed under homomorphisms, and, by Lemma 27, $g(\varphi)$ is insensitive to duplicate edges. Hence, by Theorem 30, $(*)$ is decidable.

It remains to show that query containment is equivalent to $(*)$. For the "if" direction, assume that $(*)$ holds. Consider an arbitrary database D that entails $\langle \mathtt{goal}, \mathbb{P} \rangle$. By Lemma 29, this is equivalent to the existence of an expansion graph $G \in L(\mathtt{goal}, \mathbb{P})$ for which exists a homomorphism $G \to G(\mathcal{I}(D))$. By $(*)$ we have $G \models g(\varphi)$. Therefore, $G(\mathcal{I}(D)) \models g(\varphi)$ by Lemma 28. By Lemma 26, $\mathcal{I}(D) \models \varphi$, and thus D entails φ. Since D was arbitrary, this shows the claimed query containment.

For the "only if" direction, assume that $\langle \mathtt{goal}, \mathbb{P} \rangle$ is contained in φ. Let $G \in L(\mathtt{goal}, \mathbb{P})$ be arbitrary. By Definition 18, $F(G)$ satisfies the expansion of $\langle \mathtt{goal}, \mathbb{P} \rangle$ that induced G. Therefore, $F(G) \models \mathbb{P} \to \mathtt{goal}$. It is easy to see that the (finite) interpretation $F(G)$ is equal to $\mathcal{I}(D)$ for a database D. Thus D entails $\langle \mathtt{goal}, \mathbb{P} \rangle$, and, by the assumed containment, D also entails φ. Hence, $\mathcal{I}(D) = F(G) \models \varphi$ Thus, by Lemma 26, $G \models g(\varphi)$. Since G was arbitrary, this shows $(*)$. □

Proof of Theorem 10. According to Lemmas 23 and 24, the containment problem can be reduced to the containment problem of Boolean queries without constants. It is important to note that the constructions in both cases preserve the closure of an MSO formula under homomorphisms. The claim then follows from Lemma 31. □

Proofs for Section 6

THEOREM 13 (NEMODEQ ANSWERING UNDER BTS). *Let D be a database and let Σ be a set of rules for which the treewidth of $\mathcal{I}(D \cup \Sigma)$ is bounded. Let $\mathfrak{Q}[\boldsymbol{x}]$ be a NEMODEQ.*

1. *$D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ if and only if $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ has no countable model with bounded treewidth.*
2. *If Conjecture 12 holds, the problem $D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ is decidable.*

PROOF. We start by proving the first claim. For the "only if" direction, it is obvious that $D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ implies that $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ can have no model (and therefore no model with bounded treewidth).

For proving the "if" direction, we assume that $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ has no countable model with bounded treewidth. Then it must be unsatisfiable for the following reason: toward a contradiction assume it has a model $\mathcal{I}$. Since $\mathcal{I} \models D \cup \Sigma$, there must be a homomorphism from $\mathcal{I}(D \cup \Sigma)$ to $\mathcal{I}$. Thus, by contraposition of the closedness of models of $\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ under homomorphisms (Theorem 7), $\mathcal{I}(D \cup \Sigma)$ itself satisfies $\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ and is therefore a model of $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$. Moreover, since Σ is bts, we obtain that $\mathcal{I}(D \cup \Sigma)$ has bounded treewidth. Obviously it is also countable, which yields the desired contradiction and ensures unsatisfiability of $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$.

For the second claim we start from the previous claim and note that, since D and Σ are first-order theories and $\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ is expressible as an MSO formula, $D \cup \Sigma \cup \{\neg\mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]\}$ is expressible as an MSO theory. Thus, we can invoke Conjecture 12 to obtain the desired result. □

Proofs for Section 7

For the following considerations we capitalize on the fact that a set of Datalog rules can be viewed as a (possibly infinite) collection of conjunctive queries that are obtained by expanding rules by repeated backward-chaining (cf. Definition 18). Every expansion $\varphi[\boldsymbol{x}]$ with variables $\boldsymbol{x}$ can naturally be associated with an interpretation structure $\mathcal{I}(\varphi)$: its domain $\Delta^{\mathcal{I}(\varphi)}$ is $\boldsymbol{x} \cup \mathbf{C}$, for each constant $c \in \mathbf{C}$ we set $c^{\mathcal{I}(\varphi)} := c$, and we have $\boldsymbol{t} \in p^{\mathcal{I}(\varphi)}$ exactly if $p(\boldsymbol{t})$ occurs in φ.

LEMMA 14 (REPLACEMENT LEMMA). *Consider a set Σ of TGDs, a conjunctive query $Q = \exists\boldsymbol{y}.\psi[\boldsymbol{x},\boldsymbol{y}]$, and a NEMODEQ $\mathfrak{Q}[\boldsymbol{x}]$ that is a rewriting for Σ and Q. Then Q and $\mathfrak{Q}$ are equivalent in all models of Σ, i.e., $\Sigma \models \forall\boldsymbol{x}.Q[\boldsymbol{x}] \leftrightarrow \mathfrak{Q}[\boldsymbol{x}]$.*

Let $\psi[\boldsymbol{t}/\boldsymbol{x},\boldsymbol{y}'/\boldsymbol{y}]$ be the conjunction of Q with variables $\boldsymbol{x}$ replaced by terms $\boldsymbol{t}$ and variables $\boldsymbol{y}$ replaced by variables $\boldsymbol{y}'$. We say that $\psi[\boldsymbol{t}/\boldsymbol{x},\boldsymbol{y}'/\boldsymbol{y}]$ is a match *in a Datalog rule ρ if ρ is of the form $\psi[\boldsymbol{t}/\boldsymbol{x},\boldsymbol{y}'/\boldsymbol{y}] \wedge \varphi \rightarrow \chi$ where the $\boldsymbol{y}'$ occur neither in φ nor in χ.*

Given some NEMODEQ $\mathfrak{P}[\boldsymbol{z}]$ over Σ, let $\mathfrak{P}'[\boldsymbol{z}]$ denote a NEMODEQ obtained by replacing a match $\psi[\boldsymbol{t}/\boldsymbol{x},\boldsymbol{y}'/\boldsymbol{y}]$ of Q in some rule of $\mathfrak{P}$ by $\mathfrak{Q}[\boldsymbol{t}/\boldsymbol{x}]$, where we assume w.l.o.g. that the bound variables in $\mathfrak{Q}$ do not occur in $\mathfrak{P}$. Then $\mathfrak{P}$ and $\mathfrak{P}'$ are equivalent in all models of Σ, i.e., $\Sigma \models \forall\boldsymbol{z}.\mathfrak{P}[\boldsymbol{z}] \leftrightarrow \mathfrak{P}'[\boldsymbol{z}]$.

PROOF. We first show that $\Sigma \models \forall\boldsymbol{x}.Q[\boldsymbol{x}] \leftrightarrow \mathfrak{Q}[\boldsymbol{x}]$. For the one direction, consider a model $\mathcal{I} \models \Sigma$ and a variable assignment $\mathcal{Z}$ such that $\mathcal{I}, \mathcal{Z} \models \mathfrak{Q}[\boldsymbol{x}]$. According to Theorem 6, there is an expansion $\varphi[\boldsymbol{y}]$ of $\mathsf{datalog}(\mathfrak{Q})$ such that $\mathcal{I}, \mathcal{Z} \models \varphi[\boldsymbol{y}]$ (where we assume w.l.o.g. that $\mathcal{Z}$ assigns the appropriate domain elements to the fresh variables that φ may contain). Using notation as in Section 5, we find a model $\mathcal{I}(\varphi)$ to which φ matches under the variable assignment $\mathcal{Z}_{\mathrm{id}}$ with $\mathcal{Z}_{\mathrm{id}}(y) = y$ for each y in φ. Then $\mathcal{I}(\varphi), \mathcal{Z}_{\mathrm{id}} \models \mathfrak{Q}[\boldsymbol{x}]$.

Let $D(\mathcal{I}(\varphi))$ be the model $\mathcal{I}(\varphi)$ considered as a database containing a fact for each of the finitely many relations in $\mathcal{I}(\varphi)$. Introducing finitely many new constants for this purpose is not a problem. Let $\boldsymbol{c_x}$ denote the constants in $D(\mathcal{I}(\varphi))$ that correspond to $\mathcal{Z}_{\mathrm{id}}(\boldsymbol{x})$. Then $D(\mathcal{I}(\varphi)) \models \mathfrak{Q}[\boldsymbol{c_x}/\boldsymbol{x}]$.

Since $\mathfrak{Q}$ is a rewriting of Q under Σ, we have $D(\mathcal{I}(\varphi)), \Sigma \models Q[\boldsymbol{c_x}/\boldsymbol{x}]$. Consider a universal model $\mathcal{J}$ of $D(\mathcal{I}(\varphi)), \Sigma$. Then $\mathcal{J} \models Q[\boldsymbol{c_x}/\boldsymbol{x}]$. Moreover, there is a homomorphism from $\mathcal{J}$ to $\mathcal{I}$. Indeed, the mapping $\mathcal{Z}$ induces a homomorphism π from $\mathcal{I}(\varphi)$ to $\mathcal{I}$. This mapping can be extended to a homomorphism π' from $\mathcal{J}$ to $\mathcal{I}$, since $\mathcal{I}$ is a model of Σ. Due to Theorem 7, the query match $\mathcal{J} \models Q[\boldsymbol{c_x}/\boldsymbol{x}]$ implies $\mathcal{J} \models Q[\pi'(\boldsymbol{c_x})/\boldsymbol{x}]$. Since $\pi'(\boldsymbol{c_x}) = \pi(\boldsymbol{c_x}) = \mathcal{Z}(\boldsymbol{x})$, this shows the claim $\mathcal{I}, \mathcal{Z} \models Q[\boldsymbol{x}]$.

The other direction can be shown in a similar way, somewhat simplified due to the fact that one does not need to construct an intermediate model $\mathcal{J}$ of Σ to obtain the match for $\mathfrak{Q}$.

Now the rest of the claim follows from Theorem 6. It remains to show the claimed equivalence for $\mathsf{datalog}(\mathfrak{P})$ and $\mathsf{datalog}(\mathfrak{P}')$. This is a direct consequence of the Replacement Theorem of first-order logic that allows us to replace the sub-formula $\exists\boldsymbol{y}'.\psi[\boldsymbol{t}/\boldsymbol{x},\boldsymbol{y}'/\boldsymbol{y}]$ by $p_{\mathfrak{Q}}(\boldsymbol{t})$, both of which have just shown to be equivalent. □

THEOREM 15 (SINGLE PREDICATE REWRITING CORRECTNESS). *If Σ is j-oriented and p is a head predicate, then $\mathfrak{Q}_{p,\Sigma}[\boldsymbol{z}]$ is a rewriting for Σ and $p(\boldsymbol{z})$.*

PROOF. We consider the Datalog rewriting $\mathsf{datalog}(\mathfrak{Q}_{p,\Sigma})$ of Definition 7. By Theorem 6, it suffices to show that, for every database D and list of constants $\boldsymbol{c} = c_1, \ldots, c_n$, we have $D \cup \Sigma \models p(\boldsymbol{c})$ iff $D, \mathsf{datalog}(\mathfrak{Q}_{p,\Sigma}) \models p_{\mathfrak{Q}_{p,\Sigma}}(\boldsymbol{c})$.

To establish this, we show that, for every database D, list of constants $\boldsymbol{c} = c_1, \ldots, c_n$, and predicate $\hat{U}_q$ of $\mathsf{datalog}(\mathfrak{Q}_{p,\Sigma})$, we have $D \cup \Sigma \models q(\boldsymbol{c})$ if and only if $D \cup \mathsf{datalog}(\mathfrak{Q}_{p,\Sigma}) \models \hat{U}_q(c_j, c_1, \ldots, c_n)$. The proof is by an easy induction over the derivation of $q(\boldsymbol{c})$. Clearly, $\mathsf{datalog}(\mathfrak{Q}_{p,\Sigma}) \models \hat{V}_i(\boldsymbol{d})$ iff $\boldsymbol{d}$ is of the form $d_i, d_1, \ldots, d_i, \ldots, d_n$. Moreover, whenever there is a rule $\psi \rightarrow q(t_1, \ldots, t_n) \in \Sigma$ that has an instance $\psi_c \rightarrow q(c_1, \ldots, c_n)$, then $\mathsf{datalog}(\mathfrak{Q}_{p,\Sigma})$ contains a rule $\psi' \rightarrow \hat{U}_q(t_j, t_1, \ldots, t_n)$ with an instance $\psi'_c \rightarrow \hat{U}_q(c_j, c_1, \ldots, c_n)$. It is easy to verify the claim. □

THEOREM 16 (j-ORIENTEDNESS IMPLIES REWRITABILITY). *Every j-oriented rule set is MODEQ-rewritable.*

PROOF. A rewriting for a j-oriented rule set Σ and a CQ $Q[\boldsymbol{x}] = \exists\boldsymbol{y}.p_1(\boldsymbol{t_1}) \wedge \ldots \wedge p_m(\boldsymbol{t_m})$ is obtained from $Q[\boldsymbol{x}]$ by replacing all head atoms $p_i(\boldsymbol{t_i})$ with the MODEQ $\mathfrak{Q}_{p_i,\Sigma}[\boldsymbol{t_i}/\boldsymbol{x}]$ as in Definition 11. We assume a fixed sequence of replacement steps in the construction of the rewriting. Let $\mathfrak{Q}_0$ denote the original query $Q[\boldsymbol{x}]$, let $\mathfrak{Q}_i$ with $1 \leq i$ denote the

result after each subsequent replacement step, and let $\mathfrak{Q}$ be the final result.

One can show $\Sigma \models \forall \boldsymbol{x}.Q[\boldsymbol{x}] \leftrightarrow \mathfrak{Q}_i[\boldsymbol{x}]$ by induction over i. The base case follows since $\mathfrak{Q}_0$ is clearly equivalent to Q. The induction steps follow from Lemma 14 and Theorem 15.

Hence $\Sigma \models \forall \boldsymbol{x}.Q[\boldsymbol{x}] \leftrightarrow \mathfrak{Q}[\boldsymbol{x}]$, i.e., for every interpretation $\mathcal{I} \models \Sigma$ and every variable assignment $\mathcal{Z}$ for $\mathcal{I}$, we have $\mathcal{I}, \mathcal{Z} \models Q[\boldsymbol{x}]$ iff $\mathcal{I}, \mathcal{Z} \models \mathfrak{Q}[\boldsymbol{x}]$ $(*)$.

We show the condition of Definition 9. Thus consider an arbitrary database D and a potential query answer $\boldsymbol{c}$. Without loss of generality, we assume that D contains no fact of the form $q'(\boldsymbol{d})$ for some head predicate q' of Σ. Indeed, if it does, we can replace q' by a fresh predicate q'_D and add a rule $\forall \boldsymbol{x}.q'_D(\boldsymbol{x}) \rightarrow q'(\boldsymbol{x})$ to Σ. This modification clearly preserves j-orientedness.

If $D \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ then clearly $D \cup \Sigma \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$ and thus $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$ by $(*)$. Conversely, assume that $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$. Then $D \cup \Sigma$ is satisfiable since Σ contains no constraints (rules with empty head). More precisely, for every interpretation $\mathcal{I} \models D$, there is an interpretation $\mathcal{I}' \models \Sigma, D$ that coincides with $\mathcal{I}$ on all constants and all predicates that are not head predicates in Σ, where we use that D contains no head predicates of Σ. By the assumption $\mathcal{I}' \models Q[\boldsymbol{c}/\boldsymbol{x}]$. By $(*)$ $\mathcal{I}' \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$. Besides the auxiliary unary predicates $\mathtt{U}$, $\mathfrak{Q}$ contains only predicates for which $\mathcal{I}$ and $\mathcal{I}'$ agree. Hence $\mathcal{I} \models \mathfrak{Q}[\boldsymbol{c}/\boldsymbol{x}]$. Since $\mathcal{I}$ was arbitrary, this establishes the claim. □

Proofs for Section 8

Lemma 18 (Query rewriting with cuts). *Let D be a database and let $\Sigma_1 \triangleright \Sigma_2$ be two sets of TGDs.*

1. *If $\mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{x}]$ is a NEMODEQ-rewriting of a conjunctive query $Q[\boldsymbol{x}]$, then:*

   $D \cup \Sigma_1 \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$ if and only if $D \cup \Sigma_1 \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$.

2. *If Σ_1 and Σ_2 are NEMODEQ-rewritable, then so is $\Sigma_1 \cup \Sigma_2$.*

Proof. We start by proving the first claim. For the "only if" direction, assume $D \cup \Sigma_1 \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$. By Theorem 17 there is a conjunctive query Q' with $D \cup \Sigma_1 \models Q'$ and $Q' \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$. By the definition of rewritability, we obtain $Q' \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$. Due to $D \cup \Sigma_1 \models Q'$, we obtain $D \cup \Sigma_1 \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$.

For the "if" direction, we assume $D \cup \Sigma_1 \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$ and conclude $\mathcal{I}(D \cup \Sigma_1) \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$, which in turn means that there must be an expansion $\varphi[\boldsymbol{y}]$ of $\mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$ such that $\mathcal{I}(D \cup \Sigma_1) \models \exists \boldsymbol{y}.\varphi[\boldsymbol{y}]$. Due to universality of $\mathcal{I}(D \cup \Sigma_1)$ and homomorphism-closedness of models of $\exists \boldsymbol{y}.\varphi[\boldsymbol{y}]$, we thus find $D \cup \Sigma_1 \models \exists \boldsymbol{y}.\varphi[\boldsymbol{y}]$. Since $\exists \boldsymbol{y}.\varphi[\boldsymbol{y}] \models \mathfrak{Q}_{Q,\Sigma_2}[\boldsymbol{c}/\boldsymbol{x}]$, we can conclude $\{\exists \boldsymbol{y}.\varphi[\boldsymbol{y}]\} \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$ by the definition of rewriting. Combining these observations, we obtain $D \cup \Sigma_1 \cup \Sigma_2 \models Q[\boldsymbol{c}/\boldsymbol{x}]$.

Consider any conjunctive query $Q[\boldsymbol{x}]$ and its rewriting $\mathfrak{Q}_{Q,\Sigma_2}$. For every rule ρ in $\mathfrak{Q}_{Q,\Sigma_2}$, the conjunction of all body atoms that do not use any of the auxiliary unary predicates $\mathtt{U}$ can be considered as a conjunctive query Q', that contains no existentially quantified variables. This query constitutes a match for ρ in the sense of Lemma 14. By this lemma, we can thus replace Q' by its rewriting $\mathfrak{Q}_{Q',\Sigma_1}$. The query $\mathfrak{Q}_{Q,\Sigma_1 \cup \Sigma_2}$ is obtained by performing this replacement in all rules of $\mathfrak{Q}_{Q,\Sigma_2}$. It is not hard to see that $\mathfrak{Q}_{Q,\Sigma_1 \cup \Sigma_2}$ is a rewriting for Q and $\Sigma_1 \cup \Sigma_2$, using a similar argument as in the proof of Theorem 16. □

Theorem 19 (Query answering with cuts). *Consider rule sets $\Sigma_1 \triangleright \Sigma_2$, such that NEMODEQ answering is decidable under Σ_1, and NEMODEQ rewriting is decidable under Σ_2. Then NEMODEQ answering is decidable under $\Sigma_1 \cup \Sigma_2$.*

Proof. This is a direct consequence of Lemma 18 (1). Given any conjunctive query Q, we can compute the rewriting $\mathfrak{Q}_{Q,\Sigma_2}$, and compute its answers over $D \cup \Sigma_1$. □

Theorem 20 (Fully oriented rule sets are rewritable). *For a rule set Σ, it can be detected in polynomial time if Σ is fully oriented. Every fully oriented set Σ is MODEQ-rewritable.*

Proof. Clearly, the relation $\approx_{\ll}$ can be constructed in polynomial time by checking $\rho \ll \rho'$ for each pair of rules and constructing the reflexive symmetric transitive closure. The equivalence classes $[\rho]_{\ll}$ are obtained from this in linear time. It is clear that j-orientedness can be checked for each set of rules in polynomial time.

It remains to show the second part of the claim. We say that $[\rho]_{\ll}$ is *maximal* if, for all $\rho' \in \Sigma$, we have that $\rho \ll \rho'$ implies $\rho' \in [\rho]_{\ll}$. If Σ is fully oriented and $[\rho]_{\ll}$ is maximal, then $\Sigma_1 := \Sigma \setminus [\rho]_{\ll}$ and $\Sigma_2 := [\rho]_{\ll}$ are such that $\Sigma_1 \bowtie \Sigma_2$, and thus $\Sigma_1 \triangleright \Sigma_2$. Moreover, Σ_1 again is fully oriented. Thus, one can apply Theorem 16 and Lemma 18 (2) to obtain the required rewriting. □

Theorem 21 (NEMODEQs = fully oriented Datalog). *For every NEMODEQ $\mathfrak{Q}$, the rule set $\mathsf{datalog}(\mathfrak{Q})$ of Definition 7 is fully oriented. Moreover, for every NEMODEQ-rewritable set Σ of TGDs, there is a fully oriented set of Datalog rules Σ' such that:*

- *every predicate p in Σ has a corresponding head predicate q_p in Σ' that does not occur in Σ,*
- *for every database D and conjunctive query $Q[\boldsymbol{x}]$ that do not contain predicates of the form q_p, and for every list of constants $\boldsymbol{c}$, we have $D \cup \Sigma \models Q[\boldsymbol{c}/\boldsymbol{x}]$ iff $D \cup \Sigma' \models Q'[\boldsymbol{c}/\boldsymbol{x}]$ where Q' is obtained from Q by replacing all predicates p with q_p.*

Proof. The first part of the claim follows by induction over the degree d of $\mathfrak{Q}$. The case of $d = 1$ follows by observing that the rules with head predicate $p_{\mathfrak{Q}}$ in $\mathsf{datalog}(\mathfrak{Q})$ cannot be $\prec$-smaller than any other rule, and thus form a maximal j-oriented (for any position j in $p_{\mathfrak{Q}}$) subset of $\mathsf{datalog}(\mathfrak{Q})$. Likewise, the set of rules with head of the form $\hat{U}_i(y, \boldsymbol{z})$ is clearly 1-oriented. To show the claim for $d > 1$, we observe that no rule in $\mathsf{datalog}(\mathfrak{Q})$ is $\prec$-smaller than any rule in $\mathsf{datalog}(\mathfrak{Q}')$ for some subquery $\mathfrak{Q}'$ of $\mathfrak{Q}$. The claim follows by induction.

For the second part of the claim, let $Q_p[\boldsymbol{y}]$ be the CQ $p(\boldsymbol{y})$ for each predicate p that is not of the form $q_{p'}$. There is a rewriting $\mathfrak{Q}_p[\boldsymbol{y}]$ for $Q_p[\boldsymbol{y}]$ and Σ. By Lemma 14, $\Sigma \models \forall \boldsymbol{y}.\mathfrak{Q}_p[\boldsymbol{y}] \leftrightarrow Q_p[\boldsymbol{y}]$. From Theorem 6 follows $\models \forall \boldsymbol{y}.\mathfrak{Q}_p[\boldsymbol{y}] \leftrightarrow$

$(\mathsf{datalog}(\mathfrak{Q}_p) \to p_{\mathsf{datalog}(\mathfrak{Q}_p)}(\boldsymbol{y}))$. Thus, $\Sigma \models \forall \boldsymbol{y}.Q_p[\boldsymbol{y}] \leftrightarrow (\mathsf{datalog}(\mathfrak{Q}_p) \to p_{\mathsf{datalog}(\mathfrak{Q}_p)}(\boldsymbol{y}))$.

Using arguments as in the proof of Theorem 16, we find that $D \models \forall \boldsymbol{y}.(\Sigma \to Q_p[\boldsymbol{y}]) \leftrightarrow (\mathsf{datalog}(\mathfrak{Q}_p) \to p_{\mathsf{datalog}(\mathfrak{Q}_p)}(\boldsymbol{y}))$ whenever $D \cup \Sigma$ is consistent. To cover the case that $D \cup \Sigma$ is inconsistent, let $\mathfrak{Q}_\bot$ be a MODEQ without free variables such that, for all databases D', $D' \cup \Sigma$ is inconsistent iff $D' \models \mathfrak{Q}_\bot$. To find such a $\mathfrak{Q}_\bot$, consider $\mathfrak{Q}_p$ for a predicate p that does not occur in Σ and delete from $\mathsf{datalog}(\mathfrak{Q}_p)$ all rules that use the predicate p (these rules check for occurrences of p in the input database). $\mathfrak{Q}_\bot$ can easily be obtained from this.

By the first part of the claim, $\mathsf{datalog}(\mathfrak{Q}_p)$ and $\mathsf{datalog}(\mathfrak{Q}_\bot)$ are fully oriented. We set $q_p := p_{\mathsf{datalog}(\mathfrak{Q}_p)}$. Let $\Sigma_\bot$ be the fully oriented rule set obtained from $\mathsf{datalog}(\mathfrak{Q}_\bot)$ by replacing each rule $\forall \boldsymbol{y}.\varphi \to$ that has an empty head by new rules $\forall \boldsymbol{x}, \boldsymbol{y}.\varphi \to q_p(\boldsymbol{x})$ for each of the predicates q_p where $\boldsymbol{x}$ is a list of fresh variables of the appropriate length. Thus, $\Sigma_\bot$ entails all possible facts over predicates q_p from D whenever $D \cup \Sigma$ is inconsistent.

Now we can set $\Sigma' := \Sigma_\bot \cup \bigcup_p \mathsf{datalog}(\mathfrak{Q}_p)$ where we assume w.l.o.g. that any two $\mathsf{datalog}(\mathfrak{Q}_p)$ and $\mathsf{datalog}(\mathfrak{Q}_{p'})$ use mutually disjoint sets of head predicates. It is easy to verify that the claim is satisfied for this choice. □