The Limits of Efficiency for Open- and Closed-World Query Evaluation Under Guarded TGDs

Pablo Barceló
IMC, PUC Chile & IMFD Chile
pbarcelo@ing.puc.cl

Victor Dalmau
Universitat Pompeu Fabra
victor.dalmau@upf.edu

Cristina Feier
University of Bremen
feier@uni-bremen.de

Carsten Lutz
University of Bremen
clu@uni-bremen.de

Andreas Pieris
University of Edinburgh
apieris@inf.ed.ac.uk

ABSTRACT

Ontology-mediated querying and querying in the presence of constraints are two key database problems where tuple-generating dependencies (TGDs) play a central role. In ontology-mediated querying, TGDs can formalize the ontology and thus derive additional facts from the given data, while in querying in the presence of constraints, they restrict the set of admissible databases. In this work, we study the limits of efficient query evaluation in the context of the above two problems, focussing on guarded and frontier-guarded TGDs and on UCQs as the actual queries. We show that a class of ontology-mediated queries (OMQs) based on guarded TGDs can be evaluated in FPT iff the OMQs in the class are equivalent to OMQs in which the actual query has bounded treewidth, up to some reasonable assumptions. For querying in the presence of constraints, we consider classes of constraint-query specifications (CQSs) that bundle a set of constraints with an actual query. We show a dichotomy result for CQSs based on guarded TGDs that parallels the one for OMQs except that, additionally, FPT coincides with PTime combined complexity. The proof is based on a novel connection between OMQ and CQS evaluation. Using a direct proof, we also show a similar dichotomy result, again up to some reasonable assumptions, for CQSs based on frontier-guarded TGDs with a bounded number of atoms in TGD heads. Our results on CQSs can be viewed as extensions of Grohe’s well-known characterization of the tractable classes of CQs (without constraints). Like Grohe’s characterization, all the above results assume that the arity of relation symbols is bounded by a constant. We also study the associated meta problems, i.e., whether a given OMQ or CQS is equivalent to one in which the actual query has bounded treewidth.

1 INTRODUCTION

Tuple-generating dependencies (TGDs) are a prominent rule-based formalism at the core of several areas that are of central importance to databases and artificial intelligence. They are first-order implications of the form \( \forall x \forall y \phi(x,y) \rightarrow \exists z \psi(x,z) \), where \( \phi \) and \( \psi \) are conjunctions of relational atoms, and they essentially state that some tuples (facts) in a relational instance imply the presence of some other tuples in that instance (hence the name “tuple-generating”). In query evaluation over relational databases, TGDs have two facets: (1) they can be used as ontology axioms that allow us to derive additional facts from the given data, which supports more complete query answers, and (2) they can be used as integrity constraints that restrict the set of admissible databases, which paves the way to constraint-aware query optimization. In fact, TGDs were originally introduced as a unifying framework for the large range of relational constraints introduced in the 1970s and the 1980s [1].

The Two Facets of TGDs. When a set of TGDs is used as an ontology, they are often bundled with an actual database query, typically a union of conjunctive queries (UCQ), to form a composite ontology-mediated query (OMQ) [11, 12]. As discussed above, adding the ontology serves the purpose of delivering more complete answers to queries. It also enriches the vocabulary available for querying as it may introduce new relation symbols that are not part of the data signature. An OMQ language is a pair \((\mathcal{C}, \mathcal{Q})\), where \(\mathcal{C}\) is a class of TGDs and \(\mathcal{Q}\) a query language, which collects all the OMQs in which the ontology is formulated in \(\mathcal{C}\) and the actual query comes from \(\mathcal{Q}\) [12]. The main problem of concern for an OMQ language \((\mathcal{C}, \mathcal{Q})\) is OMQ evaluation under an open-world semantics, i.e., we are looking for answers to the actual query from \(\mathcal{Q}\) that are logically entailed by the input database and the ontology from \(\mathcal{C}\).

On the other hand, when a set of TGDs is used as relational integrity constraints, the TGDs together with the actual query can be bundled together into what we call a constraint-query specification (CQS). A class of CQSs is a pair \((\mathcal{C}, \mathcal{Q})\), where, as in OMQs, \(\mathcal{C}\) is a class of TGDs and \(\mathcal{Q}\) a query language. In this setting, however, the problem of interest is CQS evaluation under a closed-world semantics, i.e., we are looking for answers to the query from \(\mathcal{Q}\) over the input database, which is promised to comply with the set of constraints from \(\mathcal{C}\). In other words, we directly evaluate the query over the database, which we know to satisfy the set of constraints.

TGDs and Guardedness. It is well-known that OMQ evaluation in \((\text{TGD}, \text{UCQ})\), where \text{TGD} is the class of arbitrary TGDs, and \text{UCQ} the class of union of conjunctive queries, is an undecidable
problem; see, e.g., [14]. This has led to an intensive research activity for identifying restrictions on TGDs that ensure the decidability of OMQ evaluation; see, e.g., [3, 14, 16, 29] – the list is by no means exhaustive. One of the most robust syntactic paradigms that emerged from this extensive effort, which is central to our work, is guardedness. A TGD is guarded if the left-hand of the implication, called the body, has an atom that contains (or guards) all the universally quantified variables [14]; let $G$ be the class of guarded TGDs. A natural generalization is to guard only the variables that appear also in the right-hand side, called the head, which leads to the class of frontier-guarded TGDs $FG$ [3]. The complexity of OMQ evaluation in $(C, UCQ)$, where $C \in \{G, FG\}$, is by now well-understood. In both cases, it is $2ExpTime$-complete [2, 14], and it remains hard even in the case of bounded-arity schemas, i.e., $2ExpTime$-complete in $(FG, UCQ)$ [2], and $ExpTime$-complete in $(G, UCQ)$ [14].

Although guarded and frontier-guarded TGDs have been proposed as ontology languages, they can also naturally serve as classes of integrity constraints [9]. Note, for example, that the important class of referential integrity constraints (or inclusion dependencies) is a very special case of guarded TGDs. Now, the complexity of CQS evaluation in $(C, UCQ)$, where $C \in \{G, FG\}$, coincides with that of query evaluation for UCQ: it is $NP$-complete, even for schemas of bounded arity [17]. This holds since CQS evaluation in $(C, UCQ)$ is, in fact, a refinement of the query evaluation problem for UCQ, with the additional promise that the input database satisfies a given (but potentially empty) set of integrity constrains coming from $C$.

The Limits of Efficiency. Thus both OMQ evaluation and CQS evaluation in $(C, UCQ)$, where $C \in \{G, FG\}$, are computationally hard problems. However, there are subclasses of $(C, UCQ)$, seen either as an OMQ language or a class of CQSs, for which the evaluation problem is efficient in the sense of being tractable or being fixed-parameter tractable (FPT) where the parameter is the size of the OMQ (resp., CQS). In particular, with $UCQ_k$ being the class of CQSs of treewidth at most $k$, we know that: (1) for each $k \geq 1$, OMQ evaluation in $(G, UCQ_k)$ is in FPT (actually, this is shown in this work – Proposition 3.3), and (2) for each $k \geq 1$, CQS evaluation in $(FG, UCQ_k)$ is in PTime. Both statements rely on the well-known result that query evaluation for UCQs of bounded treewidth is tractable [18]. In view of this, it is natural to ask whether we can precisely characterize the limits of efficient OMQ and CQS evaluation. The goal of this work is to provide such efficiency characterizations.

A seminal result by Grohe precisely characterizes the recursively enumerable classes of CQs over schemas of bounded arity that can be evaluated in polynomial time (under the assumption that $FPT \neq W[1]$) [26]: this is the case if and only if for some $k \geq 1$, every CQ in the class is equivalent to a CQ of treewidth $k$. Grohe’s result also establishes that PTIME and FPT coincide for evaluating classes of CQs. The above can be naturally generalized to UCQs.

Efficiency characterizations in the same spirit have been recently obtained for OMQ languages based on description logics (DLs). In particular, [7] precisely characterizes the recursively enumerable classes of OMQs from $(ELH^L, UCQ)$, where $ELH^L$ is an important DL, essentially a fragment of guarded TGDs. Here, OMQ evaluation is in $FPT$ (assuming $FPT \neq W[1]$) if and only if for some $k \geq 1$, every OMQ in the class is equivalent to an OMQ of treewidth $k$, i.e., an OMQ such that the UCQ in it is of treewidth $k$. Note that the equivalence is now on the level of OMQs rather than on the level of UCQs. The same work [7] precisely characterizes the recursively enumerable classes of OMQs from $(ELH^L, UCQ)$, where $ELH^L$ is a key fragment of $ELH^I$ that underpins the OWL 2 EL profile of the OWL 2 recommendation [32]. Here, evaluation is in PTIME (again assuming $FPT \neq W[1]$) if and only if for some $k \geq 1$, every OMQ in the class is equivalent to an OMQ of treewidth $k$. This also shows that PTIME and FPT coincide for evaluating classes of OMQs from $(ELH^L, UCQ)$. The classification of $(ELH^L, UCQ)$, however, is subject to the condition that the ontology does not introduce relations beyond those admitted in the database.

Our Results. Our main results are as follows, under the widely believed complexity-theoretic assumption that $FPT \neq W[1]$:

1. For a recursively enumerable class of OMQs from $(G, UCQ)$ over a schema of bounded arity, evaluation is in FPT if and only if there is $k \geq 1$ such that each OMQ in the class is equivalent to one from $(G, UCQ_k)$ (Theorem 5.3).

2. For a recursively enumerable class of CQSs from $(G, UCQ)$ over a schema of bounded arity, evaluation is in PTIME if and only if evaluation is in FPT if and only if there is $k \geq 1$ such that each CQS in the class is equivalent to one from $(G, UCQ_k)$ (Theorem 5.7).

3. Our final result concerns CQSs from $(FG, UCQ)$ in which the TGDs have a bounded number of at most $m \geq 1$ head atoms; let $FG_m$ be the obtained class. For a recursively enumerable class of CQSs from $(FG_m, UCQ)$ over a schema of bounded arity, evaluation is in PTIME if and only if evaluation is in FPT if and only if there is $k \geq 1$ such that each CQS in the class is equivalent to one from $(FG_m, UCQ_k)$ (Theorem 5.12).

It is not surprising that characterization (1) talks only about FPT and not PTIME complexity since we know from [14] that the ExpTime-hardness of OMQ evaluation in $(G, UCQ)$ in the case of schemas of bounded arity holds even if the actual query is an atomic query of the simplest form, i.e., a propositional atom. Moreover, it turns out that the notion of being equivalent to an OMQ of bounded treewidth is not enough for characterizing evaluation in FPT for frontier-guarded TGDs. We can show that OMQ evaluation in $(FG, UCQ_k)$ is in $W[1]$-hard – this is actually easily inherited from the fact that Boolean CQ evaluation is $W[1]$-hard [33].

For all the above characterizations, if the efficiency condition fails, i.e., there is no integer $k \geq 1$ such that each OMQ or CQS in the considered class is equivalent to one in which the actually query falls within $UCQ_k$, then the evaluation problem is $W[1]$-hard. Showing these lower bounds is actually the most challenging task underlying our efficiency characterizations. Together with the assumption that $FPT \neq W[1]$, they establish that efficient evaluation implies the efficiency condition. For (1) and (3), this is done via an fpt-reduction from the parameterized version of the $k$-clique problem, a well-known $W[1]$-hard problem, building on Grohe’s result which also relies on an fpt-reduction from $k$-clique. For (2), we exploit a novel connection between OMQ evaluation and CQS evaluation, which is of independent interest. In fact, we provide an fpt-reduction from OMQ evaluation to CQS evaluation.

At this point, we would like to stress that the fpt-reductions underlying (1) and (3) are not merely adaptations of the fpt-reduction
by Grohe. For (1), the fact that the ontology can introduce additional relations beyond the data signature causes serious challenges. Moreover, unlike the characterizations of [7] for OMQs based on DLs, where only unary and binary relations are used, we need to deal with relations of arity beyond two, which also causes additional non-trivial complications that require novel ideas and techniques. For (3), unlike Grohe, we are more constrained in defining the right database as it must satisfy the given set of constraints. Moreover, the useful notion of core of a CQ, which was crucial for Grohe’s proof, cannot be directly used in the presence of constraints.

We further study the complexity of the associated meta problems of deciding whether, for some fixed $k$, a given OMQ or CQS is equivalent to one in which the actual query falls within $UCQ_k$. We show that in all the considered cases the problem is $2EXP$-TIME-complete (under the mild assumption that $k$ is at least the maximum arity of the occurring relation symbols, minus one). Note that decidability of the meta problem is needed to prove the lower bounds described above, but we also consider it interesting in its own right.

2 PRELIMINARIES

We consider the disjoint countably infinite sets $C$ and $V$ of constants and variables, respectively. We refer to constants and variables as terms. For an integer $n \geq 1$, we may write $[n]$ for the set $\{1, \ldots, n\}$.

Relational Databases. A (relational) schema $S$ is a finite set of relation symbols (or predicates) with associated arity. We write $ar(R)$ for the arity of a predicate $R$, and $ar(S)$ for the arity of $S$, that is, the number $\max_{R \in S}ar(R)$. An atom over $S$ is an expression of the form $R(\bar{t})$, where $R \in S$ and $\bar{t}$ is an $(ar(R))$-tuple of terms. An instance over $S$, or simply S-instance, is (a possibly infinite) set of atoms over $S$ that contain only constants, while a database over $S$, or simply S-database, is a finite $S$-instance. We write $dom(I)$ for the set of constants in an instance $I$. For a set $T \subseteq dom(I)$, we denote by $I_T$ the restriction of $I$ to atoms that mention only constants of $T$. A homomorphism from $I$ to an instance $J$ is a function $h : dom(I) \rightarrow dom(J)$ such that $R(h(i)) \in J$ for every $R(i) \in I$. We write $I \rightarrow J$ for the fact that there is a homomorphism from $I$ to $J$.

Conjunctive Queries. A conjunctive query (CQ) over a schema $S$ is a first-order formula of the form $\exists \bar{y} (R_1(\bar{x}_1) \land \ldots \land R_m(\bar{x}_m))$, where each $R_i(\bar{x}_i)$, for $i \in [m]$, is an atom over $S$ that contains only variables, each variable mentioned in the $\bar{x}_i$s appears either in $\bar{x}$ or $\bar{y}$, and $\bar{x}$ contains all the free variables of $q$ called the answer variables. Every CQ $q$ can be naturally seen as a database $D[q]$, known as the canonical database of $q$, obtained by dropping the existential quantifier prefix and viewing variables as constants. We may simply write $q$ instead of $D[q]$. A homomorphism from a CQ $q$ to an instance $I$ is a homomorphism from $D[q]$ to $I$. A tuple $\bar{c} \in dom(I)^{|S|}$ is an answer to $Q$ over $I$ if there is a homomorphism $h$ from $q$ to $I$ with $h(\bar{x}) = \bar{c}$. The evaluation of $q(\bar{x})$ over $I$, denoted $q(I)$, is the set of all answers to $Q$ over $I$. We write $CQ$ for the class of CQs. A union of conjunctive queries (UCQ) over a schema $S$ is a first-order formula of the form $q(\bar{x}) : = q_1(\bar{x}) \lor \cdots \lor q_n(\bar{x})$, where $n \geq 1$, and $q_i(\bar{x})$, for $i \in [n]$, is a CQ over $S$. The evaluation of $q$ over an instance $I$, denoted $q(I)$, is defined as the set of tuples $\bigcup_{i \in [n]} q_i(I)$. We write $UCQ$ for the class of UCQs. The arity of a (UC)Q is defined as the number of its answer variables. A (UC)Q of arity zero is called Boolean, and it can have as an answer only the empty tuple. For a Boolean (UCQ) $q$, we may write $I \models q$, if $q(I) = \{(\)}$, and $I \not\models q$, otherwise.

Treewidth. A central notion in our work is that of treewidth, which measures the degree of tree-likeness of a graph. Let $G = (V, E)$ be an undirected graph. A tree decomposition of $G$ is a pair $(T, \chi)$, where $T = (T_0, E_T)$ is a tree, and $\chi$ is a labeling function $\chi : V \rightarrow 2^{T_0}$, i.e., $\chi$ assigns a subset of vertices to each node of $T_0$, such that:

1. $\bigcup_{t \in T_0} \chi(t) = V$.
2. If $\{u, v\} \in E$, then $u, v \in \chi(t)$ for some $t \in T_0$.
3. For each $v \in V$, the set of nodes $\{t \in T_0 \mid v \in \chi(t)\}$ induces a connected subtree of $T_0$.

The width of $\delta$ is the number $\max_{t \in T_0} (|\chi(t)| - 1)$. If the edge-set $E$ of $G$ is non-empty, then the treewidth of $G$ is the minimum width over all its tree decompositions; otherwise, it is defined to be one. Each instance $I$ is associated with an undirected graph (without self loops) $G^I = (V, E)$, called the Gaifman graph of $I$. Defined as follows: $V = dom(I)$, and for $\{a, b\} \in E$ iff there is an atom $R(i) \in I$ that mentions both $a$ and $b$. The treewidth of $I$ is the treewidth of $G^I$.

For $k \geq 1$, $TW_k$ is the class of instances of treewidth at most $k$.

The evaluation problem for (U)CQs takes as input a (UC)Q $q(\bar{x})$, a database $D$, and a candidate answer $\bar{c}$, and asks whether $\bar{c} \in q(D)$. It is well-known that (U)CQ evaluation is NP-complete [17]. On the other hand, it becomes tractable by restricting the syntactic shape of CQs. One of the most widely studied such restrictions is bounded treewidth. Formally, a CQ $q(\bar{x}) = \exists \bar{y} \phi(\bar{x}, \bar{y})$ has treewidth $k \geq 1$ if $G^q_{\emptyset}$ has treewidth $k$, where $G^q_{\emptyset}$ is the subgraph of $G^q$ induced by the elements of $\emptyset$. Note that the treewidth of $q$ is defined in a more liberal way than usual. The standard definition considers the treewidth of $G^q$, while here the treewidth of $q$ is only measured with respect to the subgraph of $G^q$ induced by its existentially quantified variables. A UCQ $q$ has treewidth $k$ if each of its disjunctions has treewidth at most $k$. We write $\text{UCQ}_k$ (resp., $\text{UCQ}_{\emptyset}$) for the class of CQs (resp., UCQs) of treewidth at most $k$.

The next well-known result illustrates the usefulness of bounding the treewidth.

Proposition 2.1 ([18]). Fix $k \geq 1$. Given a database $D$, an $n$-ary query $q \in \text{UCQ}_k$, and a tuple $\bar{c} \in dom(D)^n$, the problem of deciding whether $\bar{c} \in q(D)$ can be solved in time $O(|D|^{k+1} \cdot ||q||)$.

Tuple-generating Dependencies. A tuple-generating dependency (TGD) $\sigma$ over $S$ is a constant-free first-order sentence of the form $\forall \bar{x} \forall \bar{y} \phi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})$, where $\phi$ is a possibly empty conjunction of atoms over $S$, while $\psi$ is a non-empty conjunction of atoms over $S$. For simplicity, we write $\sigma$ as $\phi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})$, and use comma instead of $\land$ for joining atoms. We call $\phi$ and $\psi$ the body and head of $\sigma$, denoted $\text{body}(\sigma)$ and $\text{head}(\sigma)$, respectively. The frontier of $\sigma$, denoted $\text{fr}(\sigma)$, is the set of variables $\bar{x}$, i.e., the variables that appear both in the body and the head of $\sigma$. The TGD $\sigma$ above is logically equivalent to the expression $\forall \bar{x} (q_\phi(\bar{x}) \rightarrow q_\psi(\bar{x}))$, where $q_\phi(\bar{x})$ and $q_\psi(\bar{x})$ are the CQs $\exists \bar{y} \phi(\bar{x}, \bar{y})$ and $\exists \bar{z} \psi(\bar{x}, \bar{z})$, respectively. Therefore, an instance $I$ over $S$ satisfies $\sigma$, denoted $I \models \sigma$, if $q_\phi(I) \subseteq q_\psi(I)$. An instance $I$ satisfies a set $\Sigma$ of TGDs, denoted $I \models \Sigma$, if $I \models \sigma$ for each $\sigma \in \Sigma$. Henceforth, whenever we refer to a set of TGDs we mean a finite set. We write TGD for the class of TGDs, that is, the family of all possible sets of TGDs.

1As usual, given a syntactic object $O$, we write $||O||$ for its size.
Frontier-Guardedness. A TGD $\sigma$ is guarded if either body($\sigma$) is empty, or there exists an atom $\alpha$ in its body that contains all the variables occurring in body($\sigma$) [14]. Such an atom $\alpha$ is the guard of $\sigma$, denoted guard($\sigma$). We write $\varnothing$ for the class of guarded TGDs. A natural generalization of guardedness is frontier-guardedness, where only the frontier variables must be guarded. Formally, a TGD $\sigma$ is frontier-guarded if either body($\sigma$) is empty or there exists an atom $\alpha$ in its body that contains all the variables of fr($\sigma$) [3], which we call again guard and denote as guard($\sigma$). The class of frontier-guarded TGDs is denoted by FG. Clearly, $\varnothing \subseteq FG \subseteq TGD$.

The Chase Procedure. The chase is a useful tool when reasoning with TGDS [14, 22, 28, 30]. We first define a single chase step. Let $I$ be an instance over a schema $S$ and $\sigma$ a TGD of the form $\phi(x, y) \rightarrow \exists z \psi(x, z)$ over $S$. We say that $\sigma$ is applicable with respect to $I$, if there exists a tuple $(c, c')$ of constants in $I$ such that $\phi(c, c') \subseteq I$. In this case, the result of applying $\sigma$ over $I$, with $(c, c')$ is the instance $J = I \cup \{ \psi(c, c') \}$, where $c'$ is the tuple obtained from $z$ by simultaneously replacing each variable $z$ with a fresh constant distinct not occurring in $I$. For such a single chase step we write $I \xrightarrow{\sigma, (c, c')} J$.

Let $I$ be an instance and $\Sigma$ a set of TGDS. A chase sequence for $I$ under $\Sigma$ is a sequence of chase steps $I_0 \xrightarrow{\sigma_0, (c_0, c'_0)} I_1 \xrightarrow{\sigma_1, (c_1, c'_1)} \ldots$ such that (1) $I_0 = I$, (2) $\sigma_i \in \Sigma$ for each $i \geq 0$, and (3) $J \models \Sigma$, where $J = \bigcup_{i \geq 0} I_i$. The instance $J$ is the (potentially infinite) result of this chase sequence, which always exists. Since we consider the oblivious chase, i.e., a TGD is triggered whenever its body is satisfied no matter whether its head is satisfied, every chase sequence for $I$ under $\Sigma$ leads to the same result (up to isomorphism). Thus, we can refer to the result of the chase for $I$ under $\Sigma$, denoted chase($I, \Sigma$). The key property of the chase follows:

**Proposition 2.2.** Consider an instance $I$ and a set $\Sigma$ of TGDS. For every instance $J$ such that $J \supseteq I$ and $J \models \Sigma$, chase($I, \Sigma$) $\rightarrow$ $J$ via a homomorphism that is the identity on dom($I$).

Parameterized Complexity. Parameterized complexity has a central role in our work. A parameterized problem over an alphabet $\Lambda$ is a pair $(P, \kappa)$, with $P \subseteq \Lambda^*$ a decision problem and $\kappa$ a parameterization of $P$, that is, a parameterized computational function $k : \Lambda^* \rightarrow \mathbb{N}$. A prime example is $p$-Clique, where $P$ is the set of all pairs $(G, k)$ with $G$ an undirected graph that contains a $k$-clique and $\kappa(G, k) = k$.

A problem $(P, \kappa)$ is fixed-parameter tractable (fpt) if there is a computable function $f : \mathbb{N} \rightarrow \mathbb{N}$ and an algorithm that decides $P$ in time $|x|^f(O(1)) \cdot f(k(x))$, where $x$ denotes the input. We use FPT to denote the class of all fixed-parameter tractable problems that are fixed-parameter tractable. Notice that FPT corresponds to a relaxation of the usual notion of tractability: if $P$ is in PTIME, then $(P, \kappa)$ is in FPT for any $\kappa$, but the latter might also be the case when $P$ is NP-hard.

An fpt-reduction from a problem $(P_1, \kappa_1)$ over $\Lambda_1$ to a problem $(P_2, \kappa_2)$ over $\Lambda_2$ is a function $\rho : \Lambda_1^* \rightarrow \Lambda_2^*$ such that, for some computable functions $f : \mathbb{N} \rightarrow \mathbb{N}$,

1. $x \in P_1$ iff $\rho(x) \in P_2$, for all $x \in \Sigma_1^*$;
2. $\rho(x)$ is computable in time $|x|^f(O(1)) \cdot f(k_1(x))$, for $x \in \Sigma_1^*$;
3. $k_2(\rho(x)) \leq g(k_1(x))$, for all $x \in \Sigma_1^*$.

An important parameterized complexity class is W[1], defined in terms of fpt-reductions. It is generally believed that FPT $\neq$ W[1], the status of this problem being comparable to that of PTIME $\neq$ NP. Hence, if a parameterized problem $(P, \kappa)$ is W[1]-hard, then $(P, \kappa)$ is not fpt unless FPT = W[1]. A well-known W[1]-hard problem is p-Clique [21].

3 THE TWO FACETS OF TGDS IN QUERYING

As discussed in Section 1, TGDS can be used as:

1. *Ontology axioms* that enrich incomplete data with domain knowledge, which leads to more complete answers.
2. *Integrity constraints* that specify semantic properties satisfied by all databases, which paves the way to constraint-aware query optimization techniques.

Depending on how TGDS are used, we obtain different evaluation problems, which we now formalize.

3.1 TGDS as Ontology Axioms

When a set of TGDS is used as ontology axioms, or an ontology, it is typically seen, together with the actual query, as one composite query, called ontology-mediated query. Formally, an ontology-mediated query (OMQ) is a triple $Q = (S, \Sigma, q)$, where $S$ is a schema, called data schema, which indicates that $Q$ will be evaluated over $S$-databases, $\Sigma$ is a set of TGDS over an extended schema $T \supseteq S$, called ontology, and $q$ is a UCQ over $T$. In case $S = T$, we say that $Q$ has full data schema. The arity of $Q$ is defined as the arity of $q$. We write $Q(x)$ to emphasize that the answer variables of $q$ are $x$, and say that $Q$ is over $T$ to mean that the extended schema of $Q$ is $T$.

The evaluation of an OMQ $Q(x) = (S, \Sigma, q(x))$ over an $S$-database $D$ consists of all the tuples that are answers to $q$ over each model of $D$ and $\Sigma$. A model of $D$ and $\Sigma$ is an instance $I$ such that $I \supseteq D$ and $I \models \Sigma$. A tuple $\bar{e} \in \text{dom}(D)^{|x|}$ is an answer to $Q$ over $D$ if $\bar{e} \in q(I)$ for each model $I$ of $D$ and $\Sigma$. The evaluation of $Q(x)$ over $D$, denoted $Q(D)$, is the set of all answers to $Q$ over $D$. By Proposition 2.2, and the monotonicity of UCQS, we get the following useful result:

**Proposition 3.1.** For every OMQ $Q = (S, \Sigma, q(x)) \in (\text{TGD}, \text{UCQ})$, $Q(D) = q(\text{chase}(D, \Sigma))$.

We write $(C, Q)$ for the class of OMQs, called OMQL, in which the ontology is formulated in the class of TGDS $\mathcal{C}$, and the actual query is coming from the class of queries $\mathcal{Q}$; for example, we may write $(\mathcal{C}, \mathcal{Q})$, $(\mathcal{FG}, \mathcal{UCQ})$, $(\mathcal{C}, \mathcal{UCQ})$, for some $k \geq 1$, etc. This brings us to the following question problem for OMQL:

**PROBLEM:** OMQ-Evaluation($\mathcal{C}$)

**INPUT:** An OMQ $Q = (S, \Sigma, q(x)) \in \mathcal{C}$, an $S$-database $D$, and a tuple $\bar{e} \in \text{dom}(D)^{|x|}$

**QUESTION:** Is it the case that $\bar{e} \in Q(D)$?

We are also interested in the parameterized version of the above problem, dubbed p-OMQ-Evaluation($\mathcal{C}$), which is the same size of the OMQ $Q$, as customary in the literature [33]. Thus p-OMQ-Evaluation($\mathcal{C}$) is in FPT if it can be solved in time $|D|^f(O(1))$, for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

It is well-known that OMQ-Evaluation($\text{TGD}$, $\mathcal{C}$) is decidable; see, e.g., [14]. On the other hand, if we focus on OMQLs in which the ontology is formulated as a set of frontier-guarded TGDS, then the problem becomes decidable, in fact, it is 2ExpTime-complete [2]. At this point, one may wonder whether bounding the treewidth...
of the CQs will have the same positive effect as in the case of CQ evaluation (see Proposition 2.1). It is implicit in [14] that this is not the case. The next result summarizes some key facts about the evaluation problem for OMQs based on (frontier-)guarded TGDs.

**Proposition 3.2.** It holds that:

(1) OMQ-Evaluation($\emptyset$, UCQ) is 2ExpTime-complete even for schemas of bounded arity.

(2) OMQ-Evaluation($G$, UCQ) is 2ExpTime-complete, and becomes ExpTime-complete for schemas of bounded arity.

(3) For each $k \geq 1$, OMQ-Evaluation($G$, UCQ$_k$) is 2ExpTime-complete, and still ExpTime-hard for schemas of bounded arity.

Since the parameterized version of the evaluation problem for CQs is $W[1]$-hard [33], even for schemas of bounded arity, we can immediately conclude that p-OMQ-Evaluation($G$, UCQ) is $W[1]$-hard, even for schemas of bounded arity. Do we gain something if we focus on UCQs of bounded treewidth? The answer is negative for frontier-guarded TGDs, but affirmative for guarded TGDs.

**Proposition 3.3.** It holds that:

(1) p-OMQ-Evaluation($G$, UCQ) is $W[1]$-hard even for schemas of bounded arity.

(2) p-OMQ-Evaluation($\emptyset$, UCQ$_k$) is $W[1]$-hard even for schemas of bounded arity.

(3) For each $k \geq 1$, p-OMQ-Evaluation($G$, UCQ$_k$) is in FPT.

As discussed above, item (1) is a consequence of the fact that the parameterized version of CQ evaluation is $W[1]$-hard, even for schemas of bounded arity. Item (2) is a consequence of the fact that the parameterized version of CQ evaluation is $W[1]$-hard, even for Boolean CQs over a schema of bounded arity, while a Boolean CQ can be transformed into a frontier-guarded TGD. Indeed, given a Boolean CQ $\exists \phi(\bar{x}) \phi(\bar{x}) \rightarrow \text{Ans}$, where Ans is a 0-ary predicate, is trivially a frontier-guarded TGD since its frontier is empty.

We now briefly explain how item (3) is shown, while the details can be found in the appendix. A TGD is called linear if its frontier is empty. Grohe’s result follows:

**Theorem 4.1 (Grohe’s Theorem [26]).** Fix $r \geq 1$. Let $Q$ be a recursively enumerable class of CQs over schemas of arity $r$. The following are equivalent, assuming $\text{FPT} \neq \text{W}[1]$:  

(1) The evaluation problem for $Q$ is in PTime.

(2) The evaluation problem for $Q$ is in FPT.

(3) There is $k \geq 1$ such that $Q \subseteq \text{UCQ}_k^r$.

If either statement is false, then evaluation for $Q$ is $W[1]$-hard.

Interestingly, it is decidable whether a CQ is equivalent to one of treewidth $k$. This is shown by exploiting the notion of core.
Recall that the core of a CQ \( q \) is a \( \subseteq \)-minimal subquery of \( q \) that is equivalent to \( q \). It is known that, for each \( k \geq 1 \), a CQ \( q \) belongs to \( \mathbb{C}Q_k^p \) iff its core is in \( \mathbb{C}Q_k \), and that deciding this property is NP-complete [20]. There is also a natural generalization of this characterization and of Theorem 4.1 to the class of UCQs.

At this point, it is natural to ask whether it is possible to obtain a characterization of the classes of OMQs (resp., CQSs) based on (frontier-)guarded TGDs for which OMQ-Evaluation (resp., CQS-Evaluation) can be efficiently solved, in the same spirit as Grohe’s Theorem. In fact, our main question is whether the natural generalization of the notion of bounded treewidth modulo equivalence for CQs to OMQs (resp., CQSs) is a decidable notion that exhausts tractability or fixed-parameter tractability for OMQ-Evaluation (resp., CQS-Evaluation), as in the case of Grohe’s Theorem.

### 4.1 Semantic Tree-likeness for OMQs

We first concentrate on OMQs and introduce the notion of UCQk-equivalence, which essentially tells us that an OMQ can be rewritten into an equivalent one where the UCQ belongs to UCQk. Let us clarify that for OMQs the notion of equivalence is not applied at the level of the actual query, but to the whole OMQ. Formally, given two OMQs \( Q \) and \( Q' \), both with data schema \( S \), \( Q \) is contained in \( Q' \), written \( Q \subseteq Q' \), if \( Q(D) \subseteq Q'(D) \) for every \( S \)-database \( D \). We then say that \( Q \) and \( Q' \) are equivalent, denoted \( Q \equiv Q' \), if \( Q \subseteq Q' \) and \( Q' \subseteq Q \). In what follows, let \( C \) be a class of TGDs, i.e., \( C \subseteq \text{TGD} \).

**Definition 4.2 (UCQk-equivalence for OMQs).** An OMQ \( Q = (S, \Sigma, q) \in (C, \text{UCQ}) \) is UCQk-equivalent, for \( k \geq 1 \), if there exists an OMQ \( Q' = (S, \Sigma', q') \in (C, \text{UCQ}_k) \) such that \( Q \equiv Q' \). Given an OMQ language \( D = (C, \text{UCQ}) \), for each \( k \geq 1 \), let \( \text{O}_k \subseteq \text{UCQ}_k \)-equivalent.

According to the above definition, we are allowed to rewrite both the ontology and the UCQ. It is conceptually meaningful though to consider also the setting where the ontology cannot be altered. This leads to the uniform version of UCQk-equivalence.

**Definition 4.3 (Uniform UCQk-equivalence for OMQs).** An OMQ \( Q = (S, \Sigma, q) \in (C, \text{UCQ}) \) is uniformly UCQk-equivalent, for \( k \geq 1 \), if there is \( Q' = (S, \Sigma', q') \in (C, \text{UCQ}_k) \) such that \( Q \equiv Q' \). For an OMQ language \( D = (C, \text{UCQ}) \) and \( k \geq 1 \), let \( \text{O}_k \subseteq \text{UCQ}_k \)-equivalent.

The next example shows that both the ontology and the data schema can have an impact on the treewidth.

**Example 4.4.** We first illustrate that the ontology can have an impact on the treewidth. Consider the OMQ \( Q_1 = (S, \Sigma, q) \), where

\[
\begin{align*}
S &= \{ R_1, R_2, R_3, R_4, P \} \\
\Sigma &= \{ R_2(x) \rightarrow R_4(x) \} \\
q() &= P(x_2, x_1) \land P(x_4, x_3) \land R_1(x_3) \land R_2(x_3) \land R_3(x_3) \land R_4(x_4).
\end{align*}
\]

Note that \( S \) contains all the predicates in \( \Sigma \) and \( q \), while \( q \) is a Boolean CQ (the existential quantifiers are omitted). Observe that \( q \) is a core from \( \mathbb{C}Q_2 \), and thus can easily be seen not to belong to \( \mathbb{C}Q_1^p \). However, \( Q_1 \) is equivalent to the OMQ \( (S, \Sigma, q') \), where

\[
q'() = P(x_2, x_1) \land P(x_2, x_3) \land R_1(x_1) \land R_2(x_2) \land R_3(x_3).
\]

Since \( q' \in \mathbb{C}Q_1 \), we get that \( q_1 \in (G, \mathbb{U}CQ)_1^{\text{opt}} \).

We proceed to show that the data schema can also have an impact. Consider the OMQ \( Q_2 = (S', \Sigma', q) \) with full data schema, where

\[
\Sigma' = \{ S(x) \rightarrow R_1(x), S(x) \rightarrow R_3(x) \}.
\]

It is not hard to see that \( Q_2 \) does not belong to \( (G, \mathbb{U}CQ)_2^{\text{opt}} \). If, however, the predicate \( R_1 \) is omitted from the signature, then \( Q_2 \) is equivalent to the OMQ \( (S', P, R_2, R_4, \Sigma', q'') \), where

\[
q''() = P(x_2, x_1) \land P(x_2, x_3) \land R_2(x_2) \land R_3(x_1) \land R_4(x_4).
\]

And thus, it belongs to \( (G, \mathbb{U}CQ)_2^{\text{opt}} \).

### 4.2 Semantic Tree-likeness for CQSs

We now introduce UCQk-equivalence for CQSs. For this setting, only the uniform version is relevant. Indeed, given a CQS \( S = (\Sigma, q) \), by altering the set of integrity constraints \( \Sigma \), we essentially change the semantics of \( S \), which is, of course, not our intention. Given two CQSs \( S = (\Sigma, q) \) and \( S' = (\Sigma', q') \) over a schema \( T \), we say that \( S \) is contained in \( S' \), denoted \( S \subseteq S' \), if \( q(D) \subseteq q'(D) \) for every \( T \)-database \( D \) that satisfies \( \Sigma \). We then say that \( S \) is equivalent to \( S' \) if \( S \subseteq S' \) and \( S' \subseteq S \). We may also write \( q \equiv q' \) (resp., \( q \not\equiv q' \)) for the fact that \( S \not\subseteq S' \) (resp., \( S \equiv S' \)). We recall a well-known result about containment among CQSs, which will be useful for our analysis:

**Proposition 4.5.** Let \( S_1 = (\Sigma, q_1(x)) \) and \( S_2 = (\Sigma, q_2(y)) \) be CQSs from \( \text{TGD}, \text{UCQ} \) with \( |x| = |y| \). Then \( S_1 \subseteq S_2 \) iff for each \( p_i \in q_1 \), there exists \( p_2 \in q_2 \) such that \( p_2 \in \text{ch}(p_1, \Sigma) \).

The notion of uniform UCQk-equivalence for CQSs follows:

**Definition 4.6 (Uniform UCQk-equivalence for CQSs).** A CQS \( S = (\Sigma, q) \) from \( (C, \text{UCQ}) \) is uniformly UCQk-equivalent, for \( k \geq 1 \), if there exists a CQS \( S' = (\Sigma, q') \) from \( (C, \text{UCQ}_k) \) such that \( S \equiv S' \). Given a class of CQSs \( D \), for \( k \geq 1 \), let \( \Sigma_k \) be the class of all CQSs from \( D \) that are uniformly UCQk-equivalent.

Observe that the first part of Example 4.4 also works when \( (\Sigma, q) \) is viewed as a CQS. This illustrates the fact that integrity constraints can have an impact on the treewidth.

### 5 OUR RESULTS IN A NUTSHELL

Having the relevant notions in place, we can now provide an answer to our main question, that is, whether (uniform) UCQk-equivalence of OMQs (resp., uniform UCQk-equivalence of CQSs) exhausts tractability and fixed-parameter tractability of OMQ-Evaluation (resp., CQS-Evaluation) in the same spirit as Grohe’s Theorem. In this section, we give an overview of our results, observe interesting connections between the OMQ and CQS settings, and study the associated meta problems. Further details are deferred to the subsequent sections and the appendix.

### 5.1 The Guarded Case

**Ontology-mediated Queries.** We start with our results on OMQs based on guarded TGDs. We first ask whether (uniform) UCQk-equivalence is decidable for \( (G, \text{UCQ}) \).

---

1. We overload the notation again. It would be clear from the context when \( \mathbb{C}Q_k ^p \) is an OMQ language or a class of CQSs. We also avoid the superindex \( u \) since for CQSs we only consider the uniform version of UCQk-equivalence.
Theorem 5.1. For each \( k \), deciding whether a given OMQ \( Q \) from \((G, UCQ)\) over \( T \) with \( k \geq \text{ar}(T) - 1 \) is (uniformly) UCQ\(_k\)-equivalent is 2ExpTime-complete. If existant, an OMQ \( Q' \in (G, UCQ) \) such that \( Q \equiv Q' \) can be computed in double exponential time.

The above complexity result exploits a characterization of when an OMQ \( Q \) from \((G, UCQ)\) (is (uniformly) UCQ\(_k\)-equivalent, which in turn relies on what we call a UCQ\(_k\)-approximation, that is, approximations of \( Q \) from below in terms of an OMQ from \((G, UCQ)\).

In a nutshell, a UCQ\(_k\)-approximation of \( Q = (S, \Sigma, q) \) is an OMQ \( Q_k = (k, \Sigma, q_k) \), where \( q_k \) belongs to UCQ\(_k\), that behaves like \( Q \) over \( S \)-databases of treewidth at most \( k \), i.e., for every \( S \)-database \( D \) of treewidth at most \( k \), \( Q(D) = Q_k(D) \). It follows that \( Q_k \) is equivalent to \( Q \) if and only if \( Q \) is UCQ\(_k\)-equivalent. The formal definition of \( Q_k \) can be found in the appendix. It is significantly more involved than in the case of description logics [7] because there the chase only generates structures of treewidth one.

Proposition 5.2. Let \( Q \) be an OMQ from \((G, UCQ)\) over \( T \), and let \( k \geq \text{ar}(T) - 1 \). The following are equivalent:

1. \( Q \) is UCQ\(_k\)-equivalent.
2. \( Q \) is uniformly UCQ\(_k\)-equivalent.
3. \( Q \equiv Q_k \).

Since UCQ\(_k\)-equivalence and uniform UCQ\(_k\)-equivalence turn out to be equivalent, we henceforth only use UCQ\(_k\)-equivalence. Note that Proposition 5.2 also provides an approach to deciding UCQ\(_k\)-equivalence, and thus to establishing Theorem 5.1: compute the UCQ\(_k\)-approximation \( Q_k \) of \( Q \) and accept if \( Q \subseteq Q_k \) (note that \( Q_k \) holds always); otherwise, reject. We can show that \( Q_k \) can be computed in double exponential time and it is known that OMQ containment for \((G, UCQ)\) can be decided in double exponential time [6].

We remark that the case \( k < \text{ar}(T) - 1 \), excluded in Theorem 5.1 and Proposition 5.2, is somewhat esoteric as there are relation symbols whose arity is so high that they cannot be part of the UC in a UCQ\(_k\)-approximation (unless variables are reused). In the appendix, we provide a concrete example which illustrates that this case is significantly different from the case \( k \geq \text{ar}(T) - 1 \); in particular, we can see that Proposition 5.2 is provably wrong.

We now state our main result concerning guarded OMQs, which shows that UCQ\(_k\)-equivalence characterizes FPT for classes of OMQs from \((G, UCQ)\) over schemas of bounded arity.

Theorem 5.3 (Main Result I). Fix \( r \geq 1 \). Let \( \mathcal{O} \) be a recursively enumerable class of OMQs from \((G, UCQ)\) over a schema of arity \( r \). The following are equivalent, assuming FPT \( \neq W[1] \):

1. \( \text{p-OMQ-Evaluation}(\mathcal{O}) \) is in FPT.
2. There is \( k \geq 1 \) such that \( \mathcal{O} \subseteq \{G, UCQ\}^k \).

If either statement is false, then \( \text{p-OMQ-Evaluation}(\mathcal{O}) \) is \( W[1] \)-hard.

The easy direction is (2) implies (1), which exploits Theorem 5.1, and the third item of Proposition 3.3. The hard task is to show that (1) implies (2). To this end, since we assume that FPT \( \neq W[1] \), it suffices to show the following lower bound, which is our main technical result on OMQs based on guarded TGDs:

Theorem 5.4. Fix \( r \geq 1 \). Let \( \mathcal{O} \) be a recursively enumerable class of OMQs from \((G, UCQ)\) over a schema of arity \( r \), and, for each \( k \geq 1 \), \( \mathcal{O}_k \subseteq \{G, UCQ\}^k \). Then, \( \text{p-OMQ-Evaluation}(\mathcal{O}) \) is \( W[1] \)-hard.

In view of Proposition 3.2, which states that, for each \( k \geq 1 \), \( \text{OMQ-Evaluation}(\mathcal{O}) \) is \( \text{ExpTime-hard} \), it is not surprising that Theorem 5.3 does not state a pure tractability result. One may think that the above result can be easily obtained by using Grohe’s construction underlying the fpt-reduction from \( p \)-Clique that establishes the lower bound of Theorem 4.1. However, the fact that the ontology can introduce additional relations that are not part of the data schema causes serious challenges. A detailed sketch of our proof is presented in Section 6, while full details are in the appendix.

Constraint Query Specifications. We now turn our attention to CQs based on guarded TGDs. Interestingly, there is a strong connection between CQs and OMQs with full data schema, which allows us to transfer results from the OMQ to the CQ setting. Recall that a full data schema consists of all the predicates occurring in the ontology and the UCQ. Note that we can naturally convert a CQ \( S = (\Sigma, q) \) over a schema \( \mathcal{O} \) into the OMQ \( (S, \Sigma, q) \) that has full data schema, denoted \( \text{omq}(S) \). The following result relates the UCQ\(_k\)-equivalence of CQs to the UCQ\(_k\)-equivalence of OMQs.

Proposition 5.5. Consider a CQS \( S \in (G, UCQ) \) over \( T \). For each \( k \geq \text{ar}(T) - 1 \), the following are equivalent:

1. \( S \) is uniformly UCQ\(_k\)-equivalent.
2. \( \text{omq}(S) \) is UCQ\(_k\)-equivalent.

From Theorem 5.1 and Proposition 5.5, we get a 2ExpTime upper bound for deciding uniform UCQ\(_k\)-equivalence for guarded CQs, while the lower bound is inherited from [8]. As in Theorem 5.1, \( k \) should be greater than the arity of the schema, minus one.

Theorem 5.6. Let \( S \) be a CQS from \((G, UCQ)\) over a schema \( T \). For each \( k \geq \text{ar}(T) - 1 \), deciding whether \( S \) is uniformly UCQ\(_k\)-equivalent is 2ExpTime-complete.

Our main result concerning guarded CQSs shows that uniform UCQ\(_k\)-equivalence characterizes tractability and fpt for classes of CQs from \((G, UCQ)\) over schemas of bounded arity.

Theorem 5.7 (Main Result II). Fix \( r \geq 1 \). Let \( \mathcal{O} \) be a recursively enumerable class of CQs from \((G, UCQ)\) over a schema of arity \( r \). The following are equivalent, assuming FPT \( \neq W[1] \):

1. \( \text{CQS-Evaluation}(\mathcal{O}) \) is in PTIME.
2. \( \text{p-CQS-Evaluation}(\mathcal{O}) \) is in PTIME.
3. \( \text{There is } k \geq 1 \text{ such that } \mathcal{O} \subseteq \{G, UCQ\}^k \).

If either statement is false, then \( \text{p-CQS-Evaluation}(\mathcal{O}) \) is \( W[1] \)-hard.

The fact that (3) implies (1) is shown in [8], while (1) implies (2) holds by definition. The interesting task is to show that (2) implies (3). To this end, we exploit the subsequent result which relates the evaluation of CQs to the evaluation of OMQs. Given a class \( \mathcal{O} \) of CQs, we write \( \text{omq}(\mathcal{O}) \) for the class of OMQs \( \{\text{omq}(S) \mid S \in \mathcal{O}\} \).

Proposition 5.8. Consider a class \( \mathcal{O} \) of CQs from \((G, UCQ)\). There exists an fpt-reduction from \( \text{p-OMQ-Evaluation}(\text{omq}(\mathcal{O})) \) to \( \text{p-CQS-Evaluation}(\mathcal{O}) \).
The reduction exploits the fact that guarded TGDs are finitely controllable, which means that OMQ-Evaluation(G, UCQ) coincides with OMQ-Evaluation\textsubscript{nil}(G, UCQ) that considers only finite models [4]. Note that the reduction does not extend to frontier-guarded TGDs since it requires the evaluation of TGD bodies to be fpt. Actually, no such reduction is possible for frontier-guarded TGDs (unless FPT = \text{W}[1]) since there exists a class \(O\) of CQSs from \(\langle \Sigma,\Omega\rangle\) such that p-OMQ-Evaluation(omq(O)) is \text{W}[1]-hard, while p-CQS-Evaluation(\(\Omega\)) is in FPT: \(\Omega\) consists of CQSs \((\Sigma, \exists x R(x))\), where \(\Sigma\) contains TGDs of the form \(q(x, \bar{y}) \rightarrow R(x)\), i.e., frontier-guarded TGDs with only one frontier variable. Now, by Theorem 5.4, Proposition 5.5, and Proposition 5.8, we get the following result, which establishes the direction (2) implies (3) of Theorem 5.7 (since we assume that FPT \(\neq \text{W}[1]\)).

**Theorem 5.9.** Fix \(r \geq 1\). Let \(O\) be a recursively enumerable class of CQSs from \(\langle \Sigma,\Omega\rangle\) over a schema of arity \(r\), and, for each \(k \geq 1\), \(O \not\subseteq \langle G, UCQ \rangle^F_k\). Then, p-CQS-Evaluation(\(\Omega\)) is \text{W}[1]-hard.

**5.2 The Frontier-guarded Case**

By item (2) of Proposition 3.3, the notion of (uniform) UCQ\(_k\)-equivalence does not provide a characterization of fpt for classes of OMQs from \(\langle \Sigma,\Omega\rangle\). Such a characterization result for OMQs based on frontier-guarded TGDs via an alternative notion of “being equivalent to an OMQ of low treewidth” remains an interesting open problem. The situation is different for querying in the presence of constraints. In fact, uniform UCQ\(_k\)-equivalence allows us to characterize tractability and fpt for classes of CQSs from \(\langle \Sigma,\Omega\rangle\) over schemas of bounded arity \(r\) provided that the number of atoms in the head of TGDs is bounded by an integer \(m\). The latter is required in our proof for reasons that are explained in Section 7.

The first question is whether uniform UCQ\(_k\)-equivalence for CQSs based on frontier-guarded TGDs, with at most \(m \geq 1\) head atoms, over schemas of arity \(r \geq 1\), is decidable. We write \(\text{FG}_{m}\) for the class of frontier-guarded TGDs with at most \(m\) head atoms.

**Theorem 5.10.** Fix \(r, m \geq 1\). Let \(S\) be a CQS from \(\langle \text{FG}_{m},\text{UCQ} \rangle\) over a schema of arity \(r\). For each \(k \geq (r + m - 1)\), deciding whether \(S\) is uniformly UCQ\(_k\)-equivalent is \text{2ExpTime}-complete.

The above result exploits a characterization, analogous to Proposition 5.2, of when a CQS from \(\langle \text{FG}_{m},\text{UCQ} \rangle\) over a schema of arity \(r\) is uniformly UCQ\(_k\)-equivalent via a suitable notion of UCQ\(_k\)-approximation. The fact that we consider schemas of arity \(r\), and frontier-guarded TGDs with at most \(m\) atoms in the head, allows us to show that choosing a database of treewidth \(k\), with \(k \geq r + m - 1\), will lead to an instance that has treewidth \(k\). This in turn allows us to adopt a notion of UCQ\(_k\)-approximation that is significantly simpler than the one for guarded OMQs in Proposition 5.2. A contraction of a CQ \(q\) is a CQ obtained from \(q\) by identifying variables. When an answer variable \(x\) is identified with a non-answer variable \(y\), the resulting variable is \(x\), while the identification of two answer variables is not allowed. The UCQ\(_k\)-approximation of a CQS \(S = (\Sigma, q) \in \langle \Sigma, \Omega \rangle\) is the CQS \(S_k^q \in \langle \Sigma, q_k^q \rangle\), where \(q_k^q\) is the UCQ that consists of all contractions of a CQ from \(q\) that belong to UCQ\(_k\). We can then show the following:

**Proposition 5.11.** Fix \(r, m \geq 1, k \geq r + m - 1\). Let \(S\) be a CQS from \(\langle \text{FG}_{m},\text{UCQ} \rangle\) over a schema of arity \(r\). The following are equivalent:

1. \(S\) is uniformly UCQ\(_k\)-equivalent.
2. \(S \equiv S_k^q\).

From the above result, we get a procedure for deciding uniform UCQ\(_k\)-equivalence: compute the UCQ\(_k\)-approximation \(S_k^q = (\Sigma, q_k^q)\) of \(S = (\Sigma, q)\), and accept if \(S \subseteq S_k^q\) (the other direction holds by construction); otherwise, reject. Clearly, \(S_k^q\) can be computed in exponential time, while \(q_k^q\) consists, in general, of exponentially many CQs, each of polynomial size. By Proposition 4.5, \(S \subseteq S_k^q\) iff for each CQ \((\bar{x}, \bar{r}) \in q\), there exists \(p \in q_k^q\) such that \(\bar{x} = p(\text{chase}(\bar{r}, \Sigma))\). Moreover, by Proposition 3.1 and item (1) of Proposition 3.2, we get that checking whether \(\bar{x} \in p(\text{chase}(\bar{r}, \Sigma))\) is feasible in double exponential time. Summing up, to check whether \(S \subseteq S_k^q\) we need to perform exponentially many checks, while each check is feasible in double exponential time. This leads to the desired \text{2ExpTime} upper bound, while a matching lower bound is inherited from [8].

Our main result concerning frontier-guarded CQSs follows:

**Theorem 5.12 (Main Result III).** Fix \(r, m \geq 1\). Let \(O\) be a recursively enumerable class of CQSs from \(\langle \text{FG}_{m},\text{UCQ} \rangle\) over a schema of arity \(r\). The following are equivalent, assuming FPT \(\neq \text{W}[1]\):

1. OMQ-Evaluation(\(\Omega\)) is in \text{PTIME}.
2. p-CQS-Evaluation(\(\Omega\)) is in FPT.
3. There is \(k \geq 1\) such that \(O \not\subseteq \langle \text{FG}_{m},\text{UCQ} \rangle^F_k\).

If either statement is false, then p-CQS-Evaluation(\(\Omega\)) is \text{W}[1]-hard.

As for Theorem 5.7, the fact that (3) implies (1) is shown in [8], while (1) implies (2) holds by definition. Thus, the non-trivial task is to show that (2) implies (3). To this end, since we assume that FPT \(\neq \text{W}[1]\), it suffices to show the following, which is our main technical result on CQSs based on frontier-guarded TGDs:

**Theorem 5.13.** Fix \(r, m \geq 1\). Let \(O\) be a recursively enumerable class of CQSs from \(\langle \text{FG}_{m},\text{UCQ} \rangle\) over a schema of arity \(r\), and, for each \(k \geq 1, O \not\subseteq \langle \text{FG}_{m},\text{UCQ} \rangle^F_k\). Then, p-CQS-Evaluation(\(\Omega\)) is \text{W}[1]-hard.

Unlike Theorem 5.9, the above result cannot benefit from results on OMQs since (uniform) UCQ\(_k\)-equivalence does not characterize fpt for OMQs based on frontier-guarded TGDs. It is shown via a reduction from p-Clique by following the same approach as for the lower bound of Grohe’s Theorem, and exploiting the fact that frontier-guarded TGDs are finitely controllable. Let us stress, however, that our fpt-reduction is not a merely adaptation of Grohe’s reduction. Unlike Grohe, we are more constrained in defining the right database as it must satisfy the given set of constraints. Moreover, the useful notion of core of a CQ, which was crucial for Grohe’s proof, cannot be directly used in the presence of constraints.

**The Rest of the Paper.** We proceed to give some details on how our main lower bounds are shown; the full proofs are deferred to the appendix. Moreover, we defer to the appendix more details about the proofs of the results on the complexity of deciding UCQ\(_k\)-equivalence, and of the upper bounds stated in our main results. The lower bounds for guarded OMQs and CQS, i.e., Theorems 5.4 and 5.9, respectively, are discussed in Section 6, while the lower bound for frontier-guarded CQSs, i.e., Theorem 5.13, in Section 7.
6 GUARDEDNESS

6.1 Ontology-mediated Queries

We provide a proof sketch of Theorem 5.4, illustrating only the main ideas and abstracting away many technical details. We rely on a result by Grohe, stated here in a form tailored toward our proof. For $k, \ell \geq 1$, the $k \times \ell$-grid is the graph with vertex set \{(i,j) \mid 1 \leq i \leq k \text{ and } 1 \leq j \leq \ell\}, and an edge between (i,j) and (i',j') if \(|i-i'| + |j-j'| = 1\). A minor of an undirected graph is defined in the usual way, see, e.g., [26]. When $k$ is understood from the context, we use $K$ to denote \(\binom{\ell}{2}\). We say that a database $D$ is connected if the graph $GD$ is connected. A constant $a \in \text{dom}(D)$ is isolated in $D$ if there is only a single atom $R(\bar{a}) \in D$ with $\bar{a} \in \bar{a}$.

**Theorem 6.1 (Grohe).** Given an undirected graph $G$, a $k \geq 1$, a connected $S$-database $D$, and a set $A \subseteq \text{dom}(D)$ such that the restriction $GD[A]$ of the Gaisfman graph of $G$ to vertices $A$ contains the $k \times K$-grid as a minor, one can construct in time $(f(k))\cdot \text{poly}(|\Sigma|, |D|)$ an $S$-database $D_G$ with $\text{dom}(D_G) \subseteq A \subseteq \text{dom}(D_C)$ such that:

1. there is a surjective homomorphism $h_0$ from $D_G$ to $D$ that is the identity on $\text{dom}(D) \setminus A$,
2. $G$ contains a $K$-clique iff there is a homomorphism $h$ from $D$ to $D_C$ such that $h$ is the identity on $\text{dom}(D) \setminus A$ and $h(\bar{a})$ is the identity, and
3. if $R(a_1, \ldots, a_n) \in D$, $h_0(b_i) = a_i$ for each $i \in [n]$, and \(\{b_i \mid a_i \text{ non-isolated in } D\}\) is a clique in the Gaisfman graph of $D_C$, then $D_C$ contains an atom $R(c_1, \ldots, c_n)$ where $c_i = b_i$ if $a_i$ is non-isolated in $D$ and $h_0(c_i) = a_i$ for $i \in [n]$.

The condition in (3) is not considered in [26], but can easily be verified to hold. It can be viewed as an ontology requirement for the homomorphism $h_0$ from item (1) as it says that for certain atoms $R(\bar{a}) \in D$, we must find certain atoms $R(\bar{c}) \in D_C$ with $h(\bar{a}) = \bar{c}$. The set $A$ is not present in Grohe’s construction (that is, he considers the case $A = \text{dom}(D)$), but we show in the appendix that it can be taken into account by a minor change in the construction.

For the proof of Theorem 5.4, we are given a recursively enumerable class of OMQs $\mathcal{O} \subseteq (\mathcal{G}, \mathcal{UCQ})^n_k$ for all $k \geq 1$. As a preliminary, we show that for every $\mathcal{O} \subseteq \mathcal{Q}$, we can effectively find an equivalent OMQ $Q' \subseteq (\mathcal{G} \cup \mathcal{FULL})$ over the same extended language, where $\mathcal{FULL}$ is the class of full TGDs, i.e., $\mathcal{TGD}$s without existentially quantified variables. In other words, we can write away the existential quantifiers in $\mathcal{TGD}$ heads. Although $Q'$ need not be in $\mathcal{Q}$, we can switch between $\mathcal{Q}$ and $Q'$ in the proof whenever we only care about $Q$’s semantics. To simplify this proof sketch, assume that no rewriting is required as all ontologies in $\mathcal{O}$ are already from $\mathcal{G} \cap \mathcal{FULL}$, and assume that $\mathcal{O}$ contains only Boolean OMQs where the actual queries are CQs.

The proof is by fpt-reduction from p-Clique, a W[1]-hard problem. Assume that $G$ is an undirected graph and $k \geq 1$ a clique size. We aim to construct a database $D_G$ and find an OMQ $Q \in \mathcal{O}$ where

\[ (+) \ G \text{ has a $k$-clique iff } D_G \models Q. \]

We first describe how to find $Q$. By Robertson and Seymour’s Excluded Grid Theorem, there exists an $\ell$ such that every graph of treewidth exceeding $\ell$ contains the $k \times K$-grid as a minor [34]. We may assume, w.l.o.g., that $\ell$ exceeds the fixed arity $r$ of relation symbols that occur in $\mathcal{O}$. Since $\mathcal{O} \subseteq (\mathcal{G}, \mathcal{UCQ})^r_k$, there is an OMQ $Q = (\Sigma, q)$ from $\mathcal{O}$ that is not in $(\mathcal{G}, \mathcal{UCQ})^r_k$, and we can find $Q$ by enumerating $\mathcal{O}$ and relying on Theorem 5.1. It thus remains to construct $D_G$. In the special case where $\Sigma$ is empty and $\mathcal{Q}$ is full, this can be done by replacing $q$ with its core and applying Theorem 6.1, using $D[q]$ for $D$. Then, item (2) of that theorem corresponds to (+).

In the general case, however, we cannot use $D[q]$ for $D$ since $D[q]$ might use relation symbols that are not in $\mathcal{Q}$ and thus so would the resulting $D_G$. This, in fact, is what makes our proof rather delicate and difficult. The general idea is to replace $D[q]$ with an $S$-database $D_0$ such that $D_0 \models Q$ and $(D_0)^{\mathcal{Q}} \not\models Q$, where $(D_0)^{\mathcal{Q}}$ denotes the unraveling of $D_0$ into a structure of treewidth $\ell$. Ideally, we would like to apply the Grohe construction to $D_0$ instead of $D[q]$. However, of course we still have to attain (+), which is formulated in terms of $Q$ (and thus, in terms of $\Sigma$ and $q$) rather than $D_G$. This means that we need more from $D_0$ than just the stated properties and, in fact, we want $\text{chase}(D_0, \Sigma)$ and $q$ to be structurally as similar as possible. The following example illustrates a structural dissimilarity.

**Example 6.2.** Assume that $\Sigma$ is empty, that $\mathcal{Q}$ is full, and that $q$ is the $3 \times 4$-grid, that is,

\[
\bigwedge_{i \in \{1,2,3\}, j \in \{1,2,3\}} X(x_{i,j}, x_{i+1,j}) \land \bigwedge_{i \in \{1,2,3\}, j \in \{1,2\}} Y(x_{i,j}, x_{i,j+1}).
\]

Moreover, let $D_0$ be the following database, the $3 \times 3$-grid augmented with reflexive X-loops in the rightmost column:

\[
\{ X(a_{i,j}, a_{i+1,j}) \mid i \in \{1,2\}, j \in \{1,2,3\} \} \cup \\
\{ Y(a_{i,j}, a_{i,j+1}) \mid i \in \{1,2,3\}, j \in \{1,2\} \} \cup \\
\{ X(a_{3,j}, a_{3,j}) \mid j \in \{1,2,3\} \}.
\]

Clearly, $D_0 \models Q$ and $(D_0)^{\mathcal{Q}} \not\models Q$.

Example 6.2 is a particularly simple case: the ontology is empty, the data schema is full, and the query is a core. We could in fact choose $D_0 = D[q]$. In general, however, it might not be possible to make $D_0$ and $D[q]$ isomorphic, and there might not even be some $D_0$ such that $q$ maps injectively to $D_0$. What we do is choosing $D_0$ such that the following holds, where $D \models^{=\mathcal{P}} q$ means that $D \models q$ and all homomorphisms from $p$ to $D$ are injective (‘io’ stands for ‘injectively only’):

\[ (**\) \text{ if } \text{chase}(D_0, \Sigma) \models^{=\mathcal{Q}} q_c, \text{ where } q_c \text{ is a contraction of } q, \text{ then } \text{there is no } S\text{-database } D \text{ and contraction } q_c' \text{ of } q \text{ such that } D \text{ homomorphically maps to } D_0, \text{chase}(D, \Sigma) \models^{=\mathcal{Q}} q_c', \text{ and } q_c \neq q_c'. \]

In Example 6.2, $D_0$ does not satisfy this condition as witnessed by choosing $D = D[q]$, where $q_c$ is the contraction of $q$ that identifies $x_{3,j}$ with $x_{4,j}$ for $j \in \{1,2,3\}$, and $q_c' = q$. At this point, chase($D_0, \Sigma$) is still not as closely related to $q$ as we need it to be. In fact, $D_0$ might contain parts that are superfluous for $D_0 \models Q$, and parts in which the atoms are unnecessarily entangled with each other. The following example illustrates the latter.

**Example 6.3.** Assume that $q$ takes the form of an $n \times m$-grid that uses the binary relations $X$ and $Y$. Let $S = \{X', Y'\}$ and:

\[
\Sigma = \{ X'(x, y, z) \rightarrow X(x,y), Y'(x, y, z) \rightarrow Y(x,y) \}.
\]

Further, let $D_0$ be the database

\[
\{ X'(a_{i,j}, a_{i+1,j}, b) \mid 1 \leq i \leq m \text{ and } 1 \leq j \leq n \} \cup \\
\{ Y'(a_{i,j}, a_{i+1,j}, b) \mid 1 \leq i < m \text{ and } 1 \leq j \leq n \}.
\]


Clearly, $D_0 \models Q$. However, the following ‘untangled’ homomorphic preimage $D_1$ of $D_0$ also satisfies $D_1 \models Q$, and for our proof it is much preferable to work with $D_1$ rather than with $D_0$:

$$\{X'(a_{i,j}, a_{i+1,j}, b_{ij}) \mid 1 \leq i \leq m \text{ and } 1 \leq j < n \} \cup \{Y'(a_{ij}, a_{i+1,j}, b'_{ij}) \mid 1 \leq i < m \text{ and } 1 \leq j \leq n\}.$$ 

Again, the example illustrates only a very simple case. To relate $D_0$ and $q$ even closer, we apply to $D_0$ a ‘diversification’ construction that replaces each atom $R(\bar{a}) \in D_0$ with a (potentially empty) set of atoms $R(\bar{a}_1), \ldots, R(\bar{a}_n)$, where each $R(\bar{a}_i)$ can be obtained from $R(\bar{a})$ by replacing some constants with fresh isolated constants, and then attaching to the new atoms a finite initial piece of the guarded unraveling of $D_0$ that starts at $R(\bar{a})$. The latter part is necessary to not lose entailments of atoms through the ontology $\Sigma$; such entailments are preserved by guarded unraveling, which like the introduction of fresh constants is part of the ‘untangling’. This diversification operation is made precise in the appendix. We use as $D_G$ the database obtained by applying the Grohe construction to the result $D_1$ of diversifying $D_0$ in a maximal way such that $D_1 \models Q$. It then remains to show that ($\ast$) holds.

The ‘only if’ direction is rather straightforward. For the ‘if’ direction, assume that $D_G \models Q$, that is, $\text{chase}(D_G, \Sigma) \models q$. Thus, there is a contraction $q'_G$ of $q$ with $\text{chase}(D_G, \Sigma) \models q'_G$. The composition $g$ of a witnessing homomorphism with $h_0$ shows $\text{chase}(D, \Sigma) \models q$, and thus there is a contraction $q_c$ of $q$ with $\text{chase}(D, \Sigma) \models q_c$. Property ($\ast\ast$) is preserved by diversification, and thus by putting in $D_1$ for $D_0$ and $D_G$ for $D$, we obtain $q_c = q'_c$. It follows that $g$ must be injective and thus, so is $h_0$. We can also show that $h_0$ is ‘sufficiently surjective and onto’ so that from the inverse of $h_0$, we can construct a homomorphism $h$ from $D_1$ to $D_G$ that satisfies the conditions from Theorem 6.1, and then apply that theorem to conclude that $G$ contains a $k$-clique, as needed.

We remark that our proof is considerably more involved than the one in [7], where ontologies are formulated in the description logic $ELHI$, essentially a fragment of $\mathcal{O}$. In particular, the presence of relations of arity larger than two makes it necessary to use diversifications, isolated constants, and condition (iii) from Theorem 6.1, while none of this is necessary for the $ELHI$ case.

### 6.2 Constraint Query Specifications

We now discuss the fpt-reduction of Proposition 5.8, which is underlying Theorem 5.9. Recall that this reduction relies on the fact that guarded TGDs are finitely controllable. Recall also that the notion of finite controllability is crucial for establishing our main technical result on CQSs based on frontier-guarded TGDs (Theorem 5.13), for which more details are given in the next section. Let us then give some useful details around finite controllability.

**Finite Controllability.** Given an $S$-database $D$, and a set $\Sigma$ of TGDs over $S$, we write $\text{fmods}(D, \Sigma)$ for the set of all finite models of $D$ and $\Sigma$. Finite controllability guarantees that, for computing the set of tuples that are answers to a UCQ $q$ over every finite model of $D$ and $\Sigma$, i.e., the set $\bigcap_{M \in \text{fmods}(D, \Sigma)} q(M)$, it suffices to evaluate $q$ over the (possibly infinite) instance chase($D, \Sigma$).

**Definition 6.4 (Finite Controllability).** A class $\mathcal{C}$ of TGDs is finitely controllable if, for every $S$-database $D$, set $\Sigma \in \mathcal{C}$ of TGDs over $S$, and UCQ $q$ over $S$, $q(\text{chase}(D, \Sigma)) = \bigcap_{M \in \text{fmods}(D, \Sigma)} q(M)$. ▪

Actually, as we shall see, in our proofs we rely on a seemingly stronger property than finite controllability. Roughly speaking, this property states that, once we fix the number of variables that can appear in UCQs, let say $n \geq 0$, given a database $D$, and a set $\Sigma$ of TGDs, there exists a finite model $M$ of $D$ and $\Sigma$, which depends on $n$, that allows us to compute the set of tuples $q(\text{chase}(D, \Sigma))$, i.e., $q(\text{chase}(D, \Sigma)) = q(M)$, no matter what the UCQ looks like.

**Definition 6.5 (Strong Finite Controllability).** Consider a class $\mathcal{C}$ of TGDs. We say that $\mathcal{C}$ is strongly finitely controllable if, for every $S$-database $D$, set $\Sigma \in \mathcal{C}$ of TGDs over $S$, and integer $n \geq 0$, there exists an instance $M(D, \Sigma, n) \in \text{fmods}(D, \Sigma)$ such that, for each UCQ $q$ that has at most $n$ variables, $q(\text{chase}(D, \Sigma)) = q(M(D, \Sigma, n))$. If there is a computable function that takes as input $D, \Sigma$, and $n$, and outputs $M(D, \Sigma, n)$, then we say that a finite witness is realizable.

The next technical result, which is of independent interest, shows that finite and strong finite controllability are equivalent properties.

**Lemma 6.6.** Consider a class $\mathcal{C}$ of TGDs. It holds that $\mathcal{C}$ is finitely controllable iff $\mathcal{C}$ is strongly finitely controllable.

By using the above lemma and results on the guarded negation fragment of first-order logic (GNFO) from [5], we can show that:

**Theorem 6.7.** The class $\mathcal{FO}$ is strongly finitely controllable. Moreover, finite witnesses are realizable.

In particular, every satisfiable GNFO sentence has a finite model. This allows us to show that the class of frontier-guarded TGDs is finitely controllable, and thus, by Lemma 6.6, it is also strongly finitely controllable. Moreover, we know that if a GNFO sentence $\varphi$ has a finite model, then it has a finite model of size $2^{2^{\{|\varphi|\,|O(\varphi)|\}}}$.

This allows us to show, in addition, that finite witnesses are realizable.

**The FPT-Reduction.** Let us now come back to Proposition 5.8. We want to show the following: given a class of CQSs $\mathcal{O}$ from $\mathcal{C}$, UCQ), there exists an fpt-reduction from p-OMQ-Evaluation($\text{omq}(\mathcal{O})$) to p-CQS-Evaluation($\mathcal{O}$). Consider $Q = (\Sigma, q(x))$ from $\text{omq}(\mathcal{O})$, an $S$-database $D$, and a tuple $\bar{c} \in \text{dom}(D)^{|x|}$. We are going to construct an $S$-database $D^+$ such that $D^+ \models \Sigma$, and $\bar{c} \in \text{chase}(D, \Sigma)$ iff $\bar{c} \in q(D^+)$. or, equivalently (due to Proposition 2.2), $\bar{c} \in \text{chase}(D, \Sigma)$ iff $\bar{c} \in q(D^+)$. We first define $D^+$ as the database

$$D \cup \{R(\bar{a}) \in \text{chase}(D, \Sigma) \mid \bar{a} \subseteq \text{dom}(D)\}.$$ 

Let $A$ be the family of all maximal tuples $\bar{a}$ over $\text{dom}(D)$ that are guarded in $D^+$, i.e., there is an atom $R(\bar{b}) \in D^+$ such that $\bar{a} \subseteq \bar{b}$. Fix an arbitrary tuple $\bar{a} \in A$. Since, by Theorem 6.7, the class $\mathcal{C}$ is strongly finitely controllable, and also finite witnesses are realizable, we can compute an instance $M(D^+_{\bar{a}}, \Sigma, n) \in \text{fmods}(D^+_{\bar{a}}, \Sigma)$, where $n$ is the number of variables in $q$, such that for each $\mathcal{C}$-query $q'$ of arity $|\bar{a}|$ with at most $n$ variables, it holds that

$$(\ast) \quad \bar{a} \in q'(M(D^+_{\bar{a}}, \Sigma, n)) \iff \bar{a} \in q'(\text{chase}(D^+_{\bar{a}}, \Sigma)).$$

W.l.o.g., we assume that $\text{dom}(M(D^+_{\bar{a}}, \Sigma, n)) \cap \text{dom}(M(D^+_{\bar{b}}, \Sigma, n)) \subseteq \text{dom}(D)$, for every two distinct tuples $\bar{a}, \bar{b} \in A$. The database $D^+$ is

$$D^+ \cup \bigcup_{\bar{a} \in A} M(D^+_{\bar{a}}, \Sigma, n).$$

The next lemma shows that $D^+$ is the desired database.
Our proof borrows several ideas and techniques from the proof of Grohe’s Theorem. However, the adaptation of such techniques is non-trivial for a couple of reasons. First, we cannot construct an arbitrary database, but need to construct one that satisfies the given set $\Sigma$ of frontier-guarded TGDs. Second, the notion of core of a CQ, which is crucial for Grohe’s proof, cannot be directly used in our context. Recall that, for $k \geq 1$, a CQ $q$ belongs to $\text{CQ}_k$ iff its core is in $\text{CQ}_k$ [20]. This allows Grohe to work with the core of $q$ instead of $q$ itself, and thus, exploit the property that a homomorphism from the core of $q$ to itself is injective. This is far from being true in the presence of constraints. Instead, we need a technical result, which, intuitively speaking, states that some subsets of our CQs behave like cores for the sake of our proof.

**Lemma 7.2.** Fix $\ell \geq r \cdot m$. There exists a computable function that takes as input a CQS $S = (\Sigma, q) \in (\text{PG}_m, \text{UCQ})$ that is not uniformly $\text{CQ}_\ell$-equivalent, and outputs two CQs $p$ and $p'$, and a subset $X$ of the variables of $p$, such that the following hold:

1. $q \equiv_{\Sigma} p$.
2. $D[p'] \models \Sigma$.
4. $h(X) = X$, for every homomorphism $h$ from $p$ to $p'$.
5. The treewidth of $G_{p, X}^{\Sigma}$ is larger than $\ell$.

The proof of the above lemma can be found in the appendix. Let us now briefly comment on the reason why we have to bound the number of head atoms in frontier-guarded TGDs for our proof to work. In very rough terms, this is because the way in which $p'$ is constructed is by generating from $p$ the finite model $M = M[D[p], \Sigma, n]$, for some $n \geq 0$, that is obtained from Definition 6.5 given the strong finite controllability of $\text{PG}$, as established in Theorem 6.7. A property of this construction that is crucial for the proof of Lemma 7.2 is that the treewidth of $M$ is not larger than that of $D[p]$. The only way in which we can ensure this property to hold is by fixing the number of atoms in TGD heads.

Remarkably, this additional bound on the number of head atoms is not needed for obtaining Theorem 5.7. This is because there exists an fpt-reduction from the evaluation of CQs from $(\text{FG}, \text{UCQ})$ to CQs from the same class in which there are no existential quantifiers in TGD heads – this is implicit in our proofs. When such a reduction is possible, the CQ $p'$ can be obtained in a simpler way: it suffices to chase $p$ using $\Sigma$, with $p$ being the CQ with a minimal number of atoms among all the CQs that are equivalent to $q$ w.r.t. $\Sigma$. Notice that no bound on the number of head atoms in the TGDs of $\Sigma$ is required in this case. Unfortunately, when $\Sigma \in \text{PG}$, a reduction like the one described above is not possible. This prevents us from applying the previous idea since chase($p$, $\Sigma$) may be infinite.

The FPT-Reduction. We now proceed to explain how Lemma 7.2 is applied in order to prove Theorem 5.13. Let $(\Gamma, k)$ be an instance of p-Clique. It is easy to see that we can assume, w.l.o.g., that every clique of size at most $3 \cdot r \cdot m$ in $G$ is contained in a clique of size $3 \cdot r \cdot m$. From the Excluded Grid Theorem [34], there is a computable integer $\ell$ such that every simple graph $G$ of treewidth at least $\ell$ contains a $(k \times K)$-grid as a minor. By hypothesis on the class $\Omega$, there exists a CQS $S = (\Sigma, q)$ from $\Omega$ such that $S \not\in (\text{PG}_m, \text{UCQ})$. We can assume, w.l.o.g., that $\ell \geq r \cdot m$. We build from $(\Gamma, k)$ the tuple $(D', \Sigma, q)$, where $D'$ is a database defined as follows. Since, by assumption, $\ell \geq r \cdot m$, it is possible to compute from $q$ a CQ $p$, a subset $X$ of the variables of $p$, and a CQ $p'$, that satisfy the properties stated in Lemma 7.2. In particular, the treewidth of $D'_{\Sigma, q}$ is at least $\ell$, and hence, $G_{p, X}^{\Sigma}$ contains the $(k \times K)$-grid as a minor. We then define $D'$
as $D'(G, D[p], D[p'], X)$, where $D(G, D[p], D[p'], X)$ is the database that is defined in Theorem 7.1 for $D = D[p], D' = D[p'],$ and $A = X$.

It remains to show that the above is indeed an fpt-reduction from p-Clique to CQS-Evaluation($\emptyset$). To this end, we need to show that:

**Lemma 7.3.** The following statements hold:

1. $D^* \models \Sigma$.
2. $G$ has a $k$-clique iff $D^* \models q$.
3. There are computable functions $f, g : \mathbb{N} \to \mathbb{N}$ such that $(D, \Sigma, q)$ can be constructed in time $||G||^{O(1)} \cdot f(k)$ and $(||q|| + ||\Sigma||) \leq g(k)$.

The proof of (1) follows from the last item in Theorem 7.1. We now proceed to show item (2). Assume that $G$ has a $k$-clique. Then, by Theorem 7.1, there is a homomorphism $h$ from $p$ to $D^*$ such that $h_0(h(\cdot))$ is the identity. Hence $D^* \models p$. But $q \equiv_\Sigma p$, and thus, $D^* \models q$ since, by item (1), we know that $D^* \models \Sigma$. Conversely, assume that $D^* \models q$, and thus $D^* \models p$. Then, there is a homomorphism $h$ from $p$ to $D^*$. It follows that $h_0(h(\cdot))$ is a homomorphism from $p$ to $p'$. Consequently, by Lemma 7.2, we obtain that $h(X) = X$. Hence, there must exist some $n \geq 0$ such that $q = h_0(h_0(h_0(\ldots h_0(h_0(\cdot))))$ is a homomorphism from $p$ to $D^*$ with $h_0(h(g(\cdot)))$ being the identity on $X$. Then, $G$ has a $k$-clique from item (2) of Theorem 7.1.

As for items (3) and (4), first notice that the CQS $S = (\Sigma, q)$ can be computed by simply enumerating the CQSs from $p$ until we find $S$ since, by Theorem 5.10, we can check whether $D\models q$ computed by simply enumerating the CQSs from $p$. Since, by item (1), we know that $D\models \Sigma$. Theorem 7.1 states, on the other hand, that it is possible to construct $D$ in time $||G||^{O(1)} \cdot f'(k)$ for some computable function $f' : \mathbb{N} \to \mathbb{N}$. Putting all these together, we obtain the existence of computable functions $f, g : \mathbb{N} \to \mathbb{N}$ as needed.

## 8 RELATIONS OF UNBOUNDED ARITY

We take a first look at the case where the arity of relations is not bounded by a constant, focusing on linear TGDs and the OMQ language ($\mathbb{L}, \mathbb{UCQ}$). In contrast to the OMQ languages studied in previous sections, every OMQ from ($\mathbb{L}, \mathbb{UCQ}$) is rewrivable into an equivalent UCQ. We are going to exploit this property in our analysis. For simplicity, we only treat Boolean OMQs in what follows, so without further mention all queries considered in this section are assumed to be Boolean.

We start with some preliminaries. A hypergraph is a pair $H = (V, E)$ with $V$ a set of nodes and $E \subseteq 2^V \setminus \emptyset$ a set of edges. A tree decomposition of a hypergraph is defined like that of a graph in Section 2 except that the second condition now says: if $e \in E$, then $e \subseteq \chi(t)$ for some $t \in V_G$. The treewidth of a hypergraph is then defined in the expected way. Every CQ $q$ gives rise to an associated hypergraph $H[q]$ whose vertices are the variables of $q$ and which contains an edge $x$ for every atom $R(x)$ of $q$.

Let $H = (V, E)$ be a hypergraph. A function $f : 2^V \to \mathbb{R}^+$ is submodular if $f(X) + f(Y) \geq f(X \cap Y) + f(X \cup Y)$. We only consider submodular functions $f$ that satisfy $f(\emptyset) = 0$ and are edge dominated, that is, $f(e) \leq 1$ for all $e \in E$. The submodular width of $H$ is the smallest $k$ such that for every monotone submodular function $f$, there is a tree decomposition $(T, \chi)$ of $H$, $T = (V_T, E_T)$ with $f(\chi(t)) \leq k$ for all $t \in V_T$. The submodular width of a CQ $q$ is that of $H[q]$. A class of CQs $\mathcal{Q}$ has bounded submodular width if there exists a $k > 0$ such that the submodular width of each $q \in \mathcal{Q}$ is at most $k$.

A CQ $q$ is self-join free if no relation symbol occurs in more than one atom in $q$. The exponential time hypothesis (ETH) says that there is no algorithm for $n$-variable 3SAT that runs in time $2^{o(n)}$ [27]. It implies FPT $\neq$ W[1]. The following result can to some extent be viewed as a comparison of Theorem 6.1 for the case of unbounded relation arity.

**Theorem 8.1 (Marx).** Let $\mathcal{Q}$ be a recursively enumerable class of CQs such that for every $q \in \mathcal{Q}$, there is a self-join free $q' \in \mathcal{Q}$ with $H[q] = H[q']$. The following are equivalent unless ETH fails:

1. $\mathcal{Q}$ can be evaluated in FPT;
2. $\mathcal{Q}$ has bounded submodular width.

The "$(2) \Rightarrow (1)$" direction also holds without the condition on $\mathcal{Q}$ and does not depend on ETH.

This formulation is actually a slight strengthening of that of Marx who only demands that $\mathcal{Q}$ is closed under CQs that have the same hypergraph. Marx’ proof, however, also establishes this slightly stronger version. Note that, in contrast to Grohe’s Theorem, Theorem 8.1 does not speak about equivalence to CQs of bounded width. This is related to the assumption about self-join freeness: the hypergraph of every self-join free query is a core and in fact, there is no gain in transitioning from a core to an equivalent query. It is not known whether Marx’ result holds also without this assumption.

As a preparation for what follows, we first lift Marx’ result to the UCQ case. We call a UCQ canonical if the CQs in it are mutually homomorphically incomparable. It is easy to see that every UCQ is equivalent to a canonical UCQ. A class of UCQs $\mathcal{Q}$ has bounded submodular width if there exists a $k$ such that every CQ that occurs in some $q \in \mathcal{Q}$ has submodular width at most $k$.

**Theorem 8.2.** Let $\mathcal{Q}$ be a recursively enumerable class of canonical UCQs such that for every $q \in \mathcal{Q}$ and every $p \in \mathcal{Q}$ in $q$, there is a UCQ $q' \in \mathcal{Q}$ and a self-join free CQ $p'$ in $q'$ with $H[p] = H[p']$. The following are equivalent unless ETH fails:

1. $\mathcal{Q}$ can be evaluated in FPT;
2. $\mathcal{Q}$ has bounded submodular width.

The "$(2) \Rightarrow (1)$" direction also holds without the condition on $\mathcal{Q}$ and does not depend on ETH.

Proof. The "$(2) \Rightarrow (1)$" direction is trivial: one can simply evaluate each CQ in a given UCQ and consider the union of results.

As concerns the "$(1) \Rightarrow (2)$" direction, assume that $\mathcal{Q}$ has unbounded submodular width. Then for every $k$, there is a UCQ $q_k$ in $\mathcal{Q}$ and a CQ $p_k$ in $q_k$ such that $p_k$ has submodular width higher than $k$. Furthermore, there must be some UCQ $q_k$ in $\mathcal{Q}$ and some CQ $p_k$ in $q_k$ such that $p_k$ is self-join free and has submodular width higher than $k$. Let $Q$ be the class of all CQs $p_k, k > 0$. Then, each CQ in $Q$ is self-join free, and thus $Q$ meets the prerequisites from Theorem 8.1.

As $Q$ does not have bounded submodular width, it follows from Theorem 8.1 that $Q$ cannot be evaluated in FPT. We show that the same holds for $\mathcal{Q}$ by reducing the evaluation of each CQ from $Q$'s...
to the evaluation of some UCQ from \(Q\). Let \(\tilde{p}_k\) be a CQ from \(Q'\), for some \(k > 0\), and \(D\) be some database. By construction, there is a UCQ \(q_k\) in \(Q\) which contains \(\tilde{p}_k\) as a CQ and in which each two CQs are homomorphically incomparable. Then, for every CQ \(p'\) in \(q_k\) with \(p' \neq \tilde{p}_k\), \(\text{D}[\tilde{p}_k] \models p'\) (due to the fact that \(\tilde{p}_k\) and \(p'\) are homomorphically incomparable). Moreover, \(\text{D}[\tilde{p}_k] \models \tilde{p}_k\). It follows that \(D \models \tilde{p}_k\) iff \(D \times \text{D}[\tilde{p}_k] \models q_k\).

The assumption about self-join freeness also needs to be reflected in the intended classification of OMQs. It makes, however, little sense to adopt it verbatim: to obtain a self-join free CQ with the same hypergraph, one has to replace relation symbols in the query with fresh ones, but this breaks the connection to the ontology. We thus develop a more natural and mild condition. A class of OMQs \(O \subseteq (\Sigma, \text{UCQ})\) is closed under fresh implicants if whenever \((\Sigma, S, q) \in O, R \in S,\) and \(R'\) is a relation symbol of the same arity as \(R\) that does not occur in \(\Sigma\) or \(q\), then \(q' = (\Sigma \cup \{R'(x) \rightarrow R(x)\}, S \cup \{R'\}, q) \in O\) where \(x\) is a tuple of variables of suitable length without repetitions. We then also call \(R'\) a fresh implicant of \(R\) and say that \(q'\) is obtained from \(q\) by adding a fresh implicant of \(R\).

The submodular width of an OMQ \((\Sigma, S, q)\) is that of \(q\). For \(k > 0\), we use \((\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) to denote the class of OMQs from \((\text{linTGD}, \text{UCQ})\) that are of submodular width at most \(k\) and \((\Sigma, \text{UCQ})_{\text{SMW}(k)}\) to denote the class of OMQs \(Q \in (\Sigma, \text{UCQ})\) that are equivalent to an OMQ from \((\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\).

**Theorem 8.3.** For any recursively enumerable class \(Q\) of OMQs that is closed under fresh implicants, the following are equivalent unless ETH fails:

1. \(Q\) can be evaluated in FPT;
2. \(Q \subseteq (\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) for some \(k > 0\).

The \((2) \Rightarrow (1)\) direction also holds without fresh implicants and does not depend on ETH.

To prove Theorem 8.3, we exploit the fact that every OMQ \(Q \in (\Sigma, \text{UCQ})\) can be rewritten into a UCQ \(q\) that is equivalent to \(Q\) [15]. Call a UCQ core if all CQs in it are cores. For a class of OMQs \(Q \subseteq (\Sigma, \text{UCQ})\), we use \(\text{rew}(Q)\) to denote the set of all canonical core UCQs that are equivalent to some OMQ from \(Q\). The following is a central observation.

**Lemma 8.4.** For every recursively enumerable \(Q \subseteq (\Sigma, \text{UCQ})\) that is closed under fresh implicants, UCQ \(q \in \text{rew}(Q)\), and CQ \(p\) in \(q\), there is a UCQ \(q' \in \text{rew}(Q)\) and a self-join free CQ \(p'\) in \(q'\) such that \(H[p] = H[p']\).

Proof. Let \(Q, q,\) and \(p\) be as in the lemma. Also, let \(Q = (\Sigma, S, q) \in Q\) be an OMQ that is equivalent to \(q\). As \(Q\) is closed under fresh implicants, there must be an OMQ \(Q' \in Q\) that is obtained from \(Q\) by adding fresh implicants and such that for every \(R \in S\), the number of fresh implicants of \(R\) in \(Q'\) is bounded from below by the maximum size of CQs in \(q\). Let \(q'\) be the UCQ obtained from \(q\) by adding as disjuncts all CQs obtained from a CQ in \(q\) by replacing atoms \(R(x)\) with \(R'(x), R'\) a fresh implicant of \(R\) in \(Q'\). We observe the following:

1. \(q'\) is equivalent to \(Q'\);
2. \(q'\) is canonical.

Sketch the arguments.

It follows from Points 1 and 2 above that \(q' \in \text{rew}(Q)\). Furthermore, by construction \(q'\) contains a CQ \(p'\) that has been obtained from \(p\) by replacing each atom with one based on a different fresh implicant, and thus \(p'\) is self-join free and \(H[p] = H[p']\), as required.

We are now ready to prove Theorem 8.3.

\((1) \Rightarrow (2)\). Assume that \(Q\) can be evaluated in FPT. Then so can \(\text{rew}(Q)\) because for every \(q \in \text{rew}(Q)\) we can effectively find a \(Q \in Q\) that is equivalent: enumerate all OMQs from \(Q\), construct a UCQ-rewriting \(q'\) of \(Q\), and check whether \(q \equiv q'\). By Lemma 8.4 and Theorem 8.2, the submodular width of the UCQs in \(\text{rew}(Q)\) is bounded by some \(k\). We can view each query from \(\text{rew}(Q)\) as an OMQ from \((\text{linTGD}, \text{UCQ})\) in which the ontology happens to be empty. Consequently, \(Q \subseteq (\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\).

\((2) \Rightarrow (1)\). Assume that \(Q \subseteq (\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) for some \(k > 0\). It suffices to argue that \(\text{rew}(Q)\) can be evaluated in FPT as we can effectively convert every OMQ from \(Q\) into an equivalent one from \(\text{rew}(Q)\).

By Lemma A.1 and Theorem 8.2, the class \((\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) can be evaluated in FPT, and thus so can \(\text{rew}(\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\). It is therefore enough to show that \(\text{rew}(Q) \subseteq \text{rew}(\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\).

Let \(q \in \text{rew}(Q)\). Then there is a \(Q \in Q\) that is equivalent to \(q\) and \(Q_k \in (\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) that is equivalent to \(Q\) and thus to \(q\). It follows that \(\text{rew}(\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\) contains an OMQ \(q_k\) that is equivalent to \(q\). Both \(q\) and \(q_k\) consists of CQs that are homomorphically incomparable cores. It follows that for every CQ in \(q\), there is an isomorphic one in \(q_k\) and vice versa. Thus, \(q \in \text{rew}(\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\).

It seems worthwhile to point out that it follows from the above proof that if \(Q \subseteq (\text{linTGD}, \text{UCQ})_{\text{SMW}(k)}\), then the submodular width of the CQs in \((\text{UCQs in}) \text{rew}(Q)\) is also bounded by \(k\). This of course depends on the fact that we assume all such CQs to be cores.

9 CONCLUSIONS AND FUTURE WORK

We have studied the limits of efficient query evaluation in the context of ontology-mediated querying and querying in the presence of constraints, focusing on guarded and frontier-guarded TGDs, and on UCQs as the actual queries. We have obtained novel efficiency characterizations in the spirit of Grohe’s well-known characterization of the tractable classes of CQs (without constraints).

An interesting open problem that emerges from this work is whether our main result on guarded OMQs (Theorem 5.3) can be generalized to frontier-guarded TGDs. Since the notion of being equivalent to an OMQ of bounded treewidth is not enough for frontier-guarded TGDs. We would also like to investigate whether in our main result on CQs based on frontier-guarded TGDs (Theorem 5.12) the bound on the number of head atoms can be dropped. Recall that our results, similarly to Grohe’s characterization, assume that the underlying schema is of bounded arity. Establishing similar efficiency characterizations for schemas of unbounded arity...
is a challenging open problem, where the more involved notion of submodular width, introduced by Marx [31], should be considered.

REFERENCES


A PROOF OF PROPOSITION 3.3

The first two items are easy, and have been already discussed in the main body of the paper. Here we focus on item (3). The proof relies on two technical lemmas.

The first such lemma, which is implicit in [15], states that, given an S-database $D$ and an OMQ $Q = (S, \Sigma, q)$ from (L, UCQ), $Q(D)$ coincides with the evaluation of $q$ over an initial finite portion $C$ of chase($D, \Sigma$). To formalize this, we need then notion of chase level. Let $s = s_0 \rightarrow s_1 \rightarrow s_2 \rightarrow \ldots$ be a chase sequence for a database $D$ under a set $\Sigma$ of TGDs, where $s_0 = \phi_I(x, \bar{y}) \rightarrow \exists \bar{y} \phi_I(x, \bar{z})$. The $s$-level of an atom $\alpha \in$ chase($D, \Sigma$) is inductively defined as follows: levels($\alpha$) = 0 if $\alpha \in D$, and levels($\alpha$) = $l$ if max(levels($\beta$) | $\beta \in \phi_I(t_{i-1}, \bar{u}_{i-1})$) = $l - 1$, where $i > 0$ is such that $\alpha \in I_i$ and $\alpha \notin \bigcup_{0 \leq j < i} I_i$. Let chase$_i$($D, \Sigma$) be the instance $\{\alpha \in \bigcup_{1 \leq j \leq i} I_j |$ levels($\alpha$) $\leq l\}$. A chase sequence for $D$ under $\Sigma$ is level-wise if, for $i > 0$, it produces all the atoms of $s$-level $i$ before generating an atom of $s$-level $i + 1$.

**Lemma A.1.** Consider a database $D$, a set $\Sigma \subseteq \mathbb{L}$, and a UCQ $q$. There exists a computable function $g : \mathbb{N} \rightarrow \mathbb{N}$ such that

$q$(chase($D, \Sigma$)) = $q$ (chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma))

where $s$ is a level-wise chase sequence for $D$ under $\Sigma$. Furthermore, the instance chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma)$ can be computed in time $|D| \cdot f(||\Sigma||)$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

**Proof.** The existence of the computable function $g$, which is actually an exponential function, has been shown in [15]. It remains to show that chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma)$ can be computed in time $|D| \cdot f(||\Sigma||)$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$; recall that $s$ is a level-wise chase sequence for $D$ under $\Sigma$. We first provide an upper bound on the size of chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma)$. Let $H_D$ be the maximum number of atoms in the head of a TGD of $\Sigma$.

**Lemma A.2.** It holds that

[chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma) \leq |D| \cdot (|\Sigma| \cdot H_D + 1)^{g(||\Sigma|| + ||q||)}]

**Proof.** We proceed by induction on the $s$-level $i \geq 0$.

**Base case.** It is clear that chase$_0^0(D, \Sigma) = D$, and the claim follows.

**Inductive step.** Observe that, due to linearity, each TGD of $\Sigma$ can be triggered by an atom of chase$_{i-1}^1(D, \Sigma)$ at most once, and generate $H_D$ new atoms. Hence,

[chase$_i^1(D, \Sigma) \leq$ chase$_{i-1}^1(D, \Sigma) \cdot |\Sigma| \cdot $chase$_{i-1}^1(D, \Sigma) \cdot H_D$.

By induction hypothesis,

[chase$_{i-1}^1(D, \Sigma) \leq |D| \cdot (|\Sigma| \cdot H_D + 1)^{i-1}$.]

Therefore,

[chase$_i^1(D, \Sigma) \leq |D| \cdot (|\Sigma| \cdot H_D + 1)^i \cdot (|\Sigma| \cdot H_D + 1)^{i-1} \cdot H_D$]

[chase$_i^1(D, \Sigma) \leq |D| \cdot (|\Sigma| \cdot H_D + 1)^i$,

and the claim follows.]

A.1 Proof of Lemma A.3

The rest of this section is devoted to the proof of Lemma A.3. Consider an S-database $D$, and an OMQ $Q = (S, \Sigma, q) \in (G, UCQ)$. We proceed to construct a database $D^*$, and a set $\Sigma^* \subseteq G_\Sigma$ by adapting a construction from [23], such that $Q(D) = q$(chase$(D^*, \Sigma^*))$. Let us first explain the high-level idea underlying $D^*$ and $\Sigma^*$.

The High-level Idea. A key notion when reasoning with guarded TGDs is the type of an atom $\alpha$ (w.r.t. $D \Sigma$), denoted type$_D(\Sigma)(\alpha)$, defined as the set $\{\beta \in$ chase($D, \Sigma$) | dom($\beta$) $\subseteq$ dom($\alpha$))}. Roughly speaking, the key property of the type is that the atoms that can be derived during the chase from an atom $\alpha$ used as a guard is determined by its type; see [15] for further details. The main idea underlying the construction of $D^*$ is to encode an atom $\alpha$ in $D$ and its type as a single atom of the form $|r|\{\tau\}$, where $r$ is essentially a representation of the “shape” of $\alpha$ and its type. For example, given an atom $R(a, b) \in D$, and assuming that its type is $\{\tau|a, b, Si(b, a), T(b), T(a)\}$, we can encode $R(a, b)$ and its type as the atom $R(1, 2), \{\tau|1, 2\} \in \{a, b\}$. When it comes to $\Sigma^*$, the intention is, for a TGD $\sigma \in \Sigma$, to encode the shape of the type $\tau$ of the guard of $\sigma$ in a predicate $[\tau]$, and then replace $\sigma$ with a

We can now show that chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma)$ can be computed in time $|D| \cdot f(||\Sigma||)$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

Indeed, for each linear TGD $\sigma \in \Sigma$, we need to scan the instance chase$_s^1(D, \Sigma)$, and check whether the atom body(\sigma) matches with an atom of chase$_s^1(D, \Sigma)$, and, if this is the case, add to the instance under construction $H_D$ new atoms. By Lemma A.2, for each $i \geq 0$,

[chase$_s^i(D, \Sigma) \leq$ chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma) \leq |D| \cdot f(||\Sigma|| \cdot H_D + 1)^{g(||\Sigma|| + ||q||)}$.

Thus, chase$_s^{g(||\Sigma|| + ||q||)}(D, \Sigma)$ can be computed in time $|D| \cdot g(||\Sigma||)$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

Before giving the proof of the above result, let us explain how Lemma A.1 and Lemma A.3 allow us to obtain item (3) of Proposition 3.3. Given an OMQ $Q = (S, \Sigma, q(x)) \in (G, UCQ)$, for $k \geq 1$, an S-database $D$, and $\bar{c} \in$ dom($D$)$^{|k|}$, by Lemmas A.1 and A.3, we simply need to construct a finite instance $I$ in time $|D|^{O(1)} \cdot f(||\Sigma||)$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$, and then check if $\bar{c} \in q(I)$. Since $q \in UCQ_k$, by Proposition 2.1, the overall process takes time $|D|^{O(1)} \cdot f(||\Sigma||)$ for some computable function $g : \mathbb{N} \rightarrow \mathbb{N}$.
linear TGD that uses in its body an atom of the form $[r](\cdot)$. However, we need an effective way to compute the type of an atom $a$ by completing its known part, which is inherited from the type of the guard atom that generates $a$, with atoms that mention the new nulls invented in $a$. This relies on the main property of the type mentioned above. In particular, for each $a \in \varepsilon(\varepsilon, \Sigma)$, obtained due to the application of the TGD $\sigma = \varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z}(\bar{x}, \bar{z}) \in \Sigma$ with the witness $(\bar{i}, \bar{o})$, we can construct $\text{type}_{\Sigma'}(\bar{z})$ from $\varphi(\bar{i}, \bar{o})$, where $\bar{o}$ is a tuple of new null values, together with the restriction of $\text{type}_{\Sigma'}(\bar{z})$ to the terms $\bar{i}$. This is precisely how we are going to generate new types from existing ones. Before giving the formal construction, we need to introduce some auxiliary terminology.

Auxiliary Terminology. Let $\text{sch}(\Sigma)$ be the set of predicates occurring in $\Sigma$. For an atom $a$, let $\text{base}(a, \Sigma)$ be the set of all atoms that can be formed using terms from $\text{dom}(a)$ and predicates from $\text{sch}(\Sigma)$. A $\Sigma$-type $\tau$ is a pair $(\alpha, \Sigma)$, where $\alpha = \{t_1, \ldots, t_n\} \in \text{sch}(\Sigma)$ and $t_1 \leq t_2 \leq \ldots \leq t_n$ for each $i \in \{\alpha, \beta\}$, and $\Sigma \subseteq \text{base}(\{t_1, \ldots, t_n\}, \Sigma')$. We write $\text{type}(\tau)$ for the atom $a$, and $\text{atoms}(\tau)$ for $(\alpha \cup \beta)$. The arity of $\tau$, denoted $\text{ar}(\tau)$, is the maximum integer occurring in $\text{guard}(\tau)$. Intuitively, $\tau$ encodes the shape of a guard atom $\alpha$ and a set of side atoms that are “covered” by $\beta$. Consider a tuple $\bar{u} = (u_1, \ldots, u_n)$. Suppose the istomophic to $I = (t_1, \ldots, t_n)$, written $\bar{u} \simeq I$, which means that $u_i = u_j$ iff $t_i = t_j$. The instantiation of $\tau$ with $\bar{u}$, denoted $\text{chase}(\tau, \bar{u})$, is the set of atoms obtained from atoms of $\alpha$ and $\beta$ with each $t_i$ with $u_i$. The projection of $\tau$ over $P \subseteq \{1, \ldots, \text{ar}(\tau)\}$ is the set of atoms $\Pi_\tau(\alpha) = \{\beta \in \text{atoms}(\tau) \mid \text{dom}(\beta) \subseteq P\}$. The completion of an instance $I$ w.r.t. $\Sigma$, denoted $\text{complete}(I, \Sigma)$, is defined as the instance $\{(\alpha_1, \cdot) \mid \text{dom}(\alpha_1) \subseteq \text{sch}(\Sigma)\}$.

The Formal Construction. The new database is defined as $D^* = \{(r)|.\} \cup \{\bar{t}\}|.\bar{t} \in D, \forall \bar{t} \in \text{complete}(\Sigma, \Sigma)\mid \text{ar}(\Sigma)\}$. The new set $\Sigma^*$ of linear TGDs is the union of $\Sigma^G$ and $\Sigma^E$, where (1) $\Sigma^G$ is the type generator, i.e., is responsible for generating new $\Sigma$-types from existing ones. (2) $\Sigma^E$ is the expander, it expands a derived $\Sigma$-type $\tau$, i.e., it explicitly constructs the atoms over $\text{sch}(\Sigma)$ encoded by $\tau$. The type generator is defined as follows. For each $\sigma \in \Sigma$

$\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z}_1 \cdots \exists \bar{z}_m \cdot R_1(\bar{u}_1), \ldots, R_n(\bar{u}_n)$

with guard($\sigma$) $= G(\bar{u})$, and $\varphi(\bar{z}_i)$ $\in \text{base}(\Sigma)$, and for every $\Sigma$-type $\tau$ such that there is a homomorphism $h$ from $\varphi(\bar{x}, \bar{y})$ to $\text{atoms}(\tau)$ and $h(G(\bar{u})) = \text{guard}(\tau)$, we add to $\Sigma^G$

$\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z}_1 \cdots \exists \bar{z}_m \cdot \{\bar{t}_1(\bar{u}_1), \ldots, \bar{t}_N(\bar{u}_N)\}$

where, having the function $\rho$ with $\rho(h(t))$ if $t \in \bar{x}$

$\rho(\alpha) + \text{id}$ if $t \in \bar{x}$

as well as the function $\rho$ that renames the integers in $\alpha$ in order to appear in increasing order starting from 1 (e.g., $\rho(\{(2, 1, 3)\}) = \{(1, 2, 3)\}$) for each $i \in [n]$, with $\alpha_i = R_{t_i}(f(\bar{u}_i))$, the $\Sigma$-type $\tau$ is

$I = \rho(\alpha_1, \ldots, \alpha_n)|.\bigcup \{\text{atoms}(\alpha_i)\}$

with

$I = \rho(\alpha_1, \ldots, \alpha_n)|.\bigcup \{\text{atoms}(\alpha_i)\}$

The expander actually constructs, for each $\Sigma$-type $\tau$, the guard atom of $\tau$. More precisely, for each $\Sigma$-type $\tau$, we add to $\Sigma^E$

$\varphi(\bar{x}, \bar{y}) \rightarrow R(\bar{x}, \bar{y})$

where $R$ is the $k$-ary predicate of guard($\tau$).

This completes the construction of $\Sigma^*$. It is clear that $\Sigma^* \subseteq \Sigma$. It also is not difficult to show, by exploiting the properties of the type, that $\Omega(D) = q(\text{chase}(D^*, \Sigma^*))$, as needed. It remains to show that $D^*$ and $\Sigma^* \subseteq \Sigma$ can be constructed in the claimed times.

An FPT Construction. By construction, $\Sigma^*$ depends only on $\Sigma$ and thus, is clear that it can be computed in time $\text{g}(||D||)$ for some computable function $g : \mathbb{N} \rightarrow \mathbb{N}$. On the other hand, $D^*$ depends both on $\Sigma$ and $\Sigma$. It remains to show that $D^*$ can be computed in time $|D^*| \cdot |\text{ar}(\Sigma)|$ for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

To this end, we first show the following auxiliary result:

Lemma A.4. For a database $D'$ and set $\Sigma' \subseteq \mathbb{N}$, the (finite) instance chase($D', \Sigma'$) can be constructed in time $|D'| \cdot |\text{ar}(\Sigma')|$, for some computable function $g : \mathbb{N} \rightarrow \mathbb{N}$.

Proof. We provide similar analysis as in the proof of Lemma A.1. Let us first establish an upper bound on the size of chase($D', \Sigma'$). We assume that $\Sigma'$ is over the schema $T$. We can show that:

$|\text{chase}(D', \Sigma')| \leq |D'| \cdot |T| \cdot \text{ar}(\Sigma')$

Observe that, due to guardedness, two constants $c, d \in \text{dom}(D')$ can occur together in an atom of chase($D', \Sigma'$) only if they already occur together in an atom of $D'$. Thus, for each $n$-tuple $\bar{c}$ occurring in $D'$ (i.e., there is an atom $R(\bar{c})$ in $D'$), the chase can produce at most $\text{ar}(T)^n$ new tuples of constants from $\bar{c}$, where each such new tuple can be stored in a predicate of T. Since $\text{ar}(T) \leq \text{ar}(\Sigma')$, we get that chase($D', \Sigma'$) can have at most $\text{ar}(T)^n \cdot |\text{ar}(\Sigma')| \leq \text{ar}(T)^n \cdot |\text{ar}(\Sigma')|$

We can now show that indeed chase($D', \Sigma'$) can be constructed in time $|D'| \cdot |\text{ar}(\Sigma')|$. Let $B_{\Sigma'}$ (resp., $H_{\Sigma'}$) be the maximum number of atoms in the body (resp., head) of a TGD of $\Sigma'$. Observe that, for $i \geq 0$, chase$_{i-1}(D', \Sigma')$, where $s$ is a level-wise chase sequence for $D'$ under $\Sigma'$, can be computed from chase$_{i-1}(D', \Sigma')$ in time

$\text{chase}_{i-1}(D', \Sigma') \cdot \text{g}(\text{ar}(\Sigma') + 1) \cdot \text{ar}(\Sigma') + \text{g}(\text{chase}_{i-1}(D', \Sigma')) \cdot \text{ar}(\Sigma') + \text{g}(\text{chase}_{i-1}(D', \Sigma')) \cdot \text{ar}(\Sigma')$

Indeed, for each guarded TGD $\sigma \in \Sigma'$, we need to scan the instance chase$_{i-1}(D', \Sigma')$, and check whether the atom guard($\sigma$) matches with an atom $\alpha \in \text{chase}_{i-1}(D', \Sigma')$. Note that the matching of guard($\sigma$) with $\alpha$ uniquely determines the atoms $\alpha_1, \ldots, \alpha_m$ with which the atoms of body($\sigma$) apart from guard($\sigma$) should match. Then, for each such atom $\alpha_i$, we need to scan chase$_{i-1}(D', \Sigma')$, and check that indeed it is present. If all the atoms $\alpha_1, \ldots, \alpha_m$ occur in chase$_{i-1}(D', \Sigma')$, then we add $H_{\Sigma'}$ new atoms.

As shown above, for each $i \geq 0$, $|\text{chase}_{i}(D', \Sigma')| \leq |\text{chase}(D', \Sigma')| \leq |D'| \cdot |T| \cdot \text{ar}(\Sigma')$. |
Since chase($D', \Sigma'$) can have at most $n = |D'| \cdot |T| \cdot \text{ar}(T)^{\text{ar}(T)}$ s-levels, it can be computed in time
\[ n \cdot |\Sigma'| \cdot (n \cdot (\text{ar}(T) + 1) + B_{\Sigma'} \cdot (\text{ar}(T) + 1) + n \cdot B_{\Sigma'} \cdot (\text{ar}(T) + 1)) + H_{\Sigma'} \cdot (\text{ar}(T) + 1). \]
Summing up, chase($D', \Sigma'$) can be computed in time
\[ ||D'||^2 \cdot g(||\Sigma'||) \]
for some computable function $g : \mathbb{N} \to \mathbb{N}$, and the claim follows. \hfill \Box

In view of Lemma A.4, to show that $D'$ can be computed in time $||D'||^{O(1)} \cdot f(||Q||)$, it suffices to show that we can construct in time $g(||\Sigma'||)$ a set $\widehat{\Sigma} \in \mathcal{S} \cap \text{FULL}$ such that, for every database $D'$, chase($D', \widehat{\Sigma}$) $\subseteq$ chase($D, \Sigma$), where $\mathcal{S}$ consists of all the predicates of the form $[\tau]$, where $\tau$ is a $\Sigma$-type, occurring in chase($D, \Sigma$). The set $\widehat{\Sigma}$ will simply:

1. (complete the database $D$ with all the atoms of chase($D, \Sigma$)) that mention only constants from dom($D$), i.e., add to $D$ the set of atoms chase($D, \Sigma$) $\cap$ base($D, \Sigma$), and
2. (generate all the atoms of $D'$).

For achieving task (1) above, we exploit a result from [24], which states the following: given a set $\Sigma' \in \mathcal{S}$, we can construct a set $\xi(\Sigma') \in \mathcal{S} \cap \text{FULL}$ over sch($\Sigma'$) such that, for every database $D'$, chase($D', \xi(\Sigma')$) $\subseteq$ chase($D, \Sigma$) $\cap$ base($D, \Sigma$). Thus, task (1) can be done via the set $\xi(\Sigma) \in \mathcal{S} \cap \text{sch}$. Task (2) can be achieved via the set of TGDs $\Sigma_{\text{types}} \in \mathcal{S} \cap \text{FULL}$ defined as follows. For each $\Sigma$-type $\tau$, we have a TGD $\tau(x_1, \ldots, x_k) \to [\tau]\{x_f_1(i), \ldots, x_f_x(i)\}$, where $k = \text{ar}(\tau)$, $i$ is the arity of the predicate of guard($\tau$), and, assuming that $(i_1, \ldots, i_p)$ is the tuple of guard($\tau$), $x_f_x(i) = i_j$, for each $j \in [\ell]$. By abuse of notation, we use $\tau(x_1, \ldots, x_k)$ to denote the conjunction of atoms in the instantiation of $\tau$ with $(x_1, \ldots, x_k)$.

The set $\widehat{\Sigma}$ is defined as $\xi(\Sigma) \cup \Sigma_{\text{types}}$, which can be clearly constructed in time $g(||\Sigma||)$ since it depends only on $\Sigma$.

B PROOF OF THEOREM 5.1

Given a CQ $q$, we write var($q$) for the set of variables in $q$. We may also treat, as usual, a CQ $q$ as the set of atoms in $q$. The proofs in this section rely on notions and results introduced in the proof of Proposition 5.2, which we advise the reader to read first.

The 2EXPTime-hardness is inherited from [7], where the same problem for OMQs based on DLs has been studied. It remains to establish the 2EXPTime upper bound. Let $Q = (S, \Sigma, q)$. Proposition 5.2 provides a procedure for deciding (uniform) $\text{UCCQ}_k$-equivalence:

1. (compute the $\text{UCCQ}_k$-approximation) $Q_k = (S, \Sigma, q_k)$ of $Q$;
2. (accept if $Q \equiv Q_k^\ast$; otherwise, reject).

This yields decidability. It also shows that if $Q$ is $\text{UCCQ}_k$-equivalent, then an OMQ $Q'$ from $(\mathcal{G}, \text{UCCQ}_k)$ such that $Q \equiv Q'$ can be constructed in double exponential time. Actually, to find $Q'$ we can simply construct the OMQ $Q_k^\ast$.

Lemma B.1. $Q_k^\ast$ can be constructed in double exponential time.

Proof. For each disjunct $p$ of $q$, we first need to compute its specializations ($p', V$). It is easy to see that there is single exponentially many specialization that can be found in single exponential time. For each specialization $s = (p', V)$, we then have to find all $\Sigma$-groundings $g_s(\bar{x})$ of $s$, and keep the ones of treewidth at most $k$. Note that we can form double exponentially many guarded CQs, each of exponential size. Thus, there are double exponentially many candidates for being a $\Sigma$-grounding of $s$, each for exponential size, that can be found in double exponential time. Now, in the check whether such a candidate $g$ is indeed a $\Sigma$-grounding of $s$, the only non-trivial part is that we have to check for each maximally $|V|$-connected component $p\_i$ of $p'$, whether $p_i \to \text{chase}(g_i, \Sigma)$, where $g_i$ is a guarded subquery of $g$ of exponential size, via a homomorphism that is the identity on var($p\_i$) $\cap$ $V = \{y_1, \ldots, y_m\}$. This boils down to the problem whether $(y_1, \ldots, y_m) \in p\_i'$ (chase($g_i, \Sigma$)), where $p\_i'$ consists of the atoms of $p_i$. By item (2) of Proposition 3.2, this can be checked in 2EXPTime. Actually, this is not immediate from Proposition 3.2 since $g_i$ might be of exponential size. We additionally need to say that the double exponential bound obtained from Proposition 3.2 is only polynomial in the size of the database; implicit in [14]. Finally, checking whether a $\Sigma$-grounding $g_s$ of $s$ has treewidth at most $k$ is feasible in double exponential time since $g_s$ is, in general, of exponential size. \hfill \Box

Containment between OMQs from $(\mathcal{G}, \text{UCQ})$ can be decided in 2EXPTime [6]. Since, however, $Q_k^\ast$ may consists of double exponentially many CQs, it is not clear how to implement step (2) above in 2EXPTime and how to obtain the 2EXPTime upper bound in Theorem 5.1 by a direct implementation of the above procedure. We solve this problem in two steps. First, we replace $Q_k^\ast$ with an OMQ $Q'_k = (S, \Sigma, q_k')$ that is equivalent to $Q_k^\ast$, and such that $q_k'$ contains only single exponentially many CQs, each of polynomial size. And second, we modify the containment check from [6] in a mild way.

B.1 Replacing $Q_k^\ast$ with $Q_k'$

The OMQ $Q'_k = (S, \Sigma, q_k')$ is defined as follows. We introduce a fresh unary relation symbol $A$ (that is not in $S$) and obtain $\Sigma'$ from $\Sigma$ by extending every TGD head with the atom $A(x)$ whenever $x$ is an existentially quantified variable in that TGD head. Then $q_k'$ is the UCQ that includes $p_i$ for all specializations $s = (p, V)$ of a CQ $p$ in $q$ such that there exists a $\Sigma$-grounding of $s$ that is of treewidth at most $k$, extended by adding $A(x)$ for every variable $x \in \text{var}(p\_i) \cap V$.

It is clear that $q_k'$ has only single exponentially many disjuncts, each of polynomial size. Reusing the arguments from the proof of Lemma B.1, it is also clear that $Q'_k$ can be constructed in double exponential time. We further make the following important observation, which rests on the assumption that $k \geq \text{ar}(T) + 1$.

Lemma B.2. Let $s = (p\_v, V)$ be a specialization of a CQ in $q$. If there is a $\Sigma$-grounding of $s$ that has treewidth at most $k$, then all $\Sigma$-groundings of $s$ have treewidth at most $k$.

Proof. Let $p\_1, \ldots, p\_n$ be the maximally $|V|$-connected components of $p\_V$, let $g_s(\bar{x}) = \exists \bar{y} \in \text{var}(p\_v) \cap V \subseteq \chi(\bar{v})$. Now, let $g'_s(\bar{x}) = \exists \bar{y} \in \text{var}(p\_v) \cap V \subseteq \chi'(\bar{v})$. From $g_s(\bar{x})$, for $i \in [n]$, must be a $v_i \in V\_A$ with var($p\_i\_v$) $\cap V \subseteq \chi(\bar{v}_i)$. Now, let $g'_s(\bar{x}) = \exists \bar{y} \in \text{var}(p\_v) \cap V \subseteq \chi'(\bar{v}_i)$. A fresh successor $u_i$ to $v_i$ and setting $\chi'(u_i) = \text{var}(g'_s)$.
verified that $\delta'$ is a tree decomposition of $Q'_k$, $q'$ the existentially quantified variables of $Q'_k$. In particular, Condition (2) of tree decompositions is satisfied for every edge induced by an atom in some $q'_i$ due to the presence of $u_i$. Since $g'_i$ is guarded, the number of variables in $\text{var}(q'_i)$ is bounded by $\text{ar}(T)$. Since $k \geq \text{ar}(T) - 1$, the width of $\delta'$ is bounded by $k$, and the claim follows.

We now proceed to show that $Q_k^p$ and $Q'_k$ are indeed equivalent.

**Lemma B.3.** $Q_k^p \equiv Q'_k$.

**Proof.** First assume that $\bar{c} \in Q_k^p(D)$ with $D$ an $S$-database. Then, there is a homomorphism $h$ from a CQ $p(\bar{x})$ in $Q_k^p$ to chase($D, \Sigma'$) with $h(\bar{x}) = \bar{c}$, i.e., $p$ was constructed for the specialization $s = (p_c, V)$, that is, $p$ is the extension of $p_c$ with $A(x)$ for all $x \in \text{var}(p_c) \setminus V$. By construction of $\Sigma'$, $h$ thus maps all variables in $\text{var}(p)$ to $V$ to constants that the chase has introduced to satisfy existential quantifiers in TGD heads. Let $p_1, \ldots, p_n$ be the maximally $[V]$-connected components of $p(\bar{V})$. It follows from the facts that $p_1$ is $[V]$-connected, and that $h$ maps all variables in $\text{var}(p)$ to existential constants that we find a fact $\alpha_i \in D$ such that the restriction of $h$ to the variables in $p_1$ is a homomorphism to chase($\text{type}_{p_1}(\alpha_i), \Sigma'$), for $i \in [n]$. Let the $\Sigma'$-grounding $g'_k$ of $s$ by $\alpha_i$ be a full CQ. Clearly, the restriction of $h'$ to the variables in $V$ can be extended to a homomorphism $h''$ from $g'_k$ to chase($D, \Sigma'$) by mapping the variables from the CQs $g_i$ that are not in $V$ to constants from $\alpha_i$ in the expected way. Since $g'_k$ does not contain the relation $A$, $h'$ maps $g_i$ to chase($D, \Sigma$). It remains to argue that $g'_k$ is a CQ in $Q'_k$. Since $p$ is in $q'_k$, there is a $\Sigma'$-grounding of $s$ that is of treewidth at most $k$. By Lemma B.2, this implies that $g'_k$ is of treewidth at most $k$ and thus $g'_k$ is in $q'_k$. Conversely, assume that $\bar{c} \in Q'_k(D)$ with $D$ an $S$-database. Then, there is a homomorphism $h$ from a CQ $g(\bar{x})$ in $q'_k$ to chase($D, \Sigma$) with $h(\bar{x}) = \bar{c}$. Let $g_1$ be a $\Sigma'$-grounding of the specialization $s = (p_c, V)$ of a CQ $p$ in $q'_k$. Then $g_1$ is of treewidth at most $k$. Consequently, $g_1$ gives rise to a corresponding CQ $p_1$ from $q'_k$, that is, $p_1$ is the extension of $p_c$ with $A(x)$ for all $x \in \text{var}(p_c) \setminus V$. Clearly, the restriction of $h_1$ to the variables in $V$ is a homomorphism from $p_1$ to chase($D, \Sigma$). Let $p_1, \ldots, p_n$ be the maximally $[V]$-connected components of $p_1(\bar{V})$. By definition of $\Sigma'$-groundings, $p_1 \rightarrow$ chase($g_1, \Sigma$) via a homomorphism that is the identity on $\text{var}(p_1) \setminus V$, for $i \in [n]$. We can combine all these homomorphisms with $h_1$, to obtain a homomorphism $h''$ from $p_1$ to chase($D, \Sigma'$) that maps all variables outside of $V$ to constants that have been introduced by the chase to satisfy existential quantifiers in TGD heads. By construction of $p'$ and $\Sigma'$, $h''$ is also a homomorphism from $p'$ to chase($D, \Sigma'$).

**B.2 Checking Equivalence**

We now proceed to analyze the complexity of checking whether $Q \equiv Q'_k$. By item (1) of Lemma C.7 and Lemma B.3, $Q'_k \subseteq Q$. The non-trivial task is to check whether $Q \subseteq Q'_k$. We know from [6] that checking containment among OMQs from $(G, UCC)$ is in $2\text{ExpTime}$. In fact, [6] first provides a polynomial time reduction from containment among OMQs from $(G, UCC)$ to containment among CQs from $(G, \Sigma)$, denoted $\text{Cont}(G, UCC)$, to containment among OMQs from $(G, \Sigma)$, denoted $\text{Cont}(G, \Sigma)$, and then devises an automata-based procedure for $\text{Cont}(G, \Sigma)$ that runs in double exponential time in the combined size of the OMQs. However, $Q'_k$ might be of exponential size, which implies that if we first apply the reduction from $\text{Cont}(G, UCC)$ to $\text{Cont}(G, \Sigma)$, we get an OMQ that has a CQ of exponential size, and then, by using the decision procedure for $\text{Cont}(G, \Sigma)$ from [6] as a black box, we only get a $3\text{ExpTime}$ upper bound. Nevertheless, it turns out that we can reuse the automata constructions from [6] for $\text{Cont}(G, \Sigma)$ in order to devise a procedure that directly operates on OMQs from $(G, UCC)$, and decides containment in time double exponential in the size of the ontology, the schema, and the maximum size of the CQs in the UCCs of the OMQs, but only single exponential in the number of disjuncts in those UCCs. Having such a procedure in place, it is then clear that checking whether $Q \subseteq Q'_k$ is in $2\text{ExpTime}$ since, due to Lemma B.1, and since although the UCC in $Q'_k$ has exponentially many disjuncts, each one is of polynomial size. For the sake of completeness, we recall the key ingredients underlying the automata-based procedure from [6], and then explain how we get the automata-based procedure that directly operates on OMQs from $(G, UCC)$ that leads to the desired upper bound.

**The Automata-based Procedure for $\text{Cont}(G, UCC)$**

Let us first recall that we can focus, w.l.o.g., on Boolean OMQs since there is an easy reduction, originally proposed in [10], of containment for non-Boolean OMQs to containment for Boolean OMQs. In what follows, we write $\text{UCQ}$ and $\text{CQ}$ meaning the classes of Boolean UCQs and Boolean CQs, respectively.

We know from [6] that, given two OMQs $Q_1$ and $Q_2$ from $(G, \Sigma)$, if $Q_1 \not\subseteq Q_2$, then this is witnessed via a "nearly acyclic" database. To formalize this we need the notion of guarded tree decomposition. A tree decomposition $\delta = (T, \chi)$, where $T_k = (V_k, E_k)$, of $G^D$ for some database $D$ is called $[V]$-guarded, where $V \subseteq V_k$ if for every node $v \in V_k \setminus V$, there exists a fact $R(c_1, \ldots, c_n) \in D$ such that $\chi(v) \subseteq \{c_1, \ldots, c_n\}$, i.e., $\chi(v)$ is guarded in $D$. We write $\text{root}(\delta)$ for the root node of $\delta$, and $\text{root}(v)$, for $v \in V_k$, for the subset of $D$ induced by $\chi(v)$. An $S$-database is called a $C$-tree, where $C \subseteq D$, if there exists a tree decomposition $\delta$ of $G^D$ such that: (i) $D_\delta(\text{root}(\delta)) = C$, and (ii) $\delta$ is $[\{\text{root}(\delta)\}]$-guarded. Roughly, if a database $D$ is a $C$-tree, then $C$ can be viewed as the cyclic part of $D$, while the rest of $D$ is acyclic. In [6], it has been shown that for containment purposes we can focus on databases that are $C$-trees with the size of $C$ bounded and depending only on the left-hand side OMQ. Note, however, that the result of [6] talks only about OMQs from $(G, \Sigma)$. A careful inspection of the proof given in [6] reveals that the following statement for OMQs from $(G, UCC)$ holds:

**Lemma B.4.** Let $Q_1 = (S, \Sigma_1, q_1)$ and $Q_2 = (S, \Sigma_2, q_2)$ be OMQs from $(G, UCC)$. The following are equivalent:

1. $Q_1 \not\subseteq Q_2$.
2. There exists a $C$-tree $S$-database $D$ with $||\text{dom}(C)|| \leq \text{ar}(S \cup \text{sch}(G)) \cdot \max_{p \in q_1} ||p||$ such that $Q_1(D) \not\subseteq Q_2(D)$.

The goal is to devise a two-way alternating parity tree automaton (2ATA) that checks for the second item of Lemma B.4. To this end, we need a convenient encoding of a $C$-tree database as a tree over a finite alphabet that can be accepted by a 2ATA. As shown in [6], this can be done by exploiting standard encodings of bounded treewidth
CQ Specializations

Recall that a contraction of a CQ $q(x)$ is a CQ $p(x)$ that can be obtained from $q$ by identifying variables. When an answer variable $x$ is identified with a non-answer variable $y$, the resulting variable is $x$; the identification of two answer variables is not allowed. The formal definition of CQ specialization follows:

**Definition C.1.** Consider a CQ $q(x)$. A specialization of $q$ is a pair $s = (p, V)$, where $p$ is a contraction of $q$, and $x \subseteq V \subseteq \text{var}(p)$.

Intuitively speaking, a specialization of a CQ $q$ describes a way how $q$ can be homomorphically mapped to the chase of a database $D$ under a set $\Sigma$ of TGDs. First, instead of mapping $q$ to chase($D, \Sigma$), we injectively map $p$ to chase($D, \Sigma$), while the set of variables $V$ collects all the variables of $p$ that are mapped to constants of dom($D$). The next result, which relies on the properties of the type of an atom (see the proof of Lemma A.1 in Appendix ??) illustrates the usefulness of CQ specializations. A CQ $q$ is $[V]$-connected, for $V \subseteq \text{var}(q)$, if $G^q_{\mid \text{var}(q)} \setminus V$, that is, the subgraph of $G^q$ induced by $\text{var}(q) \setminus V$, is connected. We also write $q[V]$ for the subquery of $q$ obtained after dropping the atoms of $q[V]$, i.e., the atoms that contain only variables of $V$. Given a database $D$, and a set $\Sigma$ of TGDs, we write chase($D, \Sigma$) for the ground part of chase($D, \Sigma$), that is, the set consisting of all atoms in chase($D, \Sigma$) with only constants from dom($D$).

**Lemma C.2.** Consider a database $D$, a set $\Sigma \in \mathbb{G}$, a CQ $q(x)$, and a tuple $\bar{e} \in \text{dom}(D)^{|\bar{x}|}$. The following are equivalent:

1. $\bar{q}_1 \subseteq \bar{q}_2$ and $L(A) = 0$.
2. $A$ has exponentially many states in $||Q_1^{\text{max}}||$ and $\ell$, and polynomially many states in $||q_1|| + ||q_2||$. It can be constructed in double exponential time in $||Q_1^{\text{max}}||$ and $\ell$, and in polynomial time in $||q_1|| + ||q_2||$.

**Proof.** Let $\ell = \lceil \text{ar}(S \cup \text{sch}((\Sigma_1))) \cdot \max_{p \in q_1} ||p|| \rceil$. Then $Q_1 = \text{L}((C\bar{q}_1))^\ell \cap (\bigcup_{p \in q_1} A((S, \Sigma_1, p), \ell))$.

Using Lemmas B.5, B.6, and B.7, the fact that union and complementation of 2ATA are feasible in polynomial time, and the stated bound on the number of states in automata for union, it is easy to verify that indeed $A$ is the desired 2ATA.

It is now easy to see, having Lemma B.8 in place, that indeed checking whether $Q \subseteq Q'$ is feasible in double exponential time. This actually follows from the fact that the UCQ of $Q'$ has exponentially many disjuncts, each of polynomial size, and the fact that emptiness for 2ATA can be decided in exponential time in the number of states of the automaton.

**C PROOF OF PROPOSITION 5.2**

UCQ$_q$-approximations for OMQs from $(\mathbb{G}, \text{UCQ})$ rely on the notion of specialization of a CQ. We first introduce CQ specializations, and give a lemma that illustrates their usefulness. We then define the notion of grounding of a CQ specialization, and establish a technical lemma that will allow us to define UCQ$_q$-approximations. We then establish our main technical result about UCQ$_q$-approximations, which in turn allows us to complete the proof of Proposition 5.2. In what follows, given a CQ $q$, we write $\text{var}(q)$ for the set of variables in $q$. We may also treat, as usual, a CQ $q$ as the set of atoms in $q$. 

**C.1 CQ Specializations**
(1) $\varepsilon \in q(chase(D, \Sigma)).$

(2) There exists a specialization $s = (p, V)$ of $q$ such that $p \rightarrow chase(D, \Sigma)$ via an injective homomorphism $\mu$ with $(i) \mu(\bar{\varepsilon}) = \bar{\varepsilon}$, (ii) $\mu(p) \subseteq chase(D, \Sigma)$, and (iii) for every maximally $[V]$-connected component $p'$ of $p[V]$, there exists an atom $\alpha \in D$ such that $\text{dom}(\alpha) \supseteq \{ y | y \in \text{var}(p') \cap V \}$ and $\mu(p') \subseteq chase(\text{type}_{D,\Sigma}(\alpha), \Sigma)$.

C.2 Grounding Specializations

We now define the notion of grounding of a specialization $(p, V)$ of a CQ $q$. The intention is to replace each maximally $[V]$-connected component $p'$ of $p[V]$ with a guarded set of atoms, which is ought to be mapped to the ground part of the chase, that entails $p'$ under the given set of TGDs. A CQ $q$ is guarded if it has an atom, denoted $\text{guard}(q)$, that contains all the variables of $\text{var}(q)$, and full if all the variables of $\text{var}(q)$ are answer variables.

Definition C.3. Consider a set $\Sigma$ of TGDs over $T$, and a CQ $q(\bar{x})$ over $T$. Let $s = (p, V)$ be a specialization of $q$ with $p_1, \ldots, p_n$ being the maximally $[V]$-connected components of $p[V]$. A $\Sigma$-grounding of $s$ is a CQ $g_s(\bar{x})$ over $T$ of the form

$$\exists \bar{\varepsilon} \ (g_0 \land g_1 \land \cdots \land g_n)$$

where

- $g_0$ is the full CQ consisting of the atoms of $p[V]$,
- for each $i \in [n]$, $g_i$ is a guarded full CQ such that:
  - $\text{var}(g_i) \subseteq (\text{var}(p_i) \cap V) \cup \{ y_i', \ldots, y_{\text{ar}(T)}' | \}$, $V$,
  - $\text{var}(p_i) \cap V \subseteq \text{var}(\text{guard}(g_i))$, and
  - $p_i \rightarrow chase(g_i, \Sigma)$ via a homomorphism that is the identity on $\text{var}(p_i) \cap V$, and
- $\bar{\varepsilon} = (\bigcup_{0 \leq i \leq n} \text{var}(g_i)) \setminus \bar{x}$.

Before giving the main lemma concerning groundings of CQ specializations, let us establish a useful auxiliary claim.

Claim C.4. Let $\Sigma$ be a set of TGDs over $T$, and $q(\bar{x})$ a CQ over $T$. Let $g_s(\bar{x})$ be a $\Sigma$-grounding of a specialization $s = (p, V)$ of $q$, and assume that $g_s \rightarrow I$, for an instance $I$, via a homomorphism $\mu$. Then, $p \rightarrow chase(I, \Sigma)$ via a homomorphism $\mu'$ with $\mu(\bar{\varepsilon}) = \bar{\varepsilon}$.

Proof. Let $p_1, \ldots, p_n$ be the maximally $[V]$-connected components of $p[V]$. Assume that $g_s(\bar{x})$ is of the form

$$\exists \bar{\varepsilon} \ (g_0 \land g_1 \land \cdots \land g_n)$$

where $g_0$ is the full CQ consisting of the atoms of $p[V]$, and for each $i \in [n]$, $g_i$ is a guarded full CQ with the properties given in Definition C.3. Clearly, for each $i \in [n]$, there exists a mapping $\lambda_i$ that is the identity on $\text{var}(p_i) \cap V$ that maps $p_i$ to $\text{chase}(g_i, \Sigma)$. Since the sets of atoms $p[V], p_1, \ldots, p_n$ share only variables of $V$, we conclude that $\lambda = \bigcup_{i \in [n]} \lambda_i$ is a well-defined mapping that maps $p$ to $\text{chase}(g_s, \Sigma)$ with $\lambda(\bar{\varepsilon}) = \bar{\varepsilon}$. Since, by hypothesis, $g_s \rightarrow I$ via some $\mu$, it is not difficult to show that $\text{chase}(g_s, \Sigma)$ maps to $\text{chase}(I, \Sigma)$ via $\mu'$ that extends $\mu$, and such, $\mu(\bar{\varepsilon}) = \mu'(\bar{\varepsilon})$. Therefore, $\lambda = \mu' \circ \lambda$ maps $p$ to $\text{chase}(I, \Sigma)$ with $\mu(\bar{\varepsilon}) = \xi(\bar{\varepsilon})$, as needed.

C.3 UCQk-approximations

We proceed to introduce the notion of UCQk-approximation for OMQs from (G, UCQ), which relies on the notion of grounding of a CQ specialization. Roughly, the UCQk-approximation of an OMQ $Q = (S, \Sigma, q)$, for $k \geq 1$, is the OMQ $Q_k^k$ obtained from $Q$ by replacing each disjunct $p$ of $q$ with the UCQ consisting of all the $\Sigma$-groundings of all specializations of $p$ with treewidth at most $k$.

Definition C.6. Consider an OMQ $Q = (S, \Sigma, q)$ from (G, UCQ) over a schema $T$. The UCQk-approximation of $Q$, for $k \geq 1$, is the OMQ $Q_k^k = (S, \Sigma, q_k^k)$, where $q_k^k$ is obtained from $q$ by replacing each disjunct $p$ of $Q$ with the UCQ consisting of all the $\Sigma$-groundings of $p$ over $T$ with treewidth at most $k$ of all specializations of $p$.

The main technical lemma concerning UCQk-approximations follows. This relies on the notion of $k$-unraveling of a database $D$ up to a certain tuple $c$ over $\text{dom}(D)$, which can be defined in the same way as in [7]. In particular, the $k$-unraveling of $D$ up to $c$, for some $k \geq \text{ar}(S) - 1$, denoted $D_k^c$, is a possibly infinite $S$-instance for which the following properties hold:

(1) The subgraph of $G(D)$ induced by the elements of $\text{dom}(D_k^c) \setminus \tilde{c}$ has treewidth $k$. 

Lemma C.5. Consider a database $D$, a set $\Sigma \in \Theta$ over a schema $T$, a CQ $q(\bar{x})$ over $T$, and $\bar{c} \in \text{dom}(D)[\bar{x}]$. The following are equivalent:

(1) $\bar{c} \in q(chase(D, \Sigma))$.

(2) There exists a specialization $s$ of $q$ such that $\Sigma$-grounding $g_s(\bar{\varepsilon})$ of $s$ maps to $\text{chase}(D, \Sigma)$ via an injective mapping $\mu$ with $\mu(\bar{\varepsilon}) = \bar{\varepsilon}$.

Proof. (1) $\Rightarrow$ (2). By Lemma C.2, there is a specialization $s = (p, V)$ of $q(\bar{x})$, with $p_1, \ldots, p_n$ being the maximally $[V]$-connected components of $p[V]$, such that $p \rightarrow chase(D, \Sigma)$ via an injective mapping $\lambda$ with $(i) \lambda(\bar{\varepsilon}) = \bar{\varepsilon}$, (ii) $\lambda(p) \subseteq chase(D, \Sigma)$, and (iii) for every maximally $[V]$-connected component $p'$ of $p[V]$, there exists an atom $\alpha \in D$ such that $\text{dom}(\alpha) \supseteq \{ y | y \in \text{var}(p') \cap V \}$ and $\lambda(p') \subseteq chase(\text{type}_{D,\Sigma}(\alpha), \Sigma)$. Consider the CQ $g_s(\bar{x})$ of the form

$$\exists \bar{\varepsilon} \ (g_0 \land g_1 \land \cdots \land g_n)$$

where $g_0$ is the full CQ consisting of the atoms of $p[V]$, and, for each $i \in [n]$, $g_i$ is the guarded full CQ obtained from $\text{type}_{D,\Sigma}(\alpha_i)$ by converting $\lambda(y)$, where $y \in \text{var}(p_i) \cap V$, into the variable $y_i$, and each constant $c_i \in \{ c_i, \ldots, c'_{i} \} = \text{dom}(\text{type}_{D,\Sigma}(\alpha_i)) \setminus \lambda(y)$, $y \in \text{var}(p_i) \cap V$, into the variable $y_i$. Finally, let $\bar{\varepsilon} = (\bigcup_{0 \leq i \leq n} \text{var}(g_i)) \setminus \bar{x}$. Clearly, $g_s(\bar{x})$ is a $\Sigma$-grounding of $s$. Moreover, the mapping $\mu = \lambda \cup \{ y_i' \mapsto c_i \}_{i \in [n], j \in [\ell]}$ maps injectively $g_s$ to $\text{chase}(D, \Sigma)$, since $\text{type}_{D,\Sigma}(\alpha_i) \subseteq \text{chase}(I, \Sigma)$, with $\mu(\bar{\varepsilon}) = \bar{\varepsilon}$.

(2) $\Rightarrow$ (1). Consider a specialization $s = (p, V)$ of $q$. Assume that there exists a $\Sigma$-grounding $g_s(\bar{x})$ of $s$ that maps to $\text{chase}(D, \Sigma)$ via an injective mapping $\mu$ with $\mu(\bar{\varepsilon}) = \bar{\varepsilon}$, actually, the fact that $\mu$ is injective is not crucial here. By Claim C.4, we get that $p$ maps to $\text{chase}(D, \Sigma) = chase(D, \Sigma)$ via a homomorphism $\bar{\lambda}$ with $\bar{\lambda}(\bar{x}) = c$. Clearly, by the definition of contractions, there exists a homomorphism $\bar{h}$, which is the identity on $\bar{x}$, that maps $q$ to $p$. Therefore, $\lambda = \bar{\lambda} \circ \bar{h}$ maps $q$ to $\text{chase}(D, \Sigma)$ with $\lambda(\bar{\varepsilon}) = \bar{\varepsilon}$, which implies that $\bar{c} \in q(chase(D, \Sigma))$.

We are now ready to show the main lemma about groundings of CQ specializations.
Consider an \( \bar{c} \) via a homomorphism that is the identity on \( \bar{c} \).

Let us stress that item (3) relies on the assumption that \( k \geq \ar(S) - 1 \) because, otherwise, there can be single atoms in \( D \) that cannot be part of any database of treewidth \( k \). In what follows, we say that \( D \) has treewidth \( k \) up to \( \bar{c} \) if the subgraph of \( G^D \) induced by the elements of \( \text{dom}(D) \setminus \{ \bar{c} \} \) has treewidth \( k \).

**Lemma C.7.** Let \( Q \) be an OMQ from \( (G, \text{UCQ}_k) \) with data schema \( \Sigma \), \( \bar{c} \in Q(D) \) implies \( \bar{c} \in Q^k(D) \).

(1) \( Q^k_k \subseteq Q \).

(2) For every S-database \( D \) and tuple \( \bar{c} \) over \( \text{dom}(D) \) such that \( D \) has treewidth at most \( k \) up to \( \bar{c} \), \( \bar{c} \in Q^k(D) \).

(3) For every OMQ \( Q' \) from \( (G, \text{UCQ}_k) \) such that \( Q \subseteq Q' \), \( \bar{c} \in \overline{Q}^k \), which implies that \( Q \subseteq Q^k_k \). By item (1) of Lemma C.7, we also get that \( Q^k_k \subseteq Q \), and the claim follows.

**C.5 The Case \( k < \ar(T) - 1 \)**

Notice that Theorem 5.1 and Proposition 5.2 only cover the case where \( k \geq \ar(T) - 1 \), and also Lemma C.7, which gives the main technical properties of \( \overline{\text{UCQ}}_k \)-approximations, requires this condition. We view the case \( k < \ar(T) - 1 \) as a somewhat esoteric corner case. In the context of Theorem 5.1, for example, we ask for an equivalent OMQ from \( (G, \text{UCQ}_k) \) when the arity of some relations \( R \) is so high that the actual UCQ in such an OMQ cannot contain a fact \( R(x_1, \ldots, x_n) \) unless some of the \( x_i \) are identical. We show that this case is also technically very different. In particular, our \( \overline{\text{UCQ}}_k \)-approximations do not serve their purpose.

We consider the case where \( k = 1 \), while \( \ar(S) = 3 \) and \( \ar(T) = 6 \) with \( S \) being the data schema and \( T \) being the extended schema. Let \( S = \{T_1, T_2, T_3\} \) with both \( T_1 \) and \( T_2 \) ternary. We use the Boolean CQ (for brevity, the existential quantifiers are omitted)

\[
q() = R(x_1, x_2) \land R(x_1, x_3) \land R(x_2, x_4) \land R(x_3, x_4) \land A_1(x_2) \land A_2(x_3)
\]

which is a core of treewidth \( 2 > k \). Our ontology takes the form

\[
\Sigma = \{T_1 \rightarrow q()\} \cup \Sigma_1 \cup \Sigma_2
\]

where \( \Sigma_1 \) makes sure that there is an S-path of length \( 2^n \) whenever there is a \( T_1 \)-atom and \( \Sigma_2 \) make sure that there is an S-path of length \( 2^n - 1 \) whenever there is a \( T_2 \)-atom, where \( S \) is a binary relation symbol. We need auxiliary relation symbols to establish these paths, but we will make sure that they are of arity at least three so that they cannot be ‘seen’ by \( \overline{\text{UCQ}}_k \)-approximations. Note that while the presence of a \( T_1 \)-atom implies that \( q() \) is true, the presence of a \( T_2 \)-atom does not. In detail, \( \Sigma_1 \) contains the following TGDs:

- \( T_1(x_1, x_2, x_3) \rightarrow B_{ij}^0(x_1, x_2, x_3) \) for \( 1 \leq i \leq n \).
- \( T_1(x_1, x_2, x_3) \rightarrow B_{ij}^0(x_1, x_2, x_3) \) for \( 1 \leq i \leq n \).

We leave the definition of \( \Sigma_2 \) to the reader.

This defines an OMQ \( Q = (S, \Sigma, q) \). It is not hard to see that \( Q \) is equivalent to an OMQ \( Q' \) from \( (G, \overline{\text{UCQ}}_1) \), namely to \( Q' = (S, \Sigma, q') \) where \( q'() \) is the Boolean CQ that asks for the existence of an S-path of length \( 2^n \). Note that the size of \( Q' \) is exponential in the size of \( Q \). This cannot be avoided unless we use an ontology different from \( \Sigma \) that is, resort to non-uniform UCQ₁-equivalence.

**Lemma C.8.** Let \( Q'' = (S, \Sigma, q'') \in (G, \overline{\text{UCQ}}_1) \) be equivalent to \( Q \). Then \( q'' \) must contain a CQ with at least \( 2^n \) atoms.
Lemma C.8 allows us to show that the UCQ₂ₜₖₐ₈ₐₙₐ₉ₚₛ₉ₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜₜportion of length $2^\alpha$. We aim to show that $p$ has at least $2^\alpha$ atoms. To this end, it suffices to show that for every edge $S(a_1, a_2)$ in the path chase$(D_1, \Sigma) \setminus \{T_1(c_1, c_2, c_3)\}$, there is an atom $S(x_1, x_2) \in p$ such that $h(x_i) = a_i$ for $i \in \{1, 2\}$. Assume that this is not the case. Then clearly we also find a homomorphism from $p$ into an $S$-path of length $2^\alpha - 1$, thus into chase$(D_2, \Sigma)$ where $D_2 = \{T_2(c_1, c_2, c_3)\}$. But $Q$ evaluates to false on $D_2$, which is a contradiction. \qed
Lemma D.4. Let $Q = (S, \Sigma, q) \in (G \cap \text{FULL, UCQ})$ and $\bar{a}$ a tuple of distinct constants. Then for every S-database $D$ with $D \models Q(\bar{a})$, there is an S-database $\bar{D}$ such that the following are satisfied:

1. $\bar{D} \models Q(\bar{a})$ and $\bar{D}, \bar{a} \rightarrow D, \bar{a}$;
2. if chase($\bar{D}, \Sigma$) $\models^{\sigma} q_c(\bar{a})$, for $q_c$ a contraction of $q$, then there is no S-database $D'$ and contraction $q'_c$ of $q$ where $D', \bar{a} \rightarrow D, \bar{a}$, chase($D', \Sigma$) $\models^{\sigma} q'_c(\bar{a})$, and $q_c \neq q'_c$ is a contraction of $q_c$.

Proof. We start with $\bar{D} = D$. Then, of course, Condition 2 is not guaranteed to be satisfied. We thus iteratively replace $D$ with more suitable databases, as follows. As long as there is an S-database $D'$ and contractions $q_c, q'_c$ of $q$ such that chase($\bar{D}, \Sigma$) $\models^{\sigma} q_c(\bar{a})$, $D', \bar{a} \rightarrow D, \bar{a}$, chase($D', \Sigma$) $\models^{\sigma} q'_c(\bar{a})$, and $q_c \neq q'_c$ is a contraction of $q'_c$, replace $D$ by $D'$. Informally, this iterative process terminates since we lose at least one contraction $q_c$ of $q$ with chase($\bar{D}, \Sigma$) $\models^{\sigma} q_c(\bar{a})$ in every step. We refer to [7] for details.

It is clear that the resulting $D$ satisfies Condition (2). Condition (1) is satisfied as well: we have chase($\bar{D}, \Sigma$) $\models^{\sigma} q_c(\bar{a})$ for some contraction $q_c$ of $q$, thus $\bar{D} \models Q(\bar{a})$.

Lemma D.5. Given an OMQ $Q = (S, \Sigma, q)$ from $(G \cap \text{FULL, UCQ})$ and an S-database $\bar{D}$, it is decidable whether Conditions 1 and 2 from Lemma D.4 hold.

Proof. As Condition 1 is clearly decidable, we concentrate on Condition 2. We start with a central observation. Let $q'_c$ be a contraction of $q$ and let $q''_c$ be the UCQ that consists of all contractions of $q'_c$ that are distinct from $q''_c$. Then checking whether chase($D', \Sigma$) $\models^{\sigma} q'_c(\bar{a})$, for some S-database $D'$ with $D', \bar{a} \rightarrow D, \bar{a}$, is equivalent to checking that the OMQs ($S, \Sigma, q''_c$) and ($S, \Sigma, q'_c$) are not equivalent on S-databases $D'$ with $D', \bar{a} \rightarrow D, \bar{a}$. As a consequence, we decide Condition 2 by checking that there are no contractions $q_c, q'_c$ of $q$ such that $q_c$ is a contraction of $q'_c$, $q_c \neq q'_c$, chase($\bar{D}, \Sigma$) $\models q_c(\bar{a})$ and ($S, \Sigma, q''_c$) are not equivalent on S-databases $D'$ with $D', \bar{a} \rightarrow D, \bar{a}$. Note that in the statement ‘chase($\bar{D}, \Sigma$) $\models q_c'(\bar{a})$, we have dropped the $\models^{\sigma}$ that is present in Condition 2. This is not harmful since if chase($\bar{D}, \Sigma$) $\models q_c(\bar{a})$, but chase($\bar{D}, \Sigma$) $\not\models^{\sigma} q_c(\bar{a})$, we can simply replace $q_c$ by a contraction of itself that satisfies this stronger condition.

The width of an OMQ $Q = (S, \Sigma, q)$ from $(G \cap \text{FULL, UCQ})$, denoted $w(Q)$, is the maximum number of variables in the body of a rule in $\Sigma$ or in a CQ in $q$.

To establish decidability of Condition 2, it suffices to show that given OMQs $Q_1$ and $Q_2$ from $(G \cap \text{FULL, UCQ})$ over schema $S$, an S-database $D$, and a tuple of constants $\bar{a}$, it is decidable whether there exists an S-database $D$ such that $D \models Q_1(\bar{a})$, $D \models Q_2(\bar{a})$, and $D, \bar{a} \rightarrow D, \bar{a}$. We note that this is the case if there exists a database with these properties of is treewidth $w(Q_1)$.

For an OMQ $Q = (S, \Sigma, q)$ from $(G \cap \text{FULL, UCQ})$ we construct a Guarded Second Order (GSO) [25] formula $\phi_Q$ such that for every S-database $D$ and tuple of constants $\bar{a}$, $D \models Q(\bar{a})$ if and only if $\phi_Q(\bar{a})$. Let $J_1, \ldots, J_n$ be the relation symbols in $\Sigma$ and reserve fresh relation symbols $J'_1, \ldots, J'_n$ and $J''_1, \ldots, J''_n$ of the same arity. Further let:

- $\phi'_c$ (resp. $\phi''_c$) be the conjunction of all TGDs in $\Sigma$, viewed as universally quantified FO formulas, with each relation symbol $J_i$ replaced by $J'_i$ (resp. $J''_i$);
- $\phi$ be a formula which says that if $J_i$ is in $\Sigma$, then the extension of $J_i$ is included in that of $J'_i$;
- $\psi$ be a formula which says that the extension of each $J''_i$ is included in the extension of $J'_i$, with at least one inclusion being strict.

Then $\phi_Q$ is the following formula:

$$\exists J'_1, \ldots, J'_n (\phi \land \psi \land \forall J''_1, \ldots, J''_n (\psi \rightarrow \neg \phi''_c)).$$

Note that the extension of $J'_1, \ldots, J'_n$ represents the chase of the input database with $\Sigma$.

We also take a formula $\phi_D$ that evaluates to true exactly on those S-databases $D$ such that $D \rightarrow \bar{D}, \bar{a}$. Then, we have to decide whether there exists an S-database $D$ of treewidth at most $w(Q_1)$ such that $D \models \phi_Q \land \neg \phi_D \land \psi$. Note that on sparse structures, GSO has the same expressive power as MSO [13]. As databases of bounded treewidth are sparse [19], it follows that GSO over such databases is decidable.

For $k, \ell \geq 1$, the $k \times \ell$-grid is the (undirected) graph with vertex set $\{(i, j) \mid 1 \leq i \leq k \text{ and } 1 \leq j \leq \ell \}$ and an edge between $(i, j)$ and $(i', j')$ iff $\lvert i - i' \rvert + \lvert j - j' \rvert = 1$. A graph $H$ is a minor of a graph $G$ if $H$ is isomorphic to a graph that can be obtained from a subgraph of $G$ by contracting edges. Equivalently, $H$ is a minor of $G$ if there is a minor map from $H = (V_H, E_H)$ to $G = (V_E, E_G)$, that is, a mapping $\mu : V_H \rightarrow V_G$ with the following properties for all $v, w \in V_H$:

- $\mu(v)$ is non-empty and connected in $G$;
- $\mu(v)$ and $\mu(w)$ are disjoint whenever $v \neq w$;
- if $\{v, w\} \in E_H$, then there are $v', w' \in \mu(v)$ and $w'' \in \mu(w)$ such that $\{v', w''\} \in E_G$.

We say that $\mu$ is onto if $\bigcup_{v \in V_H} \mu(v) = V_G$. For a database $D$, we say that $a \in \text{dom}(D)$ is isolated in $D$ if there is only a single atom $R(\bar{a}) \in D$ with $a \in \bar{a}$. When $k$ is understood from the context, we use $K$ to denote $\bigcup_{i=1}^{k} J_i$.

Note that Point (3) of the following theorem can be viewed as a form of ontoeness requirement for the homomorphism $h_0$ from Point (1): for the fact $R(a_1, \ldots, a_n) \in D$ mentioned in that condition, it is possible to find a fact $R(c_1, \ldots, c_n) \in D_G$ that maps to it.

Theorem 6.1 (Grohe). Given an undirected graph $G$, a $k \geq 1$, a connected S-database $D$, and a set $A \subseteq \text{dom}(D)$ such that the restriction $G^D_{\text{G}}$ of the Gaifman graph of $D$ to vertices $A$ contains the $k \times k$-grid as a minor, one can construct in time $f(k)\cdot \text{poly}(\lvert G\rvert, \lvert D\rvert)$ an S-database $D_G$ with $\text{dom}(D_G) \setminus A \subseteq \text{dom}(D_G)$ such that:

1. there is a surjective homomorphism $h_0$ from $D_G$ to $D$ that is the identity on $\text{dom}(D) \setminus A$;
2. $G$ contains a $k$-clique iff there is a homomorphism $h$ from $D$ to $D_G$ such that $h$ is the identity on $\text{dom}(D) \setminus A$ and $h_0(h(i))$ is the identity, and
3. if $R(a_1, \ldots, a_n) \in D$, $h_0(b_i) = a_i$ for each $i \in [n]$, and $\{b_1 | a_1 \text{ non-isolated in } D\}$ is a clique in the Gaifman graph of $D_G$.

\footnote{We use the variant of GSO whose the syntax is identical to that of Second Order Logic, but the second-order quantifiers range only over guarded relations, see [25].}
then $D_G$ contains an atom $R(c_1, \ldots, c_n)$ where $c_i = b_i$ if $a_i$ is non-isolated in $D$ and $h_0(c_i) = a_i$ for $i \in [n]$.

**Proof.** The construction of $D_G$ is a slight extension of the original construction from [26]. The extension is necessary because we have to treat the constants from $\text{dom}(D) \setminus A$ in a special way whereas in [26] there is no set $A$ or, in other words, $A = \text{dom}(D)$. Let $\mu$ be a minor map from the $k \times K$-grid to $G_D^{[A]}$. Since $D$ is connected, we can assume w.l.o.g. that $\mu$ is onto. To attain Point (3) of Theorem 6.1, we actually have to choose $\mu$ a bit more carefully. We first proceed with the construction and argue that Points (1) and (2) are satisfied, and then give details on Point (3). Fix a bijection $\rho$ between $[K]$ and the set of unordered pairs over $[K]$. For $p \in [K]$, let $i \in p$ be shorthand for $i \in \rho(p)$. Let $G_D^{[A]} = (V,E)$.

The domain of $D_G$ is

$$(\text{dom}(D) \setminus A) \cup \{(v,e,i,p,a) \in V \times V \times [K] \times [K] \times A \mid (v \in e \Leftrightarrow i \in p), a \in \mu(i,p)\}.$$ 

Before we say what the atoms of $D_G$ are, let us define the homomorphism $h_0 : \text{dom}(D_G) \rightarrow \text{dom}(D)$ from Point (1) by taking the identity on $\text{dom}(D) \setminus A$ and the projection to the last component for all other constants from $\text{dom}(D_G)$. We extend $h_0$ to tuples over $\text{dom}(D_G)$ in the expected way.

We now define $D_G$ to contain every atom $R(\bar{b})$, $\bar{b}$ a tuple over $\text{dom}(D_G)$, such that $R(h_0(\bar{b})) \in D$ and for all constants $b, b'$ in $\bar{a}$ that are of the form $b = (v,e,i,p,a) \land b' = (v',e',i',p',a')$,

(C1) $i = i'$ implies $v = v'$ and (C2) $p = p'$ implies $e = e'$.

This finishes the construction of $D_G$.

Point (1) follows directly from the construction. The proof of Point (2) is a slight variation of the proofs of Lemma 4.3 (for the $\Rightarrow$ direction) and Lemma 4.4 (for the $\Leftrightarrow$ direction) from [26]. In fact, the homomorphism $h$ constructed in the proof of Lemma 4.3 simply needs to be extended to be the identity on $\text{dom}(D) \setminus A$. In Lemma 4.4, the assumption that $D$ is a core has been replaced here with the condition that $h_0(b(\cdots))$ is the identity. This, however, is exactly what is derived from the assumption that $D$ is a core in the original proof. No further changes to the proof are required despite the presence of the elements in $\text{dom}(D) \setminus A$.

For Point (3), assume that $R(a_1, \ldots, a_n) \in D$, $h_0(b_j) = a_j$ for $1 \leq j \leq n$, and $S = \{b_j \mid a_j \text{ non-isolated in } D\}$ is a clique in the Gaifman graph of $D_G$. As $D$ is connected, $S$ cannot be empty. We can assume w.l.o.g. that the minor map $\mu$ from the $k \times K$-grid onto $G_D^{[A]}$ is such that one of the following holds:

(1) there is an $a_j$ that is non-isolated in $D$ such that $a_j \in A$ and if $i,p$ are such that $a_j \in \mu(i,p)$, then all $a_j \in A$ that are isolated in $D$ are also in $\mu(i,p)$; in this case, $b_j$ must take the form $(v,e,i,p,a_j)$;

(2) there is no $a_j$ that is non-isolated in $D$ such that $a_j \in A$ and there are $i,p$ such that all $a_j \in A$ that are isolated in $D$ are in $\mu(i,p)$; in this case, choose $v, e$ arbitrarily such that $v \in e$ iff $i \in p$.

Define $c_1, \ldots, c_n$ by setting

$$c_j = \begin{cases} (v,e,i,p,a_j) & \text{if } a_j \in A \text{ is isolated in } D \\ b_j & \text{otherwise.} \end{cases}$$

It can be verified that $(v,e,i,p,a_j)$ is a constant in $D_G$ if $a_j \in A$ is isolated in $D$, due to our assumptions (1) and (2). Moreover, $h_0(c_j) = h_0(b_j)$ for $1 \leq i \leq n$ and Conditions (C1) and (C2) are satisfied for the fact $R(c_1, \ldots, c_n)$, which is thus in $D_G$. Finally, $h_0(c_i) = a_i$ for $1 \leq i \leq n$, by definition of $h_0$.}

**D.2 The Reduction**

Recall that our goal is to establish the following lower bound:

**Theorem 6.4.** Fix $r \geq 1$. Let $\mathcal{Q}$ be a recursively enumerable class of OMQs from $(\mathcal{G}, \text{UCQ})$ over a schema of arity $r$, and, for each $k \geq 1$, $\mathcal{Q} \not\subseteq (\mathcal{G}, \text{UCQ})^k_\text{c}$. Then, $\text{p-OMQ-Evaluation}(\mathcal{G})$ is W[1]-hard.

The proof is by fpt-reduction from $\text{p-Clique}$, a W[1]-hard problem. Assume that $\mathcal{G}$ is an undirected graph and $k \geq 1$ a clique size, given as an input to the reduction. By Robertson and Seymour’s Excluded Grid Theorem, there is an $\ell$ such that every graph of treewidth exceeding $\ell$ contains the $k \times K$-grid as a minor [34]. We may assume, w.l.o.g., that $\ell$ exceeds the fixed arity $r$. By our assumption on $\mathcal{O}$, we find an OMQ $Q = (\mathcal{S}, \Sigma, q)$ from $\mathcal{O}$ that is not in $(\mathcal{G}, \text{UCQ})^\ell_\text{c}$. Since the choice of $\mathcal{Q}$ is independent of $\mathcal{G}$ and since it is decidable whether an OMQ from $(\mathcal{G}, \text{UCQ})$ is in $(\mathcal{G}, \text{UCQ})^\ell_\text{c}$ by Theorem 5.1, we can enumerate the OMQs from $\mathcal{O}$ until we find $Q$.

By Theorem 6.1, we can assume, w.l.o.g., that $Q = (\mathcal{S}, \Sigma, q)$ is from $(\mathcal{G} \cap \text{FULL, UCQ})$. It might of course be that the OMQ from $(\mathcal{G} \cap \text{FULL, UCQ})$ is not in $\mathcal{O}$, but we will see that this is not a problem. Let $q = q_1 \cup \cdots \cup q_n$ and let $Q_i = (\mathcal{S}, \Sigma, q_i)$ for $1 \leq i \leq n$. We can assume w.l.o.g. that $Q_i \not\subseteq Q_j$ for all $i \neq j$ because this is not the case, then we can drop $q_j$ from $Q$ and the resulting OMQ is equivalent to $Q$. Also, we exhaustively replace CQs $q_i$ in $Q$ with $(q_i)^+ \text{ as defined in the context of UCQ}_r$-approximations (see Definition C.6), whenever the resulting OMQ is equivalent to $Q^+$. Since $Q$ is not in $(\mathcal{G}, \text{UCQ})^\ell_\text{c}$, the final OMQ must contain a $q_w$ that has not been replaced by $(q_w)^+$ and, in particular, we must have $Q \not\subseteq Q'$ where $Q'$ is obtained from $Q$ by replacing $q_w$ with $(q_w)^+$. We thus find an $S$-database $D_0$ and tuple of constants $\bar{a}_0$ such that $D_0 \models Q_w(\bar{a}_0)$, $D_0 \not\models (Q_w)^+(\bar{a}_0)$, and $D_0 \not\models Q(\bar{a}_0)$ for all $i \neq w$. By duplicating constants, we can achieve that all constants in $\bar{a}_0$ are distinct. We also assume that Condition 2 of Lemma 6.4 is satisfied for $Q = Q_w$, $q = q_w$, $D = D_0$, and $\bar{a} = \bar{a}_0$. If it is not, we can apply that lemma and replace $D_0$ with the resulting $D_0$. By Condition (1), $D_0 \models Q_w(\bar{a}_0)$, $D_0 \not\models (Q_w)^+(\bar{a}_0)$, and $D_0 \not\models Q(\bar{a}_0)$ for all $i \neq w$.

Since the properties of $D_0$ are independent of $\mathcal{G}$ and due to Lemma 5.5, we can find $D_0$ by enumeration. However, $D_0$ is still not as required and needs to be manipulated further to make it suitable for the reduction. Before we can carry out the actual manipulation, we need some preliminaries.

Recall that for a guarded set $\bar{a} \subseteq \text{dom}(D_0)$, $D_0^{\bar{a}}$ denotes the guarded unraveling of $D_0$ starting at $\bar{a}$. An atomic query (AQ) takes the form $R(\bar{x})$, i.e., it has a single atom and no quantified variables.

**Lemma 6.7.** $D_0 \models (\mathcal{S}, \Sigma, p)(\bar{a}^\prime)$ iff $D_0^{\bar{a}} \models (\mathcal{S}, \Sigma, p)(\bar{a}^\prime)$ if $p$ is an AQ and all constants in $\bar{a}^\prime$ are also in $\bar{a}$.

Of course, $D_0^{\bar{a}}$ can be infinite. By compactness, however, there is a finite $D_0^{\bar{a}} \subseteq D_0^{\bar{a}}$ such that Lemma 6.7 is satisfied for all (finitely

---

1. We rely on the key property of UCQ$_r$-approximations given by item (2) of Lemma C.7.
We next observe some basic properties of $D_0$. Then domain $\bar{a}$ is a diversification of $R$ such that $x \in \text{dom}(D)$ is old if it already occurs in $D_0$ and fresh otherwise. For a fresh constant $b$, we use $b^1$ to denote the constant in $D_0$ that it was introduced for. For old constants $a$, $a^1 \equiv a$. This extends to tuples of constants from $D$ in a componentwise fashion. Note that all fresh constants in $D$ are isolated in $D$ and that all constants from $a_0$ still occur in $D$ as old constants. If $D_1$ and $D_2$ are diversifications of $D_0$, we write $D_1 \subseteq D_2$ if there is a homomorphism from $D_1$ to $D_2$ that is the identity on all old constants.

Example D.8. Assume that $D = \{R(a, b, c), S(a, b, d)\}$ and $\bar{a}_0 = ()$. Then $D_1 = \{\bar{R}(a, b', c'), S(a, b, d)\}$ and $D_2 = \{\bar{R}(a, b, c'), S(a, b, d)\}$ are diversifications of $D$ and $D_1 \subseteq D_2$.

If $D$ is a diversification of $D_0$, we use $D^*$ to denote the database obtained from $D$ by adding each database $D_1$ such that for some atom $R(a)$ in $D_0$, we obtain $D_1$ from $D_a$ by renaming the constants in $\bar{a}$ to those in $\bar{a}^1$. A constant $a \in \text{dom}(D^*)$ is old if it is an old constant from dom($D$), and fresh otherwise.

For what follows, we choose a $\subseteq$-minimal diversification $D_1$ of $D_0$ such that $D_1^* \models Q_w(\bar{a}_0)$.

Example D.9. Assume that $Q_w$ is a Boolean CQ that takes the form of an $n \times m$-grid using the binary relations $X$ and $Y$. Let $S = \{X, Y\}$ with both relations ternary and let $\Sigma = \{X(x, y, z) \rightarrow X(x, y), Y(x, y, z) \rightarrow Y(x, y)\}$. Further let $D_0$ be the database

$$\{X'(a_{i,j}, a_{i,j+1}, b) \mid 1 \leq i \leq m \text{ and } 1 \leq j < n\} \cup \{Y'(a_{i,j}, a_{i+1,j}, b) \mid 1 \leq i < m \text{ and } 1 \leq j \leq n\}$$

Then the following database is a $\subseteq$-minimal diversification of $D_0$ such that $D_1^* \models Q_w(\bar{a}_0)$:

$$\{X'(a_{i,j}, a_{i,j+1}, b_{i,j}) \mid 1 \leq i \leq m \text{ and } 1 \leq j < n\} \cup \{Y'(a_{i,j}, a_{i+1,j}, b'_{i,j}) \mid 1 \leq i < m \text{ and } 1 \leq j \leq n\}.$$ 

We next observe some basic properties of $D_1$.

**Lemma D.10.**

1. $D_1^* \models Q_w(\bar{a}_0)$ and $D_1^* \not\models Q(\bar{a}_0)$ for all $i \neq w$.
2. $D_1$ has treewidth exceeding $\ell$ up to $\bar{a}_0$.

**Proof.** For Point 1, we have $D_1^* \models Q_w(\bar{a}_0)$ by choice of $D_1^*$. By construction of $D_1^*$, we have $D_1^*, \bar{a}_0 \rightarrow D_0, \bar{a}_0$ and thus $D_1^* \not\models Q(\bar{a}_0)$ for all $i \neq w$. We proceed with Point 2, noting that we can argue in the same way to show that $D_1^* \not\models Q(\bar{a}_0)$ for all $i \neq w$. Thus Point 2 of Lemma C.7 implies that the treewidth of $(D_1^*)^+$ up to $\bar{a}_0$ exceeds $\ell$. Note that the treewidth of $D_1^* \setminus D_1$ is bounded by the fixed arity $r$. As we assume $\ell > r$, $D_1$ has treewidth exceeding $\ell$ up to $\bar{a}_0$.

Now for Point 3, that is, $D_1^*$ satisfies Condition 2 of Lemma D.4. Assume to the contrary that there are an $S$-database $D'$ and contractions $q_c, q'_c$ of $q_w$ such that $\text{chase}(D_1^*, \Sigma) \models\equiv q_c(\bar{a}_0)$, $D', \bar{a}_0 \rightarrow D_1^*, \bar{a}_0$, $\text{chase}(D', \Sigma) \models\equiv q'_c(\bar{a}_0)$, and $q_c \neq q'_c$ is a contraction of $q'_c$.

We use a case distinction:

- $\text{chase}(D_0, \Sigma) \models\equiv q_c(\bar{a}_0)$.
  From $D', \bar{a}_0 \rightarrow D_1^*, \bar{a}_0$ and $D_1^*, \bar{a}_0 \rightarrow D_0, \bar{a}_0$, we obtain $D', \bar{a}_0 \rightarrow D_0, \bar{a}_0$. This together with $\text{chase}(D_0, \Sigma) \equiv\equiv q_c(\bar{a}_0)$ and $\text{chase}(D', \Sigma) \models\equiv q'_c(\bar{a}_0)$ yields a contradiction to $D_0$ satisfying Condition 2 of Lemma D.4.

- $\text{chase}(D_0, \Sigma) \equiv\equiv q'_c(\bar{a}_0)$.
  From $\text{chase}(D_1^*, \Sigma) \equiv\equiv q_c(\bar{a}_0)$ and $D_1^*, \bar{a}_0 \rightarrow D_0, \bar{a}_0$, we obtain $\text{chase}(D_0, \Sigma) \equiv\equiv q_c(\bar{a}_0)$. By Lemma D.3, there is a contraction $q''_c$ of $q_c$ with $\text{chase}(D_0, \Sigma) \equiv\equiv q''_c(\bar{a}_0)$. But then we must have $q''_c = q'_c$, as otherwise we obtain a contradiction to $D_0$ satisfying Condition 2 of Lemma D.4 (instantiated with $D = \bar{D} = D_0$). Contradiction.

This completes the proof of Lemma D.10. 

The next lemma states a more intricate property of $D_1$ that is crucial for the remaining proof:

**Lemma D.11.** If $h$ is a homomorphism from $q_w(\bar{x}_0)$ to $\text{chase}(D_1, \Sigma)$ with $h(\bar{x}_0) = \bar{a}_0$, $R(\bar{a}) \in D_1$, and $a_1, a_2$ are old constants in $\bar{a}$, then there is a path $x_1, \ldots, x_m$ in the Gaifman graph $\mathcal{G}^{D[Q_w]}$ of $D[q_w]$ such that $h(x_1) = a_1, h(x_m) = a_2$, and $h(x_2), \ldots, h(x_{m-1})$ are in dom($D_a$).

**Proof.** Assume to the contrary that there is a homomorphism $h$ from $q_w(\bar{x}_0)$ to $\text{chase}(D_1^*, \Sigma)$ with $h(\bar{x}_0) = \bar{a}_0$, an $R(\bar{a}) \in D_1$, and old constants $a_1, a_2$ in $\bar{a}$ such that there is no path in the Gaifman graph $G^{D[Q_w]}$ of $D[q_w]$ as required by Lemma D.11. We argue that we can then find an $S$-database $D_2$ with $D_2 \subseteq D_1$ and $D_2^* \models Q_w(\bar{a}_0)$, in contradiction to the choice of $D_1$.

Let $D_2$ be obtained from $D_1$ by replacing $R(\bar{a})$ with the atoms $R(\bar{a}_1), R(\bar{a}_2)$ where $\bar{a}_1$ is obtained from $\bar{a}$ by replacing $a_1$ with a fresh constant $b_1$ and each fresh constant with a new fresh constant. Clearly, $D_2$ is a diversification of $D_0$ and $D_2 \subseteq D_1$. We convert $h$ into a homomorphism $h'$ from $q_w$ to $\text{chase}(D_1^*, \Sigma)$ such that $h'(\bar{x}) = \bar{a}_0$, which contradicts the choice of $D_1$.

As a preparation, let us discuss the construction of $D_2$ and $D_2^*$. During the construction of $D_2^*$, we have attached a copy $\mathcal{D}_a$ of $D_4$ at $\bar{a}$. In the construction of $D_2^*$, we have attached copies $\mathcal{D}_a$, and $\mathcal{D}_b$, of the same $\mathcal{D}_a$ at $\bar{a}_1$ and $\bar{a}_2$. The constants in dom($D_2$) \ dom($D_1$) are the only constants in $D_2^*$ that may not be in $D_2^*$. For $i \in \{1, 2\}$, fix an isomorphism $i_i$ between $D_a$ and $D_{a_i}$. Note that $i_i$ is the identity on $a_2 - a_1$ and in fact on all old constants in $\bar{a}$ except $a_1$.

Now define $h'(x)$ as follows, for every variable $x$ in $q$:

1. If $h(x) \in \text{dom}(D_1^*) \setminus \text{dom}(D_a)$, then $h'(x) = h(x)$
2. If $h(x) \in \text{dom}(D_a)$:

   a special case is that there is an atom $S(\bar{x})$ in $q_w$ such that $a_1$ and $a_2$ are both in $h(\bar{x})$. Also note that it follows that all old constants are in the range of $h$.

   This is no longer true of $\Sigma$ is formulated in FO and this is in fact the reason why our proof does not extended from FO to PFO.
Claim 1. Let $b$ be a guarded set in $D_i$, $j \in \{1,2\}$. Then $R(\bar{b}) \in \text{chase}(D^*_i, \Sigma)$ if and only if $R(\bar{b}^+ |) \in \text{chase}(D^*_i, \Sigma)$ for all facts $R(\bar{b})$.

Here is the second claim.

Claim 2. Let $b$ be a guarded set in $D_a$. Then $R(\bar{b}) \in \text{chase}(D^*_i, \Sigma)$ implies $R(\bar{i}(\bar{b})) \in \text{chase}(D^*_i, \Sigma)$ for all facts $R(\bar{b})$ and $i \in \{1,2\}$.

Assume to the contrary that $R(\bar{i}(\bar{b})) \notin \text{chase}(D^*_i, \Sigma)$. Set

$$D = \text{chase}(D^*_i, \Sigma)|_{\text{dom}(D^*_i) \cap \text{dom}(D_a)} \cup \{R(\bar{i}(\bar{c})) | R(\bar{c}) \in \text{chase}(D^*_i, \Sigma) \}_{\text{dom}(D_a)}(\bar{c}) \}.$$

Using the fact that $\Sigma$ is guarded, it can be shown that $D$ is a model of $\Sigma$ and of $D^*_i$. Moreover, $R(\bar{b}) \notin D$ and, consequently, $R(\bar{b}) \notin \text{chase}(D^*_i, \Sigma)$, which finishes the proof of the claim.

The proof of the third claim is identical to that of Claim 2.

Claim 3. Let $c$ be a guarded set in some tree $D_2$, $i \in \{1,2\}$, $\bar{c} \neq \bar{a}$. Then $R(\bar{c}) \in \text{chase}(D^*_i, \Sigma)$ implies $R(\bar{i}(\bar{c})) \in \text{chase}(D^*_i, \Sigma)$ for all atoms $R(\bar{c})$.

Now for the proof that $h'$ is a homomorphism from $q_w$ to $\text{chase}(D^*_i, \Sigma)$. Let $S(\bar{x}) \in q_w$.

First assume that $h(x)$ contains only constants from $\text{dom}(D_i)$. It follows from (two applications of) Claim 1 that $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$ implies $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$.

Next assume that $h(x)$ contains a constant from a tree $\text{dom}(D_k)$, $\bar{c} \neq \bar{a}$, that is not in $\text{dom}(D_1)$. By construction of $D^*_i$ and by definition of the chase, this implies that $h(\bar{x}) \subseteq \text{dom}(D_2)$. We can thus apply Claim 3 to infer that $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$ implies $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$.

Finally assume that $h(\bar{x})$ contains a constant from $\text{dom}(D_2) \setminus \text{dom}(D_1)$. Then $h(x) \subseteq \text{dom}(D_a)$. Thus $h(x) = i(h(x))$ for some $i \in \{1,2\}$. We can thus apply Claim 2 to infer that $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$ implies $S(h(\bar{x})) \in \text{chase}(D^*_i, \Sigma)$.

Clearly, neither $D_1$ nor $D^*_i$ are guaranteed to be connected. In particular, $D_0$ might be disconnected and even if it is connected the diversification might replace all constants in an atom with fresh constants and in this way make $D_1$ disconnected. By Point 2 of Lemma D.10, however, the restriction of $G_{D_1}$ to $\text{dom}(D_1) \setminus \bar{a}_0$ has a maximal connected component that is of treewidth exceeding $\ell$. Let $A$ be the set of constants in that component. Moreover, let $D$ be the maximal connected component of $D_1$ that contains all constants from $A$ (it might additionally contain constants from $\bar{a}_0$) and let $D^*_i$ denote the part of $D^*_i$ that is disconnected from $D \subseteq D^*_i$. By choice of $D, A$, and $\ell$, the restriction of $G^*_i$ to $A$ contains the $k \times k$-grid as a minor. We can thus apply Theorem 6.1 to $G, k$, and $D$, and $A$, obtaining an $\mathcal{S}$-database $G_D$ and a homomorphism $h_0$ from $G_D$ to $D$ such that Points 1 to 3 of that theorem are satisfied. Let

1. $D^*_i$ be obtained by starting with $D_G$ and then disjointly adding, for each guarded set $\bar{a}$ in $D$ such that the restriction of $h_0$ to $\bar{a}$ is injective, a copy of the database $D_{\bar{a}}(\bar{a})$ that was attached to $R(h_0(\bar{a})) \in D_1$ during the construction of $D^*_i$, identifying the root of this copy with $\bar{a}$.

2. $D^*_i$ be obtained by disjointly adding $D^*_i$.

The fpt-reduction of $p$-Clique then consists in computing $Q$ and $D^*_i$ from $G$ and $k \geq 1$. We mean here the original $Q$ that is guaranteed to be in $\mathcal{O}$ and where $\Sigma$ is from $\mathcal{G} \cap \text{FULL}$.

D.3 Correctness of the Reduction

We show in the subsequent lemma that $D^*_i \models Q(\bar{a})$ if and only if $G$ has a $k$-clique. Clearly, we can work with the version of $Q$ here in which $\Sigma$ is from $\mathcal{G} \cap \text{FULL}$ since it is equivalent to the original $Q$.

Lemma D.12. $G$ has a $k$-clique iff $D^*_i \models Q(\bar{a})$.

Proof. The 'only if' direction is easy. If $G$ has a $k$-clique, then by Point 2 of Theorem 6.1 and since $A$ contains none of the constants from $\bar{a}_0$, there is a homomorphism $h$ from $D$ to $D_G$ such that $h(\bar{a}_0) = \bar{a}_0$ and $h_0(h(\bar{c}))$ is the identity. It is straightforward to extend $h$ to a homomorphism from $D^*_i$ to $D^*_i \models D^*_i \models D^*_i \models D^*_i$; in particular, if $R(\bar{a})$ is an atom in $D$, then the restriction of $h_0$ to $\bar{a}$ must be injective since $h_0(h(\bar{c}))$ is the identity and thus a copy of the subtree that was attached to $h(\bar{a})$ in the construction of $D^*_i$ from $D_1$ was attached to $h(\bar{a})$ in the construction of $D^*_i$ from $D_G$. Consequently, $D^*_i \models Q(\bar{a})$ implies $D^*_i \models Q(\bar{a})$.

For the 'if' direction, assume that $D^*_i \models Q(\bar{a})$. The surjective homomorphism $h_0$ from $D_G$ to $D$ given by Theorem 6.1 can be extended to a homomorphism from $D^*_i \models D^*_i \models D^*_i \models D^*_i$ and from $\text{chase}(D^*_i, \Sigma)$ to $\text{chase}(D^*_i, \Sigma)$. For brevity, we denote this extension also with $h_0$. For later use, we observe that

1. $b \in \text{dom}(D_G)$ iff $h_0(b) \in \text{dom}(D)$;
2. $\bar{b}_1, \bar{b}_2$ are in the same tree of $D^*_i$ iff $h_0(b_1), h_0(b_2)$ are in the same tree of $D^*_i$.

So $h_0$ witnesses $D^*_i \models D^*_i$. Thus $D^*_i \not\models Q(\bar{a})$ for all $i \neq w$ implies $D^*_i \not\models Q(\bar{a})$ for all $i \neq w$. It follows that $D^*_i \models Q(\bar{a})$ and thus $\text{chase}(D^*_i, \Sigma) \models Q(\bar{a})$.

By Lemma D.3, we find a contraction $q_\ell$ of $q_w$ such that $\text{chase}(D^*_i, \Sigma) \models q_\ell(\bar{a})$. As witnessed by $h_0$, $\text{chase}(D^*_i, \Sigma), \bar{a}_0 \models \text{chase}(D^*_i, \Sigma)$, $\bar{a}_0$. Consequently $\text{chase}(D^*_i, \Sigma) \models q_\ell(\bar{a})$. Thus we find a contraction $q_\ell$ of $q_w$ such that $\text{chase}(D^*_i, \Sigma) \models q_\ell(\bar{a})$. We must have $q_\ell = q_\ell$ since $D^*_i$ satisfies Condition 2 of Lemma D.4, via Lemma D.10.

\footnote{The injectivity requirement implies that we can add these copies as they are, without duplicating any constants in them.}
Let $h$ be a homomorphism from $q_c$ to chase($D^+_G$, $\Sigma$). The composition $g = h_0(h(\cdot))$ is a homomorphism from $q_c$ to chase($D^+_G$, $\Sigma$). Since chase($D^+_G$, $\Sigma$) |= $\alpha$, $q_c$, this homomorphism must be injective. Consequently, also $h_0$ is injective.

Since $h_0$ is injective, we can construct a function $h_0^-$ as follows:

1. start with the restriction of $h_0$ to the range of $h$;
2. take the inverse;
3. restrict to old constants in $D$.

It follows from Lemma D.11 that the range of $h_0$ includes all old constants in $D$, and thus so does the domain of $h_0^-$. In particular, it thus also contains all constants in $\bar{a}_0$. Trivially, $h_0(h_0^-(\cdot))$ is the identity and $h_0^-$ is the identity on all constants from $\text{dom}(D) \setminus A \subseteq \bar{a}_0$ (since $h_0$ is the identity on such constants).

We are going to show that a certain extension of $h_0^-$ is a homomorphism from $D$ to $D_G$ that is the identity on $\text{dom}(D) \setminus A$, and thus Point 2 of Theorem 6.1 yields that $G$ contains a $k$-clique.

Let $R(\bar{a}) \in D$. By definition of $g$ and since the range of $g$ contains all old constants in $D$, we find for each old $a \in \bar{a}$ a variable $x_a \in \bar{x}$ and a $b_a \in \text{dom}(D^+_G)$ such that $h(x_a) = b_a$ and $h_0(b_a) = a$. Let $\Gamma$ be the set of all $b_a$. By Lemma D.11, for any two old constants $a_1, a_2$ in $\bar{a}$, there is a path $x_1, \ldots, x_n$ in the Gaifman graph of $D(q_w)$ such that $g(x_1) = a_1, g(x_m) = a_2$, and $g(x_2), \ldots, g(x_{m-1})$ are in $\text{dom}(D^+_G)$. By Properties P1 and P2, it follows that for any two $b_1, b_2$ in $\Gamma$, there is a path $x_1, \ldots, x_n$ in the Gaifman graph of $D(q_w)$ such that $h(x_1) = b_1, h(x_m) = b_2$, and $h(x_2), \ldots, h(x_{m-1})$ are in the same tree of $D^+_G$. The fact that $h_0$ maps all elements of $\Gamma$ to constants in $D$ and Property P1 further imply $b_1, b_2 \in \text{dom}(D_G)$. By construction of $D^+_G$, there must thus be a fact in $D_G$ that contains both $b_1$ and $b_2$ (the one for which the tree was added that $h(x_1), \ldots, h(x_{m-1})$ are). It follows that $\Gamma$ forms a clique in the Gaifman graph of $D_G$.

Since (the non-extended) $h_0$ is surjective, for each fresh $a \in \bar{a}$ we find some $b_a \in \text{dom}(D_G)$ such that $h_0(b_a) = a$. We can now apply Point 3 of Theorem 6.1 to $R(\bar{a})$ and the $b_a$ to obtain an extension of $h_0^-$ to the fresh constants in $\bar{a}$ such that $R(h_0^-(\bar{a})) \in D_G$. Note that the set $\{b_a \mid a$ is non-isolated in $D\}$ can only contain old constants since all fresh constants in $D$ are isolated; it is thus a subset of $\Gamma$ and forms a clique in the Gaifman graph of $D_G$. The extension is such that if $a$ is fresh and covered by the extension, then $h_0^-(a) = c$ for some $c$ with $h_0(c) = a$. Thus, $h_0(h_0^-(\cdot))$ is still the identity and $h_0^-$ is still the identity on $\text{dom}(D) \setminus A$. Also note that this extension can be done independently for all $R(\bar{a}) \in D$ because all fresh constants in $D$ are isolated in $D$.

Finally, Point 2 of Theorem 6.1 yields that $G$ has a $k$-clique.

**Lemma E.1.** Consider two CQs $S = (\Sigma, q(\bar{x}))$ and $S' = (\Sigma, q'(\bar{x}'))$ from $(G, \text{UCQ})$. Then, $S \equiv S'$ iff $S \equiv_{\text{unr}} S'$.

**Proof.** The $(\Rightarrow)$ direction holds trivially. For the $(\Rightarrow)$ direction, assume that $S \equiv_{\text{unr}} S'$. There are two cases: (i) $S \equiv_{\text{unr}} S'$, or (ii) $S' \equiv_{\text{unr}} S$. We consider only the first case; the second case is shown analogously. Since $S \equiv_{\text{unr}} S'$, we get that $\bar{x} \notin q'(\text{chase}(q, \Sigma))$. With $Q'$ being the OMQ $\text{omq}(S')$, we get that $\bar{x} \notin Q'(D[q])$. Since, by Lemma 6.6 and Theorem 6.7, guarded TGDs are finitely controllable, we conclude that there exists a finite model $M$ of $D[q]$ and $\Sigma$ such that $\bar{x} \notin q'(M)$. However, it holds trivially that $\bar{x} \notin q(M)$, and thus, $S \equiv S'$, which in turn implies that $S \equiv S'$, as needed.

Having the above lemma in place, we can now give the proof of Proposition 5.5. Assume that $S = (\Sigma, q)$ is over a schema $S$. Since, by Proposition 5.2, an OMQ from $(G, \text{UCQ})$ is $\text{UCQ}_k$-equivalent iff it is uniformly $\text{UCQ}_k$-equivalent, it suffices to show that, for each $k \geq 1$, the following are equivalent:

1. $\exists q' \in \text{UCQ}_k$ over $S$ such that $S \equiv (\Sigma, q')$.
2. $\exists q' \in \text{UCQ}_k$ over $S$ such that $\text{omq}(S) \equiv (S, \Sigma, q')$.

$(1) \Rightarrow (2)$. By hypothesis, there exists $q' \in \text{UCQ}_k$ over $S$ such that $S \equiv (\Sigma, q')$. It is easy to show that $\text{omq}(S) \equiv (S, \Sigma, q')$; for brevity, let $Q = \text{omq}(S)$ and $Q' = (S, \Sigma, q')$. Consider an arbitrary $S$-database $D$. Since, by construction, chase($D$, $\Sigma$) is a model of $\Sigma$, and, by Lemma E.1, $S \equiv_{\text{unr}} (\Sigma, q')$, we get that $q(\text{chase}(D, \Sigma)) \equiv q'(\text{chase}(D, \Sigma'))$. Therefore, $Q(D) = Q'(D)$, as needed.

$(2) \Rightarrow (1)$. By hypothesis, there exists $q' \in \text{UCQ}_k$ over $S$ such that $\text{omq}(S) \equiv (S, \Sigma, q')$; again, let $Q = \text{omq}(S)$ and $Q' = (S, \Sigma, q')$. Consider an arbitrary S-database $D$ such that $D$ is a model of $S$. Observe that $q(D) = Q(D)$ and $q'(D) = Q'(D)$. Since $Q \equiv Q'$, we get that $q(D) = q'(D)$, and the claim follows.

**F PROOF OF PROPOSITION 5.8**

Consider an OMQ $Q = (S, \Sigma, q(\bar{x}))$ from omq($\mathcal{O}$), an $S$-database $D$, and a tuple $\bar{c} \in \text{dom}(D)^{|\bar{x}|}$. We are going to construct an $S$-database $D^*$ such that $D^* \models S$, and $\bar{c} \in Q(D)$ iff $\bar{c} \in q(D^*)$, or, equivalently (due to Proposition 2.2), $\bar{c} \in \text{chase}(D, \Sigma)$ iff $\bar{c} \in q(D^*)$.

**F.1 Construction of the Database $D^*$**

Note that the definition of $D^*$ has been already given in the main body of the paper (Section 6.2). However, we repeat it here for the sake of readability. We first define $D^+$ as the database

\[ D \cup \{ R(\bar{a}) \in \text{chase}(D, \Sigma) \mid \bar{a} \subseteq \text{dom}(D) \}. \]

Let $A$ be the family of all maximal tuples $\bar{a}$ over $\text{dom}(D)$ that are guarded in $D^+$, i.e., there is an atom $R(\bar{b}) \in D^+$ such that $\bar{a} \subseteq \bar{b}$. Fix an arbitrary tuple $\bar{a} \in A$. Since, by Theorem 6.7, the class $\mathcal{O}$ is strongly finitely controllable, and also finite witnesses are realizable, we can compute an instance $M(D^+_{|\bar{a}|, \Sigma, n}) \in \text{fmods}(D^+_G, \Sigma)$, where $n$ is the number of variables in $q$, such that for each CQ $q'$ of arity $|\bar{a}|$ with at most $n$ variables, it holds that

\[ \bar{a} \in q'(M(D^+_{|\bar{a}|, \Sigma, n})) \implies \bar{a} \in q'(\text{chase}(D^+_{|\bar{a}|, \Sigma})). \]
Let \( S = (\Sigma, q(x)) \) be a CQS from \((\text{FO}_m, \text{UCQ})\) over a schema \( S \) of arity \( r \). Via a proof similar to that of Proposition 5.5, we get that:

- \( S \) is uniformly \( \text{UCQ}_k \)-equivalent iff \( \text{omq}(S) \) is.
- \( S \equiv S_k^k \iff \text{omq}(S) \equiv \text{omq}(S_k^k) \).

Let \( Q_S = \text{omq}(S) \) and \( Q^k_S = \text{omq}(S_k^k) \). Thus, for showing Proposition 5.11, it suffices to show that the following are equivalent:

1. \( Q_S \) is uniformly \( \text{UCQ}_k \)-equivalent.
2. \( Q_S \equiv Q^k_S \).

The proof of the next lemma is along the lines of the proof of Lemma C.7. The only difference is that we rely on the following fact: given a database \( D \) of treewidth at most \( k \) up to \( \bar{c} \), \( \text{chase}(D, \Sigma) \) has treewidth at most \( k \) up to \( \bar{c} \). The latter heavily exploits the bound \( m \) on the number of head atoms in the TGDs of \( \Sigma \).

**Lemma G.1.** For every \( \text{OMQ} \ Q' = (\Sigma, \varphi') \) from \((\text{FO}_m, \text{UCQ})\) such that \( Q' \subseteq Q_S \), it holds that \( Q' \equiv Q^k_S \).

Having Lemma G.1 in place, it is now easy to establish the above equivalence, i.e., \( Q_S \) is uniformly \( \text{UCQ}_k \)-equivalent iff \( Q_S \equiv Q^k_S \). Observe that the \( (=) \) direction holds trivially. It remains to show the \( (\Rightarrow) \) direction. By hypothesis, there exists an \( \text{OMQ} \ Q'' = (\Sigma, \varphi'' \) from \((\text{FO}_m, \text{UCQ}_k)\) such that \( Q'' \equiv Q^k_S \). By Lemma G.1, \( Q'' \subseteq Q^k_S \), which in turn implies that \( Q_S \subseteq Q^k_S \). The fact that \( Q^k_S \subseteq Q_S \) holds by construction, and the claim follows.

**H PROOF OF THEOREM 5.13**

We provide a complete proof of Theorem 5.13 here. Our proof does not use Grohe’s database as defined in [26], but a variant of it which we define below. We start by introducing several important notions.

Let \( G = (V, E) \) and \( H = (V', E') \) be simple graphs. A minor map from \( H \) to \( G \) is a mapping \( \mu : V' \rightarrow V^2 \) such that:

1. For every \( v' \in V', \mu(v') \) is nonempty and connected in \( G \).
2. For every \( v', u' \in V' \), with \( v' \neq u', \mu(v'), \mu(u') \) are disjoint.
3. The sets of the form \( \mu(v') \), for \( v' \in V' \), are pairwise disjoint.
4. For every edge \( \{v', u'\} \in E' \), there are nodes \( v \in \mu(v') \) and \( u \in \mu(u') \) such that \( \{v, u\} \in E \).

Such a minor map is said to be onto if, in addition, \( \bigcup_{v' \in V'} \mu(v') = V \).

Let \( G \) and \( H \) be simple graphs. It is easy to see that \( H \) is a minor of \( G \) iff there is a minor map \( \mu \) from \( H \) to \( G \). If, in addition, \( G \) is connected, then there exists an onto minor map \( \mu \) from \( H \) to \( G \).

**H.1 A Variation of Grohe’s Database**

Let \( G = (V, E) \) be an undirected graph, \( k \geq 1 \) be an integer, and \( K = (k) \). Suppose that \( D \) and \( D' \) are \( S \)-databases with \( D \subseteq D' \) such that there exists a set \( A \subseteq \text{dom}(D) \) for which there is a minor map \( \mu \) from the \((k \times K)\)-grid onto \( D|A \). We fix a bijection \( x \) that assigns to each \( 2 \)-element subset of \( [k] \) a value in \([K]\). We define a database \( D^* : = D^*(G, D, D', A, \mu) \) as follows.

The domain. Every element of \( D^* \) is either an element in \( \text{dom}(D') \) or a tuple of the form \( (v, e, i, p, z) \) where \( v \in V, e \in E, i \in [k], p \) is a two-element subset of \([k]\), and \( z \in A \). We note here that \( \text{dom}(D^*) \) does not necessarily contain all elements of the previous form since some of them may not appear in any of its atoms.

The facts. A labelled clique in \( G \) is any partial mapping \( \eta \) from \([k]\) to \( V \) such that, for every different \( i, j \) in the domain of \( \eta \), we have that \( \eta(i) \) and \( \eta(j) \) are adjacent in \( G \). We say that an element \( z \in \text{dom}(D') \) is covered by a labelled clique \( \eta \) if \( z \in A \), or there exist \( i, j, k \) in domain of \( \eta \) such that \( z \in \mu(i, \gamma((j, k))) \).

Then, for every fact \( R(\bar{x}) \in D' \) and every labelled clique \( \eta \)
covering every element mentioned in \(\tilde{\epsilon}\), it is the case that \(D^*\) contains the fact \(R(\tilde{z}_p)\) where \(\tilde{z}_p\) is obtained by replacing, in \(\tilde{\epsilon}\), every element \(z \in A\) with

\[(\eta(i), \eta(j), \eta(\ell)), i, [j, \ell], z),\]

where \(\mu(i, \chi(j, \ell)) = z\).

Additionally, we define the projection

\[h_0 : \text{dom}(D^*) \rightarrow \text{dom}(D^*)\]

such that

\[h_0(v, e, i, p, z) = z, \quad \text{if } z \in A, \text{ and }\]
\[h_0(y) = y, \quad \text{otherwise.}\]

Although we avoid to introduce Grohe’s original database [26], it is useful to compare it with ours. In particular, consider the case when \(D = D'\) and \(A = \text{dom}(D)\), and thus we have a database \(D' = D(G, D, D, \text{dom}(D), \mu)\). Grohe’s database \(D_{\text{Grohe}} = D_{\text{Grohe}}(G, D, \mu)\) then satisfies the following:

- \(\text{dom}(D_{\text{Grohe}})\) is contained in the set of tuples of the form \((v, e, i, \ell, p, z)\), for \(v \in V, e \in E, i \in [k], p\) is a two-element subset of \([k]\), and \(z \in \text{dom}(D)\). Therefore, \(h_0\) is well-defined on \(\text{dom}(D_{\text{Grohe}})\).
- \(\text{Also, } D^* \subseteq D_{\text{Grohe}}\).

In addition, the following crucial property holds for \(D_{\text{Grohe}}\):

**Lemma H.1.** ([26]) Let \(h\) be an homomorphism from \(D\) to \(D_{\text{Grohe}}\) such that \(h_0 \circ h\) is the identity. Then \(G\) contains a \(k\)-clique.

Our variant still enjoys many of the good properties of Grohe’s database, and some additional properties concerning the satisfaction of frontier-guarded TGDs, as shown in the following Theorem (which is essentially an equivalent reformulation of Theorem 7.1) for the case when CQs allow for free variables.

**Lemma H.2.** Let \(G, k, D, D', A, D'' = D'(G, D, D', A, \mu)\), and \(h_0\) be as defined above. The following statements hold:

1. There is a polynomial-time algorithm that, given \(G, D, D', A, \mu\), computes \(D''\).
2. \(h_0\) is a surjective homomorphism from \(D^*\) to \(D'\).
3. \(G\) contains a \(k\)-clique if there is a homomorphism \(h\) from \(D\) to \(D''\) such that \(h_0(h(\cdot))\) is the identity on \(A\).
4. If \(D' \models \Sigma\), and every clique of size at most \(3 \cdot r\) in \(G\) is contained in a clique of size \(3 \cdot r \cdot m\), then \(D'' \models \Sigma\).

**Proof.** Items (1) and (2) follow directly from the definition of \(D''\) and \(h_0\). However, for item (3) it is necessary to note that since the arity of the schema \(S\) is bounded by the fixed integer \(r\), in the construction of \(D\) it is enough to consider only all labelled cliques of size at most \(r\).

Let us prove (3). Although \((\equiv)\) can be obtained easily by mimicking the proof in [26], we shall give a proof using Lemma H.1. Assume that \(h\) is a homomorphism from \(D\) to \(D''\) such that \(h_0 \circ h\) is the identity on \(A\). Then the restriction \(h_{1A}\) of \(h\) to \(A\) is a homomorphism from \(D_{1A}\) to \(D''_{1A}\) such that \(h_0 \circ h_{1A}\) is the identity. But then \(h_{1A}\) corresponds to a homomorphism from \(D_{1A}\) to \(D(G, D_{1A}, D_{1A}, \text{dom}(D_{1A})), \mu)\) with \(h_0 \circ h_{1A}\) being the identity. Since \(D(G, D_{1A}, D_{1A}, \mu)\) is contained in the original Grohe’s database, \(D_{\text{Grohe}}(G, D_{1A}, \mu)\), it follows from Lemma H.1 that \(G\) contains a \(k\)-clique. For the converse \((\Rightarrow)\) assume that \(G\) has a \(k\)-clique, or alternatively, that there is a labelled clique \(\eta\) in \(G\) with domain \([k]\). Then \(\eta\) covers all elements in \(\text{dom}(D'')\) and, hence, the mapping \(h : \text{dom}(D) \rightarrow \text{dom}(D'')\) with \(h(z) = \tilde{z}_\eta\), for every \(z \in \text{dom}(D)\), defines a homomorphism from \(D\) to \(D''\) satisfying the conditions.

Let us finally show (4). Consider an arbitrary frontier-guarded TGD \(\phi(x, \tilde{y}) \models \exists z \psi(x, \tilde{z})\) in \(\Sigma\), and let \((\tilde{a}, \tilde{b})\) be an arbitrary tuple in the evaluation of \(\phi(\tilde{x}, \tilde{y})\) over \(D''\). Since \(h_0\) is a homomorphism from \(D'\) to \(D''\), we have that \(D''\) contains all facts in \(\phi(h_0(\tilde{a}, \tilde{b}))\). Therefore, since \(D'' \models \Sigma\), there is a tuple \(\tilde{d}\) of elements in \(\text{dom}(D'')\), for \(|\tilde{d}| = |\tilde{z}|\), such that \(D''\) contains every atom in \(\psi(h_0(\tilde{a}), \tilde{d})\). Since there is an atom guarding \(\tilde{x}\) in \(\phi\), it follows from the definition of \(D''\) that \(\tilde{a} = h_0(\tilde{a})\). Therefore, since \(\tilde{a}\) is a labelled clique \(\eta\) covering \(h_0(\tilde{a})\). In addition, \(\eta \in \Sigma_{\text{FG}}\), and hence \(\eta\) can have at most \(r \cdot m\) elements. It follows from the assumptions of item (4) that there is a labelled clique \(\eta'\) in \(G\), extending \(\eta\), that covers all elements mentioned by \(\tilde{d}\) (in addition to all elements in \(h_0(\tilde{a})\) already covered by \(\eta\)). Thus, \(D''\) contains every atom in \(\psi(h_0(\tilde{a}), \tilde{d})\).

**H.2 A Crucial Lemma**

Our proof borrows several ideas and techniques from the proof of Grohe’s Theorem. However, the adaptation of such techniques is non-trivial for the reasons that we have already discussed in Section 5.2, and we briefly recall here again. First, we cannot construct an arbitrary database \(D\), but we need to construct one that satisfies the given set \(\Sigma\) of frontier-guarded TGDs. Second, the notion of core of a CQ, which is crucial for the proof of Grohe, cannot be directly used in our context. Instead, we need to develop a technical lemma, which, intuitively speaking, states that some subsets of our CQs behave like cores for the sake of our proof.

**Lemma H.3.** Fix \(\ell \geq r \cdot m\). There is a computable function that takes as input a CQS \(\Sigma = (\Sigma, \varphi(x))\) from \((\Sigma_{\text{FG}}, \Sigma_{\text{CQ}})\) that is not uniformly \(\Sigma_{\text{FG}}\)-equivalent, and a positive integer \(s \geq 1\), and outputs a CQ \(\varphi(x_\Sigma)\), a subset \(X\) of the existentially quantified variables of \(p\), and a CQ \(p''(x_\Sigma)\), such that:

1. \(\varphi \equiv \varphi_{\Sigma}\).
2. \(|\varphi(p''(x_{\Sigma}))| = |\Sigma|\).
3. \(|\varphi(p)\| \subseteq |\varphi(p'')|\).
4. For every homomorphism \(h\) from \(p\) to \(p''\) with \(h(\tilde{x}) = \tilde{x}\), we have that \(h(X) = X\).
5. The treewidth of \(\varphi_{\Sigma}\) is larger than \(\ell\).
6. For every CQ \(p''(x_\Sigma)\) with at most \(s\) variables, if \(\tilde{x} \in p''(\varphi(p''))\) then also \(\tilde{x} \in p''(\varphi(\varphi(p, \Sigma)))\).

**Proof.** Let us assume that \(\varphi\) is defined over schema \(S\), that \(q(\tilde{x})\) has \(n > 0\) variables, and that \(t\) is the maximum number of variables appearing in the body of some TGD in \(\Sigma\). We set \(s' = \max\{s, n, t\}\). First of all, we can compute by exhaustive search a CQ \(p(\tilde{x})\) with a minimum number of variables that satisfies \(q(\tilde{x})\). This is because the notion \(\equiv_{\Sigma}\) is decidable, for \(\Sigma \in \Sigma_{\text{FG}}\). In fact, to check whether \(p \models q\) one can construct the finite instance \(M = (\varphi(D), \Sigma, n)\) from Definition 6.5, which exists by the strong finite controllability of \(\Sigma\) as stated in Theorem 6.7, and then check whether \(\tilde{x} \in q(M)\). We know that the latter holds if \(\tilde{x} \in q(\varphi(\varphi(p, \Sigma)))\). From Proposition 4.5 we get that \(p \models q\). Analogously, one can check \(q \models p\).

We have then that item (1) holds by definition. We now explain how to build \(p\) and \(X\). We need some preparation. First, we shall
Theorem 6.7, each $R \in \bar{\Sigma}$, and set $\bar{\Sigma}'$ of frontier-guarded TGDs in the following way. Initially, set $S = S$ and $\bar{\Sigma}' = \bar{\Sigma}$. Then, for each $\sigma = \phi(x', y) \rightarrow \exists \psi(x', \bar{w})$ in $\bar{\Sigma}$, for every partition $\chi \cap \chi' = \phi$ of the atoms in $\phi$, and for every CQ $q'(\bar{z}) = \exists \psi(\chi, \bar{z}, \bar{w})$ obtained by quantifying existentially some variables in $\chi$, we include in $\bar{\Sigma}'$ a new predicate $T_{q'}(\bar{z})$ with the same arity as $\bar{z}$.

Furthermore, for every atom $R(\bar{u})$ where $R \in S$ and $\bar{u}$ might contain variables occurring in $\phi$ and new fresh variables, we include in $\bar{\Sigma}'$ two new TGDs $\sigma'$ and $\sigma''$ (provided they are in $\bar{\Sigma}$) such that $\sigma''$ is obtained by replacing, in $\phi$, $\chi(\bar{u}, \bar{z})$ by $T_{q'}(\bar{z}) \wedge R(\bar{u})$. The following claim follows easily from the definition.

**Claim H.4.** For every $S$-instance $D$, and every atom $R(\bar{c})$ where $R \in S$ and $\bar{c}$ only mentions elements from $D$, 

$$R(\bar{c}) \in \text{chase}(D, \Sigma) \iff R(\bar{c}) \in \text{chase}(D, \bar{\Sigma}')$$

We use $A$ to denote the collection of all subsets $A$ from $\text{dom}(p)$ that are guarded, i.e., there exists an atom $R(\bar{p})$ of $p$ such that $\bar{a}$ mentions all elements in $A$. For every $A \in \mathcal{A}$, let $E_A$ be the $A$-database that contains all facts $R(\bar{b}) \in \text{chase}(D[p], \bar{\Sigma}')$ for which it is the case that all elements from $\bar{b}$ appear in $A$. Notice that, by Theorem 6.7, each $E_A$ can be constructed from $D[p]$, $\bar{\Sigma}'$, and $\mathcal{A}$. It follows also from Theorem 6.7 that we can construct a finite $\bar{\Sigma}'$-instance $M_A = \mathcal{M}(E_A, \bar{\Sigma}', x')$ with the property that, for every CQ $p''(\bar{x})$ with at most $s$ variables,

$$\bar{a} \in p''(M_A) \iff \bar{a} \in p''(\text{chase}(E_A, \bar{\Sigma}'))$$

for every tuple $\bar{a}$ of elements in $\text{dom}(E_A)$ of the same arity as $\bar{x}''$.

Let us define $D_A$ by removing from $M_A$ those atoms that are not over $\bar{\Sigma}$. We can assume (renaming elements if necessary) that $\text{dom}(D_A) \cap \text{dom}(D_A) = A \cap A'$ for every $A, A' \in \mathcal{A}$. We then define $p''$ as $\bigcup_{A \in \mathcal{A}} D_A$. We have by construction that $D[p] \subseteq D[p'']$, i.e., item (3) holds.

We now show item (6). We actually prove something stronger:

**Claim H.5.** Let $p''(\bar{x})$ be a CQ with at most $s$ variables and $\bar{c} \in \text{dom}(p)[\bar{x}]$ such that $\bar{c} \in p''(D[p''])$. Then $\bar{c} \in p''(\text{chase}(D[p], \bar{\Sigma}))$.

**Proof.** Assume that $\bar{c} \in p''(D[p''])$, i.e., there exists a homomorphism $h$ from $p''$ to $p'$ with $h(\bar{x}) = \bar{c}$. Let $p''(\bar{x}) = \exists \psi(\bar{x}, \bar{y})$ and for every $A \in \mathcal{A}$, let $\phi_A$ be the conjunction of all atoms, $R(\bar{z})$, in $\phi$ such that $h(\bar{z})$ is entirely contained in $\text{dom}(D_A)$. Notice that if we define $p''_{A'}$ as the CQ $\exists \psi(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are all variables of $\phi_A$ occurring in $\bar{x}$ and $\bar{y}$, respectively, then $p''_{A'} \equiv \bigwedge_{A \in \mathcal{A}} p''_{A'}$.

Since $h(\bar{x}_A) \in p''_{A'}(M_A)$, we have by definition that $h(\bar{x}_A) \in p''_{A'}(M_A)$. Since $p''_{A'}$ has at most $s \leq s'$ variables, it follows then from the definition of $M_A$ that $h(\bar{x}_A) \in p''_{A'}(\text{chase}(E_A, \bar{\Sigma}'))$, and hence $h(\bar{x}_A) \in p''_{A'}(\text{chase}(D[p], \bar{\Sigma}'))$. Since $A$ was arbitrarily chosen, we can then conclude that $h(\bar{x}) \in p''(\text{chase}(D[p], \bar{\Sigma}))$. Consequently, since $p''$ only contains predicates in $S$ it follows directly from Claim H.4 that $h(\bar{x}) \in p''(\text{chase}(D[p], \bar{\Sigma}))$.

We then show that item (2) holds.

**Claim H.6.** It is the case that $D[p'] \models \Sigma$.

**Proof.** Let $\sigma = \phi(x', \bar{y}) \rightarrow \exists \psi(x', \bar{w})$ be any TGD in $\Sigma$ and let $h$ be any mapping of the variables in $\phi$ to elements in $\text{dom}(p')$ such that $h(\bar{x}') \subseteq p'. \sigma$ Our goal is to show that $h(\bar{x})$ belongs to the evaluation of $\exists \psi(x', \bar{w})$ in $D[p']$. We can assume that $h$ is injective (no two variables from $\phi$ are mapped to the same element in $\text{dom}(p')$) since the non-injective case is reduced to the injective one. Indeed, if $h$ is non-injective we only need to replace $\sigma$ by the TGD obtained by identifying variables in the body that are mapped by $h$ to the same element (which by assumption also belongs to $\bar{\Sigma}'$) and also modify $h$ accordingly.

Since $\Sigma \in \mathcal{P}_G$, it follows that there exists some $A \in \mathcal{A}$ such that $h(\bar{x}')$ is entirely contained in $\text{dom}(D_A)$. If $h(\bar{y})$ is also contained in $\text{dom}(D_A)$ then there is nothing else to prove since, by construction, $D_A$ satisfies $\Sigma$.

Assume, then, that $h(\bar{y})$ is not contained in $\text{dom}(D_A)$. Let $R(\bar{a})$ be the atom in $p$ that guards $A$. Let $\bar{a} = (a_1, \ldots, a_i)$ and let $\bar{u} = (u_1, \ldots, u_i)$ be a tuple of variables where for every $1 \leq i \leq s$, $u_i$ is a new fresh variable if $a_i$ does not belong to the image of $h$, and $u_i = h^{-1}(a_i)$ otherwise (note that here we are using the fact that $h$ is injective).

Let $\bar{y}'$ be a tuple containing all variables in $\bar{y}$ that are not mapped by $h$ inside $\text{dom}(D_A)$, let $\chi(\bar{y}', \bar{z})$ be the conjunction of all atoms in $\phi$ containing some variable in $\bar{y}'$ (where $\bar{z}$ is a tuple that contains the rest of the variables in $\chi$), and let $q'(\bar{z}) = \exists \psi(\bar{y}', \bar{z})$. Note that $h(\bar{z}) \subseteq A$ as, by construction, if a tuple in $p'$ contains elements in $\text{dom}(D_A) \setminus A$ then all of its elements must necessarily be contained in $A$. Since every element mentioned in $h(\bar{z})$ belongs to $A$, it follows that the TGD $\sigma' = \chi(\bar{y}', \bar{y}) \wedge R(\bar{a}) \rightarrow R(\bar{z})$ is frontier-guarded by $R(\bar{a})$ and, consequently, it belongs to $\bar{\Sigma}'$. Let $h'$ be the extension of $h$ so that every fresh variable $u_i$ in $\bar{u}$ is mapped by $h'$ to $a_i$. Note that the image of $h'$ is contained in $\text{dom}(p')$ and hence it follows, by applying TGD $\sigma'$ with mapping $h'$, that $R_{q'}(h(\bar{z})) \in (\text{chase}(D[p'], \bar{\Sigma}'))$. Consequently, $R_{q'}(h(\bar{z}))$ belongs to $E_A$ and, hence, to $M_A$, as well.

Consider the TGD $\sigma'' = \chi(\bar{y}', \bar{y}) \wedge R(\bar{a}) \in \text{chase}(D[p], \bar{\Sigma})$. Then $h(\bar{x}') \in \text{chase}(D[p], \bar{\Sigma})$. In the latter case there is nothing else to prove so we can assume that $\chi$ contains some atom guarding $\bar{x}'$. By definition, this atom contains some elements not in $D_A$ which implies that $\bar{x}'$ is entirely contained in $A$. Consequently $\bar{x}'$ is guarded by $R(\bar{u})$.

Now, since $R_{q'}(h(\bar{z})) \in M_A$ it follows, by applying $\sigma''$ with mapping $h'$, that $h(\bar{x}')$ belongs to the evaluation of $\exists \psi(x', \bar{w})$ over $M_A$. Since all atoms in $\psi$ involve only predicates from $S$, it follows that $h(\bar{x}')$ belongs to the evaluation of $\exists \psi(x', \bar{w})$ over $D_A$, and hence over $D[p']$.

Before continuing, we prove the following important properties of the homomorphisms from $p$ to $p'$.

**Claim H.7.** Let $h$ be any homomorphism from $p$ to $p'$ with $h(x') = \bar{x}$. The following statements hold:

1. If $\bar{c} \models h(\bar{x})$, then $\bar{c} \models p'(\text{dom}(h(p)))$.
proving that at least some connected component of $G$ we have that at least some connected component of $G$ is linked by an edge. Moreover, since $A$ is a guarded in $\text{dom}(p)$, the elements in $A \cap h(Y)$ define a clique on $H_A$. Hence, there is a node $v \in T_A$ such that $A \cap h(Y) \subseteq \chi_A(v)$. Likewise, there is $u \in T$ such that $A \cap h(Y) \subseteq \chi(u)$. Thus, we can construct a new tree decomposition by joining $u$ and $v$ with an edge. Iterating this process for all $A \in \mathcal{A}$ we get a tree decomposition of $G'_{|h(Y)}$, whose width is the maximum between $r \cdot m$ and the width of $T$. Since $h$ is injective by item (b), $G'_{|h(Y)}$ has a tree decomposition of the same width as $G'_{|h(Y)}$.

Let $k$ be the treewidth of $G'$. Since $q \equiv_p p$ and $q$ is not uniformly $\text{CQ}_l$-equivalent, it follows that $k > \ell$. A set $Y \subseteq \text{dom}(p)$ \( \chi \) is nice if $G'_{|h(Y)}$ has treewidth $k$ and $|Y|$ is minimal with such property. We construct $X$ as the set containing all elements that belong to some nice set. It follows directly from the definition that property (5) holds. To finish, we show that property (4) is satisfied.

Claim H.8. Let $h$ be a homomorphism from $p$ to $p^\prime$ with $h(x) = \bar{x}$. Then $h(X) = X$.

Proof. Let $Y$ be any nice set. From Claim H.7(c), the treewidth of $G'_{|h(Y)}$ is at most the maximum between $r$, and the treewidth of $G'_{|h(Y) \cap \text{dom}(p)}$. Since $G'_{|h(Y)}$ has treewidth $k$ and $k > \ell \geq r \cdot m$, the treewidth of $G'_{|h(Y) \cap \text{dom}(p)}$ is at least $k$, implying that it is, in fact, $k$ since the treewidth of $G'$ is $k$. By the minimality of $|Y|$, it must be the case that $h(Y) \subseteq \text{dom}(p)$. Therefore, $h(Y)$ is also a nice set and, hence, $h(X) \subseteq X$. Since $h$ is injective from Claim H.7(b), it must be the case that $h(X) = X$.

This finishes the proof of Lemma H.3.

We now proceed to explain how Lemma H.3 is applied in order to prove Theorem 5.13.

H.3 The FPT-reduction

We have all the ingredients needed to define the fpt-reduction used in the proof of Theorem 5.13. Let $(G, k)$ be an instance of p-Clique. It is easy to see that we can assume, without loss of generality, that every clique of size at most $3 \cdot r$ in $G$ is contained in a clique of size $3 \cdot r \cdot m$; recall that $r$ is the maximum arity of the predicates occurring in $\text{CQSs}$ of $\mathcal{O}$, while $m$ is the maximum number of atoms in the head of the TGDs occurring in $\text{CQSs}$ of $\mathcal{O}$.

Claim H.9. p-Clique is $W[1]$-hard, even if restricted to instances that satisfy the above restriction.

Henceforth, for $k \geq 1$ we let $K = \left\{ \frac{1}{2k} \right\}$. From the Excluded Grid Theorem [34], there is a computable function $F : \mathbb{N} \rightarrow \mathbb{N}$ such that, for each $k \geq 1$ and simple graph $G$ of treewidth at least $F(k)$, we have that at least some connected component of $G$ contains a $(k \times k)$-grid as a minor. By hypothesis on the class $\mathcal{O}$, there exists a $\text{CQS}$ $S = (\Sigma, \alpha')$ from $\mathcal{O}$ such that $S \not\in \langle \text{FG}_m, \text{UCQ}_l \rangle_{F(k)}$, where $F$ is as defined above. W.l.o.g., we assume that $F(k) \geq r \cdot m$. We build from $(G, k)$ an instance $(D', \Sigma, q'(x))$ of $p$-$\text{CQS-Evaluation}(\mathcal{O})$, where $D'$ is a database defined as follows.

First, observe that there must be a $\text{CQ}$ $q(x)$ in $q'(x)$ such that $(\Sigma, q(x)) \not\in \langle \text{FG}_m, \text{UCQ}_l \rangle_{F(k)}$, while $q \not\preceq \gamma$ for every disjunct $q'$ other than $q$. In fact, if we remove from $q'$ every disjunct that is not maximal with respect to $\preceq_{\Sigma}$, then we obtain an equivalent $\text{UCQ}$ $q''(\bar{x})$ under $\Sigma$. Therefore, $(\Sigma, q''(x)) \not\in \langle \text{FG}_m, \text{UCQ}_l \rangle_{F(k)}$ and, hence, there is at least one disjunct $q$ in $q''$ such that $(\Sigma, q(x)) \not\in \langle \text{FG}_m, \text{UCQ}_l \rangle_{F(k)}$. 

\( \square \)
Since, by assumption, $F(k) \geq r \cdot m$, it is possible to compute from $q(x)$ and $s$, where $s$ is the maximum number of variables over all CQs of $q'(x)$, a CQ $p(q)$, a subset $X$ of the existentially quantified variables of $p$, and a CQ $p'(x)$, that satisfy the properties stated in Lemma H.3 for $\ell = F(k)$. In particular, the treewidth of $G_{\bar{\Sigma}}^D$ is at least $F(k)$, and hence, by the Excluded Grid Theorem, there exists a connected component $H$ of $G_{\bar{\Sigma}}^D$ and a minor map $\mu$ from the $(k \times k)$-grid onto $H$. We then define $D' = D'(G,D[p],D[p'],X,\mu)$, where $D'(G,D[p],D[p'],X,\mu)$ is our modified version of Grohe’s database that satisfies the properties stated in Lemma H.2.

H.4 Correctness of the Reduction

It remains to show that the above is an fpt-reduction from $p$-Clique to CQs-Evaluation(\emptyset). To this end, we need to show that

**Lemma H.10.** The following statements hold:

1. $D' \models \Sigma$.
2. $G$ has a $k$-clique iff $\bar{x} \in q'(D')$.
3. There are computable functions $f,g : \mathbb{N} \to \mathbb{N}$ such that $(D',\Sigma,q(\bar{x}))$ can be constructed in time $f(k) \cdot \|G\|^{O(1)}$ and $(\|q'(\bar{x})\| + \|\Sigma\|) \leq g(k)$.

**Proof.** The proof of (1) follows from the last item in Theorem H.2 since $D[p'] \models \Sigma$ and every clique of size at most $3 \cdot r$ in $G$ is contained in a clique of size $3 \cdot r \cdot m$. It is worth noticing though that this holds for our variant of Grohe’s database, but not necessarily for the one that Grohe originally defined in [26].

We now proceed to show item (2).

$(\Rightarrow)$ Assume that $G$ has a $k$-clique. Then, by Lemma H.2, there is a homomorphism $h$ from $p$ to $D'$ such that $h_0 \circ h$ is the identity on $X$. It follows directly from the definition of $D'$ that the mapping $h'$ defined as $h'(x) = h(x)$, when $x \in X$, and $h'(x) = x$, otherwise, is also a homomorphism from $p$ to $D'$. In particular, since $X$ contains only quantified variables we have that $h'(x) = \bar{x}$, and hence $\bar{x} \in p(D')$. But $q \equiv \Sigma p$, and thus, $\bar{x} \in q(D')$ since, by item (1), $D' \models \Sigma$. Therefore, $\bar{x} \in q'(D')$ as $q$ is a disjunct of $q'(D')$.

$(\Leftarrow)$ Conversely, assume that $\bar{x} \in q'(D')$. First observe that $\bar{x} \in q(D')$. By contradiction, assume otherwise. Then, there is a CQ $\tilde{q}(\bar{x})$ in $q'(\bar{x})$, other than $q$, such that $\bar{x} \in \tilde{q}(D')$. Clearly, $D'$ homomorphically maps to $p'$ via the mapping $h_0$, and hence, there is a homomorphism from $\tilde{q}$ to $p'$ mapping $\bar{x}$ to $\bar{x}$. By Lemma H.3, $\bar{x} \in \tilde{q}(\text{chase}(p,\Sigma))$, which in turn implies that $\bar{x} \models q(D')$. Since $\bar{x} \models q(D')$, and hence, there is a homomorphism $h$ from $p$ to $D'$ such that $h(\bar{x}) = \bar{x}$. It follows that $h_0 \circ h$ is a homomorphism from $p$ to $p'$ that maps $\bar{x}$ to $\bar{x}$. In consequence, from Lemma H.3 we obtain that $h(X) = X$. Therefore, there must exist some $m \geq 0$ such that $g = h \circ (h_0 \circ h)^m$ is a homomorphism from $p$ to $D'$ that satisfies that $h_0 \circ g$ is the identity on $X$. It follows from Lemma H.2 that $G$ has a $k$-clique.

As for item (3), first notice that the CQs $S = (\Sigma,q')$ can be computed by simply enumerating the CQs from $\emptyset$ until we find $S$ since, by Theorem 5.10, we can check whether $S \not\equiv (FG,\text{CQ})^{\text{UCQ}}_{(k)}$. The same holds for $G$. From $G$ we can construct the CQs $p$ and $p'$, as well as the set of variables $X$, by applying Lemma H.3. We can then compute $\mu$ via an exhaustive search over $G^P$. Notice that the construction of $q'$, $q,p,p',X$, and $\mu$ depends only on $k$. Lemma H.2 states, on the other hand, that it is possible to construct $D'$ in time polynomial, given $p',p,X,\mu$, and $G$. Putting all these together, we obtain the existence of computable functions $f,g : \mathbb{N} \to \mathbb{N}$ as required in item (3), and the claim follows.

1 PROOF OF LEMMA 6.6

For brevity, let $\text{finans}(q,D,\Sigma) = \bigcap_{M \in \text{fmods}(D,\Sigma)} q(M).

$(\Rightarrow)$ Consider an $S$-database $D$, a set $\Sigma \in C$ of TGDs over $S$, and an integer $n \geq 0$. Assuming that $\text{dom}(D) = \{d_1, \ldots, d_f\}$, for $\ell \geq 1$, let $D' = D \cup \{\text{Dom}(d_1), \ldots, \text{Dom}(d_f)\}$, and $\bar{d} = (d_1, \ldots, d_f)$.

**Lemma L.1.** There exists an instance $M^* \in \text{fmods}(D^*,\Sigma)$ such that, for every UCQ $q^*$ over $S \cup \{|\text{Dom}\}$ of arity $\ell$ with at most $n + \ell$ variables, $\bar{d} \in q^*(M^*)$ implies $\bar{d} \in q^*(\text{chase}(D^*,\Sigma))$

**Proof.** Let $q$ be the UCQ of arity $\ell$ consisting of all the CQs $q'$ with at most $n + \ell$ variables such that $\bar{d} \not\in q'(\text{chase}(D^*,\Sigma))$. Clearly, $\bar{d} \not\in q'(\text{chase}(D^*,\Sigma))$. Since $C$ is finitely constructible, we conclude that $\bar{d} \not\in \text{finans}(q,\text{Dom}(D^*),\Sigma)$. Thus, there exists an instance $M^* \in \text{fmods}(D^*,\Sigma)$ such that $\bar{d} \not\in q(M^*)$. Fix an arbitrary UCQ $q^*$ of arity $\ell$ with at most $n + \ell$ variables. Observe that $\bar{d} \not\in q^*(\text{chase}(D^*,\Sigma))$ implies $\bar{d} \not\in q^*(M^*)$ since, by construction, each CQ in $q^*$ occurs also in $q$, and we know that $\bar{d} \not\in q(M^*)$. The claim follows.

We claim that the desired finite model $M(D,\Sigma,n)$ is the instance $M^*$ where $M^*$ is the $(S \cup \{|\text{Dom}\})$-instance of $\text{fmods}(D^*,\Sigma)$ provided by Lemma L.1. To this end, we need to show that, for every UCQ $q(\bar{x})$ over $S$ of arity $r \geq 0$ with at most $n$ variables, $\bar{d} \in q(\text{chase}(D,\Sigma))$ implies $\bar{d} \in q(\text{chase}(D,\Sigma))$. Clearly, $\bar{d} \in q(\text{chase}(D,\Sigma))$.

Fix an arbitrary tuple $\bar{d} \in \text{dom}(D)^\ell$. Clearly, $\bar{d} \in q(\text{chase}(D,\Sigma))$ since $M^*$ is a model of $D$ and $\Sigma$. Assume now that $\bar{d} \in q(M^*)$. We proceed to show that $\bar{d} \in q(\text{chase}(D,\Sigma))$. It is not difficult to see that there exists a UCQ $\tilde{q}$ of arity $\ell$ over $(S \cup \{|\text{Dom}\})$ with at most $n + \ell$ variables such that $\bar{d} \not\in q(M^*)$.

Since $\bar{d}$ has at most $n + \ell$ variables, by Lemma L.1, we get that $\bar{d} \not\in q(\text{chase}(D^*,\Sigma))$.

By hypothesis, $\bar{d} \not\in q(\text{chase}(D^*,\Sigma))$, and hence, by (2), we conclude that $\bar{d} \not\in q(\text{chase}(D,\Sigma))$, as needed.

$(\Leftarrow)$ Consider an $S$-database $D$, a set $\Sigma \in C$ of TGDs over $S$, and a UCQ $q^*$ over $S$. Fix an arbitrary tuple $\bar{d} \in \text{dom}(D)^\ell$. We need to show that $\bar{d} \not\in q(\text{chase}(D,\Sigma))$ if $\bar{d} \not\in \text{finans}(q,D,\Sigma)$. Clearly, the $(\Rightarrow)$ direction holds trivially. It remains to show that $\bar{d} \not\in q(\text{chase}(D,\Sigma))$ implies $\bar{d} \not\in \text{finans}(q,D,\Sigma)$. Let $n \geq 0$ be the number of variables occurring in $q$. Since $C$ is strongly finitely constructible, there exists $M(D,\Sigma,n) \in \text{fmods}(D,\Sigma)$ such that $\bar{d} \not\in q(M(D,\Sigma,n))$. This immediately implies that $\bar{d} \not\in \text{finans}(q,D,\Sigma)$.
PROOF OF THEOREM 6.7

To show that $\mathcal{F}\mathcal{G}$ is strongly finitely controllable, by Lemma 6.6, it suffices to show that it is finitely controllable. This can be easily shown by exploiting the fact that the guarded negation fragment of first-order logic (GNFO) [5] enjoys the finite model property, i.e., if a GNFO sentence has a model, then it has a finite one. But let us first recall the guarded negation fragment.

GNFO restricts first-order logic by requiring that all occurrences of negation are of the form $\alpha \land \neg \varphi$, where $\alpha$ is an atom containing all the free variables of $\varphi$. Formally, the formulas of GNFO are generated by the recursive definition

$$\varphi ::= R(t_1, \ldots, t_n) \mid t_1 = t_2 \mid \varphi_1 \land \varphi_2 \mid \varphi_1 \lor \varphi_2 \mid \exists x \varphi \mid \alpha \land \neg \varphi,$$

where each $t_i$ is a term (constant or variable), and in the last clause, $\alpha$ is an atomic formula containing all free variables of $\varphi$. We know that GNFO enjoys the finite model property, and we also have an upper bound on the size of finite models:

**Proposition J.1 ([5]).** Consider a GNFO sentence $\varphi$. If $\varphi$ has a model, then it has a finite one of size $2^{2^{2^{O(1)}}}$.

We proceed to show that $\mathcal{F}\mathcal{G}$ is finitely controllable. Consider a database $D$, a set $\Sigma \in \mathcal{F}\mathcal{G}$, and a UCQ $q(\bar{x})$. We need to show that

$$q(\text{chase}(D, \Sigma)) = \bigcap_{M \in \text{fmods}(D, \Sigma)} q(M).$$

It should be clear that the (⊆) direction holds trivially. It remains to show the (⊇) direction. Consider a tuple $\bar{c} \in \text{dom}(D)^{|\bar{x}|}$. Our goal is to devise a GNFO sentence $\Phi$ such that

1. $\bar{c} \in \bigcap_{M \in \text{fmods}(D, \Sigma)} q(M) \implies \Phi$ does not have a finite model.
2. $\Phi$ does not have a model $\implies \bar{c} \in q(\text{chase}(D, \Sigma)).$

Having such a sentence in place we get the (⊇) direction since, by Proposition J.1, we can conclude that if $\Phi$ does not have a finite model, then it does not have a model at all, which in turn implies that $\bar{c} \in q(\text{chase}(D, \Sigma))$, as needed. Let us now explain how $\Phi$ is constructed. We first observe that a frontier-guarded TGD $\sigma$ of the form $\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})$ can be equivalently rewritten as

$$\phi_\sigma = \neg \left( \exists \bar{x} \exists \bar{y} (\varphi(\bar{x}, \bar{y}) \land \neg \exists \bar{z} \psi(\bar{x}, \bar{z})) \right),$$

which is a GNFO sentence since all the free variables of $\exists \bar{z} \psi(\bar{x}, \bar{z})$ occur in the guard of $\sigma$. We also observe that the sentence $\neg q(\bar{c})$ is trivially a GNFO sentence. Therefore,

$$\Phi = D \land \bigwedge_{\sigma \in \Sigma} \phi_\sigma \land \neg q(\bar{c}),$$

which is clearly the desired GNFO sentence.

It remains to show that a finite witness is realizable, i.e., there is a computable function that takes as input an $S$-database $D$, a set $\Sigma \in \mathcal{F}\mathcal{G}$ over $S$, and an integer $n \geq 0$, and outputs the finite model $M(D, \Sigma, n)$. In fact, this follows from the proof of Lemma 6.6 and Proposition J.1. In the proof of Lemma 6.6, we actually show that $M(D, \Sigma, n)$ is the model of a GNFO sentence $\varphi$, and thus, due to Proposition J.1, we can assume that is of size $2^{2^{O(1)}}$. Therefore, we can compute the finite model $M(D, \Sigma, n)$ as follows:

1. we first compute the set $S_n$ of all the non-isomorphic UCQs over $S$ with at most $n$ variables, which is clearly finite; and