On the Structure of Queries in Constraint Query Languages

Michael Benedikt and Leonid Libkin

On the Structure of Queries in Constraint Query Languages

Michael Benedikt
AT&T Bell Laboratories
1000 E Warrenville Rd
Naperville, IL 60566
E-mail: benedikt@research.att.com

Leonid Libkin
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
E-mail: libkin@research.att.com

Abstract

We study the structure of first-order and second-order queries over constraint databases. Constraint databases are formally modeled as finite relational structures embedded in some fixed infinite structure. We concentrate on problems of elimination of constraints, reducing quantification range to the active domain of the database and obtaining new complexity bounds. We show that for a large class of signatures, including real arithmetic constraints, unbounded quantification can be eliminated. That is, one can transform a sentence containing unrestricted quantification over the infinite universe of discourse to get an equivalent sentence in which quantifiers range over the finite relational structure. We use this result to get a new complexity upper bound on the evaluation of real arithmetic constraints. We also expand upon techniques in [20] and [4] for getting upper bounds on the expressiveness of constraint query languages, and apply it to a number of first-order and second-order query languages.

1 Introduction

Techniques of finite model theory have found applications in a number of areas such as database theory [1] and descriptive complexity [17]. Database applications of finite model theory stem from one of the basic results of relational database theory: Classical query languages, such as relational algebra, have precisely the power of first-order logic. Since relational databases can be viewed as finite models in the language of the relational schema, this basic observation allows us to apply the tools of finite model theory to study expressibility of relational query languages. Early work in that direction includes [2, 7, 10]; for a survey see [1].

In recent years, various extensions of the basic relational model have been studied. Two most notable ones are extensions to complex objects, or nested relations [27], which underlie most object-oriented datamodels, and extensions to *constraint databases* [18, 19], which are used as the basis for geographical and temporal datamodels. For the nested relations most basic questions about expressive power and the structure of queries have been answered (see [27] and references therein), but only very recently has some progress been made for constraint databases.

The framework of constraint databases assumes some underlying model $\mathcal{M} = \langle \mathbb{U}, \Omega \rangle$ where $\mathbb{U}$ is a set (always assumed to be infinite in this paper), and Ω is a signature that consists of a number of interpreted functions and predicates over $\mathbb{U}$. For instance, the domain very often considered for geographical databases is $\langle \mathbb{R}, +, *, 0, 1, < \rangle$, with the intention that databases represent some regions on the real plane. In the classical framework of [19], (generalized) databases over $\mathcal{M}$ are given by quantifier-free formulae $\phi(x_1, \ldots, x_n)$ in the language of Ω; such a database represents the set

$$\mathcal{M}_\phi = \{\vec{a} = (a_1, \ldots, a_n) \mid \vec{a} \in \mathbb{U}^n,\ \mathcal{M} \models \phi(\vec{a})\}.$$

For example, a convex polygon with known set of vertices can easily be represented in such a way. The model $\mathcal{M}$ is typically chosen to admit quantifier elimination. Then the query evaluation process reduces to application of the quantifier elimination procedure, see [19].

Recently, attention has shifted from finitely represented models (that is, those that arise as $\mathcal{M}_\phi$) to finite ones, cf. [4, 20, 23, 24]. Since queries arising in geographical applications often involve regions that are determined by a fixed finite number of points (i.e. a convex polygon can be given by its vertices), we can convert most interesting questions involving finitely-represented queries to ones involving finite ones [21]. For example, it was conjectured that first-order logic with polynomial inequality constraints cannot express topological connectivity. A result of [23] reduced that problem to connectivity of finite graphs whose nodes come from $\mathbb{R}$, and the problem for graph connectivity was recently solved in [4]. In addition, the setting of finite databases embedded in a fixed infinite structure enables one to study constraint databases via the tools of finite (and infinite) model theory, and allows for helpful characterizations of the expressive power of classes of queries even in settings where quantifier elimination does not hold in the underlying structure.

We will work in the setting of finite databases embedded in infinite fixed structures. We start with the underlying model $\mathcal{M} = \langle \mathbb{U}, \Omega \rangle$ and add a number of predicate symbolds $R_1, \ldots, R_k$, R_i being of arity τ_i, for finite database relations. These will be interpreted as τ_i-ary relations over $\mathbb{U}$. Following the database tradition, we will call $R_1, \ldots, R_k$ a *schema*, and denote it by SC. As our main language we take $\mathcal{FO}(\mathcal{M}, SC)$, the first-order logic over the language $L(SC, \Omega)$ that contains Ω and SC. If SC is understood, we often omit it. We will mostly deal with sentences, since most of the results for sentences can be extended (as we'll show) for arbitrary formulae.

Let $\mathit{Inst}(\mathbb{U}, SC)$ be the set of k-tuples of finite relations over $\mathbb{U}$, the ith one being of arity τ_i (that is, the set of possible finite instances of SC over $\mathbb{U}$). For $D \in \mathit{Inst}(\mathbb{U}, SC)$ and ϕ a $L(\mathcal{M}, SC)$ sentence, we define $D \models \phi$ in the usual way.

Although the syntax of our formulas can be straightforwardly adapted from classical predicate logic, the properties of queries can depend in subtle ways on the domain of quantification. For $D \in \mathit{Inst}(\mathbb{U}, SC)$, the minimum possible range of quantification is the *active domain* of D: $\mathit{adom}(D)$ is the set of all elements of $\mathbb{U}$ that occur in relations in D, cf. [1, 16]. For $A \subseteq \mathbb{U}$, we write $D \models_A \phi$ if $D \models \phi'$ where ϕ' is obtained from ϕ by replacing each quantifier $\mathrm{Q}x$ by its bounded version $\mathrm{Q}x \in \mathit{adom}(D) \cup A$.

There are two cases of the $\models_A$ relations that are of special interest: $\models_{\mathbb{U}}$ is the usual entailment relation $\models$ (which is sometimes called the *natural interpretation* of queries), and $\models_\emptyset$ restricts quantification to the active domain of the finite D (this is sometimes called the *active*, or *active-domain interpretation* of queries). To see the difference between the two, assume that $\mathcal{M} = \langle \mathbb{R}, < \rangle$, and ϕ is $\exists x \forall y.(x < y) \vee (x = y)$. Then, for any nonempty $D \in \mathit{Inst}(\mathbb{R}, SC)$, $D \models \neg\phi$ but $D \models_\emptyset \phi$.

Much of what has been done in constraint databases addresses the problem of evaluating constraint queries; that is, queries in $\mathcal{FO}(\mathcal{M}, SC)$, or in a language based on another logic (e.g. fixpoint logic). To find satisfactory query evaluation algorithms for a query language, we have to address the following issues.

$\angle$ Expressive Power. Classical relational query languages have been studied in great depth, and their expressive power is well known. Prior results allow one to infer what sorts of recursion constructs are necessary to express properties such as parity, connectivity and others that arise in database applications. We also know much about the impact of adding these programming constructs on query optimization. For constraint query languages, an understanding of many of the fundamental expressivity questions is still lacking. In particular, there is a need for tools to assist in getting upper bounds for the expressivity of constraint languages. In this paper we continue the work of [4, 20, 24] in getting techniques to bound the expressivity of constraint query languages. We show how to extend the results of [4] to show equivalence in expressive power for many first-order and second-order constraint query languages. We prove several kinds of *collapse results*, which say that adding new predicates or functions to the signature Ω does not significantly increase expressive power.

$\angle$ Range of Quantification. While databases themselves are finite, the natural range of quantification for constraint queries is the whole universe $\mathbb{U}$. Thus, we need tools to reduce the problem of query evaluation to a finitary process. By choosing $\mathcal{M}$ to be decidable, we guarantee the ability to evaluate constraint queries for a fixed database, since given a query ϕ, we can replace each occurence of $R_i(\vec{x})$ in ϕ by $\vec{x} = t_1 \vee \ldots \vee \vec{x} = t_m$ where $R_i = \{t_1, \ldots, t_m\}$ and apply the decision procedure to the resulting formula. However, this still forbids us from doing important compile-time query optimizations that are possible in the classical database setting. In particular, we lack the ability to reorder quantifiers based on the range of quantification, or to evaluate expressions 'bottom-up' by retrieving stored values of subexpressions. One possible solution seems to be this: try to show that unbounded quantification can be eliminated. This is equivalent to showing that every query ϕ has an equivalent one under the active interpretation. That is, there is a φ such that $D \models \phi$ iff $D \models_{\emptyset} \varphi$. We will prove that this is possible for many constraint query languages of interest, and that it holds for the real ordered field.

$\angle$ Complexity of Constraints. There are many fundamental questions to be answered about the complexity of query evaluation for constraint query languages. Since the relational algebra and calculus are equivalent to pure first-order logic, they have AC^0 complexity [1]. Adding constraints increases this complexity. For instance, if multiplication is in the signature, the AC^0 complexity bound is lost, cf. [6]. As an upper bound for complexity, it is known that if $\mathcal{M} = \langle \mathbb{R}, +, *, 0, 1, < \rangle$, then data complexity of first-order queries is NC, see [19, 5]. Since $AC^0 \subset NC$, one could hope for more precise information about the complexity of constraint queries over the real field. We would also like to know something about the effect of adding other interpreted structure on these complexity bounds, both for first-order and higher-order logics. In this paper we will use results on equivalence of signatures and on bounding quantification to get tighter bounds on query evaluation for the real field, and to get complexity bounds for a variety of other first-order and second-order languages. One tool for doing this will be *partial collapse results*: results that show that a certain set of operations in the signature can be eliminated, assuming that our databases have all their elements coming from a certain infinite set. Using these results, we will be able to get tighter complexity and expressive bounds for queries that are *generic* [1] (those invariant under certain endomaps on $\mathbb{U}$), since their behavior over any infinite set determines their behavior globally.

Organization and quick summary In this paper we offer a detailed study of the structure of constraint queries, addressing the three issues described above.

In section 2 we introduce the 'approximate collapse' arrow relation $\mathcal{FO}(\mathcal{M}) \Rightarrow \mathcal{FO}(\mathcal{M}')$, for two models $\mathcal{M}$ and $\mathcal{M}'$ on the same set $\mathbb{U}$, meaning (informally) this: for every query ϕ in $\mathcal{FO}(\mathcal{M})$, one can find an infinite set $X \subseteq \mathbb{U}$ and a query ψ in $\mathcal{FO}(\mathcal{M}')$ such that for every $D \in Inst(\mathbb{U}, SC)$ with $adom(D) \subset X$, $D \models_X \phi$ iff $D \models_X \psi$. We are interested in the case when $\mathcal{M}'$ is a reduct of $\mathcal{M}$. The arrow relation shows that we have a means for reducing expressivity and complexity questions about $\mathcal{M}$ to ones concerning $\mathcal{M}'$, a tool we will use later on in the paper.

In section 3 we state approximate collapse results for query languages, along the lines of [24],[4],[20]: we show that one can get an infinite set on which all constraints in a query can be reduced to constraints in smaller languages. We prove such results for first-order logic and for fragments of second-order logic, and show how they can be used to get expressivity bounds for first- and second-order queries.

In section 4 we study the ability to eliminate unbounded quantification in queries in favor of quantification bounded by the active domain. Our main result is Theorem 4, which shows that unbounded quantification can be removed for all models that admit quantifier elimination and satisfy the condition of *o-minimality* [25]. This class includes both cases for which the elimination result is known [16, 24], and also the important case of $\langle \mathbb{R}, +, *, 0, 1, < \rangle$, thus solving the open problem from [24].

In section 5 we show that collapse results and bounded quantification results can always be extended from boolean queries to nonboolean queries (that is, from sentences to arbitrary formulae). We apply these results to get expressivity and complexity bounds for nonboolean queries.

In section 6, we apply the results of section 4 to prove a TC^0 complexity bound for first-order logic with polynomial constraints, thus improving the NC bound of [19]. We also establish some complexity bounds for active-domain second-order constraint queries.

Section 7 contains concluding remarks.

All proofs can be found in the appendix.

2 Notations

Assume that the domain is an infinite set $\mathbb{U}$. A schema is a nonempty collection $SC = \langle R_1, \ldots, R_k \rangle$ of relation names, R_i being of arity τ_i. A *database instance* D of schema SC is given by an interpretation of each relational symbol R_i as a finite τ_i-ary relation over $\mathbb{U}$. The set of all instances is denoted by $Inst(\mathbb{U}, SC)$. The *active domain* of D, $adom(D)$ is the set of all elements in $\mathbb{U}$ that in relations in D.

Let Ω be a signature, that is, a collection of interpreted functions and predicates on $\mathbb{U}$. The language that contains the schema predicates, equality and the symbols in Ω is denoted by $L(SC, \Omega)$[1]. A *boolean query* is a first-order sentence in $L(SC, \Omega)$. That is, it is built up from atomic formulae via the usual logical connectives and quantifiers of the form $\forall x$ and $\exists x$.

[1] All languages we consider are assumed to have equality, so we will not mention this explicitly anymore.

Let A be a subset of $\mathbb{U}$. Under the *A-interpretation* of queries, we assume that for every D the quantifiers range over $A \cup adom(D)$. That is, the $\emptyset$-interpretation is the active domain interpretation, and the $\mathbb{U}$-interpretation is the natural interpretation. We use $D \models_A \phi$ to mean that ϕ is satisfied by D under the A-interpretation.

The class of Boolean queries (maps from instances of schema SC to $\{\mathbf{T}, \mathbf{F}\}$) under the A-interpretation is denoted by $\mathcal{FO}^A(\mathbb{U}, \Omega, SC)$. If $A = \mathbb{U}$ we omit it, i.e. we use $\mathcal{FO}(\mathbb{U}, \Omega, SC)$ for $\mathcal{FO}^{\mathbb{U}}(\mathbb{U}, \Omega, SC)$ and $\models$ for $\models_{\mathbb{U}}$. We write $\mathcal{FO}^A(\mathbb{U}, \Omega)\{=, \subseteq\}\mathcal{FO}^A(\mathbb{U}, \Theta)$ to mean that $\mathcal{FO}^A(\mathbb{U}, \Omega, SC)\{=, \subseteq\}\mathcal{FO}^A(\mathbb{U}, \Theta, SC)$ holds for any schema SC.

The kind of unbounded-quantifier elimination result we are interested in can be written as $\mathcal{FO}(\mathbb{U}, \Omega) = \mathcal{FO}^{\emptyset}(\mathbb{U}, \Omega)$. Two such results are known:

Fact 1 (see [16, 24]) *$\mathcal{FO}(\mathbb{U}, \emptyset) = \mathcal{FO}^{\emptyset}(\mathbb{U}, \emptyset)$; $\mathcal{FO}(\mathbb{R}, +, -, 0, 1, <) = \mathcal{FO}^{\emptyset}(\mathbb{R}, +, -, 0, 1, <)$.* $\square$

In section 4 we will extend these results.

We are also interested in equivalence of constraint languages, which in our notation can be written as $\mathcal{FO}(\mathbb{U}, \Omega) = \mathcal{FO}(\mathbb{U}, \Theta)$ where generally Θ is "simpler" than Ω.These kinds of results are hard to achieve, and often we can only find a certain approximation to equality. For this, we need the following notation.

Definition 1 *We write*

$$\mathcal{FO}^A(\mathbb{U}, \Omega, SC) \Rightarrow \mathcal{FO}^B(\mathbb{U}, \Theta, SC)$$

if, for every $L(SC, \Omega)$ sentence ϕ, we can find an infinite set $X \subseteq \mathbb{U}$ and a $L(SC, \Theta)$ sentence ψ such that for any D with $adom(D) \subseteq X$,

$$D \models_{A \cap X} \phi \;\text{ iff }\; D \models_{B \cap X} \psi.$$

We write $\mathcal{FO}^A(\mathbb{U}, \Omega) \Rightarrow \mathcal{FO}^B(\mathbb{U}, \Theta)$ if $\mathcal{FO}^A(\mathbb{U}, \Omega, SC) \Rightarrow \mathcal{FO}^B(\mathbb{U}, \Theta, SC)$ holds for any schema SC.

Reformulating this definition, $\mathcal{FO}^A(\mathbb{U}, \Omega) \Rightarrow \mathcal{FO}^B(\mathbb{U}, \Theta)$ means that for every query Q in

$\mathcal{FO}^{A\cap X}(\mathbb{U},\Omega)$, we can find an infinite $X \subseteq \mathbb{U}$ and $Q' \in \mathcal{FO}^{B\cap X}(\mathbb{U},\Theta)$ such that $Q(D) = Q'(D)$ for any D with $adom(D) \subseteq X$.

The relation $\Rightarrow$ is an approximation to inclusion; it is an approximation in the sense that it is the inclusion restricted to models from a certain infinite set. We also further restrict quantification to that set X. For instance, if $A = B = \mathbb{U}$ (and this is the situation we encounter most often), then $\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}(\mathbb{U},\Theta)$ means that there is an infinite set X such that $\mathcal{FO}(X,\Omega) \subseteq \mathcal{FO}(X,\Theta)$, i.e. we have the inclusion for the natural interpretation over an infinite X.

Thus, the desired collapse results would be $\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}^{\emptyset}(\mathbb{U},\Omega)$, or $\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}(\mathbb{U},\Theta)$, when X happens to be $\mathbb{U}$.

We can analogously define similar arrow notation for nonboolean queries. The framework of nonboolean queries assumes the *output schema* $SC' = \{T_1,\ldots,T_l\}$, $l > 0$. Then a first-order query is given by a formula $\phi(x_1,\ldots,x_n)$, for each n-ary output relation. Such a formula defines the relation (under A-interpretation) given by $\{\vec{a} \in \mathbb{U}^n \mid D \models_A \phi(\vec{a})\}$ for each input $D \in Inst(\mathbb{U}, SC)$. We denote the class of nonboolean queries with input schema SC and output schema SC' by $\mathcal{FFO}_{SC'}(\mathbb{U},\Omega,SC)$. (Here $\mathcal{FFO}$ stands for "full first-order", as opposed to sentences only.)

The definition of the arrow relation generalizes straightforwardly. That is, we must say $D \models_{A\cap X} \phi(\vec{a})$ iff $D \models_{B\cap X} \psi(\vec{a})$ for any $\vec{a} \in (adom(D) \cup X)^n$. Finally, we write

$$\mathcal{FFO}(\mathbb{U},\Omega)\ \{\Rightarrow,=\}\ \mathcal{FFO}(\mathbb{U},\Theta)$$

if $\mathcal{FFO}_{SC'}(\mathbb{U},\Omega,SC) \Rightarrow \mathcal{FFO}_{SC'}(\mathbb{U},\Theta,SC)$ (respectively for $=$) holds for any pair of input and output schemas SC and SC'.

3 Collapsing signatures

3.1 First-order logic

The goal of this section is to investigate the approximate collapse relation. We are interested in results collapsing queries over signature Ω to queries over Θ, where Θ is much simpler than Ω. We start by reviewing the first-order case. We state a generalization of the result from [4] and [20], which can be used to get expressivity bounds on first-order constraint languages (this technique is already implicit in [24]). For completeness, we include a full proof of these results in the appendix. We will then make use of the techniques developed in these proofs to extend the arrow relation results to second-order logic and to existential second-order. In the full paper, we will also be considering the case for various infinitary logics.

Our first approximate collapse result from [4] shows that any signature approximately collapses to the order relation.

Theorem 1 *Let $\mathbb{U}$ be ordered by $<$. Then for any $L(SC,\Omega)$ sentence ϕ, we can find an infinite set $X \subseteq \mathbb{U}$ and a $L(SC,<)$ sentence ψ such that for any $D \in Inst(SC,X)$, and any $Y \subseteq X$, it is the case that*

$$D \models_Y \phi \quad \textit{iff} \quad D \models_Y \psi.$$

Corollary 1 *Let $\mathbb{U}$ be ordered by $<$. Then, for any signature Ω,*

$$\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}(\mathbb{U},<).$$

Theorem 1 was proved for the case $Y = \emptyset$ in [4]. Results of this kind are particularly useful for studying expressibility under the active-domain semantics, as demonstrated in [4] and [20]. For example, many queries of interest are generic, that is, independent under permutations of the underlying domain. For such queries, their behavior on an infinite subset of $\mathbb{U}$ fully determines their behavior on $\mathbb{U}$. For example, it can be immediately derived from theorem 1 that for any Ω, there is no $L(SC,\Omega)$ sentence ϕ such that $D \models_{\emptyset} \phi$ iff $adom(D)$ has even cardinality.

It is generally impossible to eliminate the order relation from the right hand side of the arrow relation. However, it was shown in [4] that for signatures over the reals satisfying certain smoothness conditions, collapse results to pure relational algebra are available.

Definition 2 *A signature Ω on $\mathbb{R}$ is called* analytic *if it consists of restrictions of analytic functions to real arguments.*

In other words, $\Omega = (f_i)_{i\in I}$ is analytic if there is a set of analytic functions $(F_i)_{i\in I}$ such that each f_i is the restriction of F_i to the real arguments. For example, $(+, *, \Gamma(x), e^x)$ is an analytic signature.

Theorem 2 *Let Ω be an analytic signature on the reals. Then for any $L(SC, \Omega)$ sentence ϕ, we can find an infinite (in fact, uncountable) set $X \subseteq \mathbb{U}$ and a $L(SC)$ sentence ψ such that for any $D \in Inst(SC, X)$, and any $Y \subseteq X$, it is the case that*

$$D \models_Y \phi \quad \textit{iff} \quad D \models_Y \psi.$$

A roughly analagous result was proved in [4], although the results there contained extra hypotheses and provided no cardinality information. As a corollary, we obtain the following fact about the arrow relation.

Corollary 2 *Let Ω be an analytic signature on the reals. Then*

$$\mathcal{FO}(\mathbb{R}, \Omega) \Rightarrow \mathcal{FO}(\mathbb{R}, \emptyset).$$

3.2 Second-order logic

The goal of this section is to generalize approximate collapse results to second-order logic and its fragments. When we deal with sentences, we assume that they are converted into normal form. That is, sentences

$$\mathrm{Q}'_1 P_1 \ldots \mathrm{Q}'_m P_m \mathrm{Q}_1 x_1 \ldots \mathrm{Q}_n x_n . \phi(x_1, \ldots, x_n)$$

where $\mathrm{Q}'_i P_i$ are second-order quantifiers, and ϕ is a first-order formula in the language that contains $P_1, \ldots, P_m$. Now, suppose that $D \in Inst(SC, \mathbb{U})$. For Φ a second-order sentence, we define $D \models \Phi$ in the usual way. Furthermore, for $X \subseteq \mathbb{U}$ we define $D \models_X \Phi$ by letting all first-order quantifiers range over $adom(D) \cup X$ and letting each second-order quantifier $\mathrm{Q}'_i P_i$ range over $2^{(adom(D)\cup X)^k}$, if P_i is of arity k.

For a schema SC, a signature Ω on $\mathbb{U}$, the class of second-order constraint queries, under the X interpretation, will be denoted by $\mathcal{SO}^X(\mathbb{U}, \Omega, SC)$. That is, $\mathcal{SO}^X(\mathbb{U}, \Omega, SC)$ is the class of queries Q over some schema SC for which there exists a second-order sentence Φ such that $Q(D) = \mathbf{T}$ iff $D \models_X \Phi$. As before, we write $\mathcal{SO}$ instead of $\mathcal{SO}^{\mathbb{U}}$. Similary to the first order case, omitting the schema in a statement about equation or arrow relation means "for all schemas".

We shall also consider fragments of second-order logic given by the quantifier prefixes of second-order quantifiers. Formally, a prefix is a finite sequence of pairs $(i, \forall)$ or $(i, \exists)$ where $i > 0$. Then a (normal form) second-order sentence conforms to the prefix $(i_1, \mathrm{Q}^1), \ldots, (i_s, \mathrm{Q}^s)$ if it has s second-order quantifiers, the jth one is Q^j and it binds predicates of arity i_j.

A fragment F is the set of sentences that conform to some collection of prefixes. Examples of fragments are full second-order logic, existential second-order logic and monadic Σ^1_1. The fragment associated with a set of prefixes F (under the X-interpretation) will be denoted by $\mathcal{SO}^X_F(\mathbb{U}, \Omega, SC)$, or $\mathcal{SO}_F(\mathbb{U}, \Omega, SC)$ if $X = \mathbb{U}$.

We call a fragment given by a collection of prefixes F *orderable* if $F \neq \emptyset$ and for every $f \in F$, $(2, \exists)$ followed by f is still in F. Examples are full second-order, and existential second-order $\mathcal{ESO}$. A non-orderable fragment is monadic Σ^1_1. Now we can prove a result that extends approximate collapse theorems to second-order constraint queries, and further generalizes them for orderable fragments of second-order logic.

Theorem 3 *1) Let F be an arbitrary fragment. Then $\mathcal{SO}_F(\mathbb{U}, \Omega) \Rightarrow \mathcal{SO}_F(\mathbb{U}, <)$. Furthermore, $\mathcal{SO}_F(\mathbb{R}, \Omega) \Rightarrow \mathcal{SO}_F(\mathbb{R})$ if Ω is analytic.*

2) Let F be an orderable fragment of second-order logic. Then

$$\mathcal{SO}_F(\mathbb{U}, \Omega) \Rightarrow \mathcal{SO}_F(\mathbb{U})$$

In particular, $\mathcal{SO}(\mathbb{U}, \Omega) \Rightarrow \mathcal{SO}(\mathbb{U})$ and $\mathcal{ESO}(\mathbb{U}, \Omega) \Rightarrow \mathcal{ESO}(\mathbb{U})$. These results are true for the active-domain interpretation as well; that is, $\mathcal{SO}^{\emptyset}_F(\mathbb{U}, \Omega) \Rightarrow \mathcal{SO}^{\emptyset}_F(\mathbb{U})$. □

As we saw earlier, any approximate collapse result completely describes the behavior of generic queries (those invariant under permutations of the domain, such as parity test or transitive closure). Thus, we obtain

Corollary 3 *If F is an orderable fragment, and Ω is an arbitrary signature, then every generic query in $\mathcal{SO}_F^{\emptyset}(\mathbb{U}, \Omega)$ is expressible in $\mathcal{SO}_F^{\emptyset}(\mathbb{U})$.* □

From this we get some expressivity bounds. For example, connecitivity of directed graphs is not definable under the active interpretation as a monadic Σ_1^1 constraint query, no matter what operations are in the signature. Similarly, any query that is complete for exponetial space cannot be defined as a second-order constraint query under the active interpretation.

4 Eliminating unbounded quantification

In this section we prove that unbounded quantification can be eliminated in favor of quantification bounded by the active domain for a large class of structures. This is equivalent to saying that the active-domain interpretation and the natural interpretation coincide for a large class of structures.

Recall that a structure $\langle \mathbb{U}, \Omega \rangle$, where the order relation $<$ is in Ω, is called *o-minimal* [25] if every definable set $\{c \in \mathbb{U} \mid \langle \mathbb{U}, \Omega \rangle \models \phi(c)\}$ is composed of a finite union of intervals. Here ϕ is a formula in the language that includes all symbols of Ω and constants for elements of $\mathbb{U}$. Examples of o-minimal structures are $\langle \mathbb{R}, < \rangle$, $\langle \mathbb{R}, +, *, 0, 1, < \rangle$ (this follows from quantifier elimination [8]) and $\langle \mathbb{R}, +, *, e^x \rangle$ [28].

Our main result is as follows:

Theorem 4 *Let Ω be a signature on $\mathbb{U}$ such that $\langle \mathbb{U}, \Omega \rangle$ is o-minimal and admits quantifier elimination. Then*

$$\mathcal{FO}(\mathbb{U}, \Omega) = \mathcal{FO}^{\emptyset}(\mathbb{U}, \Omega)$$

That is, every first-order query using symbols from Ω and the schema relations is equivalent to a formula where the quantifiers are bounded by the active domain. □

The proof in the appendix merely establishes the existence of an active-domain query that is equivalent to a query using unbounded quantification. However, the process of transforming an unbounded-quantifier sentence into a bounded-quantifier sentence can be done effectively assuming that the quantifier elimination procedure for the underlying model is effective. We shall present such a procedure in the full version.

Now, using the o-minimality of $\langle \mathbb{R}, +, *, 0, 1, < \rangle$ and theorem 4, we settle the open problem from [24].

Corollary 4 *Every first-order query in the language of the schema relations and $+, *, 0, 1, <$ can be expressed by a formula in the same language with all quantifiers bounded by the active domain. That is, $\mathcal{FO}(\mathbb{R}, +, *, 0, 1, <) = \mathcal{FO}^{\emptyset}(\mathbb{R}, +, *, 0, 1, <)$.* □

For example, consider $\langle \mathbb{R}, +, *, 0, 1, < \rangle$, and let our schema have a binary predicate $S(x, y)$. The sentence ϕ states that all elements of S lie on some line:

$$\phi \;=\; \exists a \exists b \forall x \forall y. (S(x, y) \rightarrow y = a * x + b)$$

This gets converted to the equivalent active-domain sentence $\psi = (card(S) < 3) \vee ((card(S) \geq 3) \wedge \psi')$ where the consitions on cardinality of S are written as first-order sentences in the language of S, and ψ' is

$$\exists x_1 \exists y_1 \forall x_2 \forall y_2 \; \forall x_3 \forall y_3 . S(x_1, y_1) \;\wedge\; (S(x_2, y_2) \;\wedge\; S(x_3, y_3) \rightarrow (x_2 - x_1)(y_3 - y_1) = (y_2 - y_1)(x_3 - x_1))$$

Then $D \models \phi$ iff $D \models_{\emptyset} \psi$ for any $D \in \mathit{Inst}(\mathbb{R}, \{S\})$.

The analog of corollary 4 for linear constraints was proved in [24]. These result stand in sharp contrast to the results of [13], who showed that elimination of unbounded quantifiers fails for integer arithmetic constraints.

The ability to convert natural semantics queries to bounded-quantifier queries is important for achieving efficient query evaluation. In addition, corollary 4 gives us an alternative proof of the conjecture that parity test cannot be defined by first-order queries that use polynomial inequality constraints (this conjecture was recently confirmed in [4]). Indeed, assume that parity is definable in such a way; then it is definable under the active semantics, and we know from the previous section that this is not the case.

Some other corollaries can be derived from theorem 4.

Corollary 5 *Let Ω be an o-minimal signature on $\mathbb{U}$. Then we can find an expansion of Ω to Ω' such that*

$$\mathcal{FO}(\mathbb{U},\Omega) = \mathcal{FO}(\mathbb{U},\Omega') = \mathcal{FO}^{\emptyset}(\mathbb{U},\Omega').$$

Corollary 6 *Assume that $\mathbb{U}$ is a dense order without endpoints. Then for an arbitrary signature Ω we have $\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}^{\emptyset}(\mathbb{U},<)$. Also, $\mathcal{FO}(\mathbb{R},\Omega) \Rightarrow \mathcal{FO}^{\emptyset}(\mathbb{R})$ for any analytic signature Ω on the reals.* □

Other consequences of Thoerem 4 will be given in section 6.

We note the hypothesis that the underlying order is dense in corollary 6 cannot be removed:

Proposition 1 *Let $(\mathbb{U},<)$ be a scattered linear ordering. Then $\mathcal{FO}(\mathbb{U},<) \not\Rightarrow \mathcal{FO}^{\emptyset}(\mathbb{U},<)$.* □

The coincidence of the active and natural interpretations proved for the first-order logic with polynomial constraints does not extend to the second-order case.

Proposition 2 *For any fragment F that allows existential quantifiers over unary predicates (e.g., full second-order, existential second-order, monadic Σ_1^1), $\mathcal{SO}_F(\mathbb{R},+,*,0,1,<) \neq \mathcal{SO}_F^{\emptyset}(\mathbb{R},+,*,0,1,<)$.* □

5 Extension to nonboolean first-order queries

In this section we show the following: *all results concerning boolean queries extend to arbitrary nonboolean queries.* In other words, all results we proved for sentences can be also proved for formulae with free variables. We show this by proving "transfer" results that extend a $\Rightarrow$ or an equality result from boolean queries to arbitrary ones.

Since we are now interested in arbitrary queries, we need two schemas: the *input schema* SC_1 and the *output schema* SC_2. Given the underlying model $\langle \mathcal{M},\Omega\rangle$, a query is given by a formula $\phi(x_1,\ldots,x_n)$, in the language $L(SC_1,\Omega)$, for each n-ary predicate symbol in SC_2. For each input $D \in \mathit{Inst}(SC,\mathbb{U})$ and each set X, under the X-interpretation such a query defines the n-ary relation $\phi_X[D] = \{\vec{a} \mid \vec{a} \in \mathbb{U}^n, D \models_X \phi(\vec{a})\}$. The class of first-order queries between the schemas SC_1 and SC_2 is denoted by $\mathcal{FO}^X_{SC_2}(\mathbb{U},\Omega,SC_1)$, and $\mathcal{FFO}^X(\mathbb{U},\Omega,)$ denotes the set of all nonboolean first-order queries. Recall that all of these are sets of *semantic objects.*

Note that $\phi_X[D]$ need not be finite. That is, $\mathcal{FO}^X_{SC_2}(\mathbb{U},\Omega,SC_1)$ is actually a set of maps from $\mathit{Inst}(SC_1,\mathbb{U})$ to $\mathit{Inst}_\infty(SC_2,\mathbb{U})$, where $\mathit{Inst}_\infty(\cdot)$ is the class of finite and infinite instances. Since we are often interested in the class of maps from $\mathit{Inst}(SC_1,\mathbb{U})$ to $\mathit{Inst}(SC_2,\mathbb{U})$, we define a restriction on queries that guarantees finiteness.

We call a query Q *domain-preserving* if for any input D, $\mathit{adom}(Q(D)) \subseteq \mathit{adom}(D)$. That is, no element of $\mathbb{U}$ can be present in $Q(D)$ unless it is present in D itself. (Every query expressed in relational algebra or safe relational calculus is such.) For a class of queries $\mathcal{C}$, we denote the subclass of domain-preserving queries in $\mathcal{C}$ by $\mathsf{dp\_}\mathcal{C}$.

The arrow notation extends to nonboolean queries and to domain preserving queries in the natural way. Now we are ready to prove the first transfer theorem that allows us to extend the arrow results to arbitrary queries.

Theorem 5 *For any signatures Ω and Θ,*

$\mathcal{FO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}(\mathbb{U},\Theta)$ *implies* $\mathcal{FFO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FFO}(\mathbb{U},\Theta)$.

Furthermore, $\mathcal{FO}^A(\mathbb{U},\Omega) \Rightarrow \mathcal{FO}^B(\mathbb{U},\Theta)$ *implies* $\mathsf{dp\_}\mathcal{FFO}^A(\mathbb{U},\Omega) \Rightarrow \mathsf{dp\_}\mathcal{FFO}^B(\mathbb{U},\Theta)$. $\square$

From this we immediately obtain:

Corollary 7 *1)* $\mathcal{FFO}(\mathbb{U},\Omega) \Rightarrow \mathcal{FFO}(\mathbb{U},<)$.
2) $\mathcal{FFO}(\mathbb{R},\Omega) \Rightarrow \mathcal{FFO}(\mathbb{R},\emptyset)$*, if* Ω *is analytic.*
3) If $\mathbb{U}$ *is a dense order without endpoints, then* $\mathsf{dp\_}\mathcal{FFO}(\mathbb{U},\Omega) \Rightarrow \mathsf{dp\_}\mathcal{FFO}^\emptyset(\mathbb{U},<)$.
4) If Ω *is analytic, then* $\mathsf{dp\_}\mathcal{FFO}(\mathbb{R},\Omega) \Rightarrow \mathsf{dp\_}\mathcal{FFO}^\emptyset(\mathbb{R},\emptyset)$.

Note that the right hand sides in 3) and 4) are the classes of queries well known in the classical relational theory. Indeed, $\mathsf{dp\_}\mathcal{FFO}^\emptyset(\mathbb{U},SC)$ is precisely the class of queries that can be defined by the relational algebra, and $\mathsf{dp\_}\mathcal{FFO}^\emptyset(\mathbb{U},<,SC)$ is the class of queries definable by the relational algebra with $<$ comparisons allowed in selections.

Next, we prove a transfer theorem that allows us to extend elimination of unbounded qunatification to nonboolean queries.

Theorem 6 *Suppose that for some signature* Ω *it is the case that* $\mathcal{FO}(\mathbb{U},\Omega) = \mathcal{FO}^\emptyset(\mathbb{U},\Omega)$*. Then,*

$$\mathcal{FFO}(\mathbb{U},\Omega) = \mathcal{FFO}^\emptyset(\mathbb{U},\Omega).$$

Corollary 8 *Let* Ω *be o-minimal and admit quantifier elimination. Then* $\mathcal{FO}(\mathbb{U},\Omega) = \mathcal{FFO}^\emptyset(\mathbb{U},\Omega)$*. In particular, the classes of arbitrary nonboolean queries that are first-order expressible with polynomial ineqaulity constraints over the reals, are the same under both natural and active-domain interpretations.* $\square$

We can now use the above results to get some expressivity bounds.

Corollary 9 *The following cannot be expressed by any first-order constraint query under the active interpretation, nor by any first-order query with polynomial inequality constraints under the natural interpretation: transitive, or deterministic transitive closure of a graph; maximal matching in a bipartite graph; Eulerian cycle.* $\square$

6 Complexity of constraint queries

As was mentioned in the introduction, the results on elimination of unbounded qunatification allow us to prove new low complexity bounds. In this section we use theorem 4 to give a new complexity bound for $\mathcal{FO}(\mathbb{R},+,*,0,1,<)$. We also use corollary 3 to establish complexity bound on generic second-order queries.

We are dealing with data complexity, that is, the complexity of evaluating a given query for instances that vary. We only look at boolean queries here, but all results generalize easily for nonboolean queries. Assume some encoding of instances, for example, the encoding of [1]. Given $D \in \mathit{Inst}(SC,\mathbb{U})$, we denote its encoding by $enc(D)$. Then for each boolean query ϕ we define the language $L_\phi = \{enc(D) \mid D \in \mathit{Inst}(SC,\mathbb{U}), D \models \phi\}$. The data complexity of ϕ is the conventional complexity of L_ϕ. In particular, for any complexity class $\mathcal{C}$ we say that ϕ has $\mathcal{C}$ data complexity if the language L_ϕ is in $\mathcal{C}$.

It was previously known [19] that every query from $\mathcal{FO}(\mathbb{R},+,*,0,1,<)$ has NC data complexity. In fact, this follows from the NC complexity bound for the first-order theory of real closed fields with a fixed number of variables [5]. However, pure first-order logic queries, as well as first-order queries with linear constraints, have AC^0 data complexity [1, 14] and we know that $AC^0 \subset NC$ [6]. So the question arises if we can improve the data complexity bound for $\mathcal{FO}(\mathbb{R},+,*,0,1,<)$ using the elimination of unbounded quantification result proved in section 4. This is indeed possible. We prove below a TC^0 complexity bound. Recall that AC^0 is the class of problems definable with unbounded fan-in constant depth circiuts that use *and*, *or* and *not* gates, and the number of gates is polynomial in the size of input. The class TC^0 extends AC^0 by allowing threshold gates, or equivalently majority gates [3, 22]. It is known that $AC^0 \subset TC^0 \subseteq NC^1 \subseteq L \subseteq NL \subseteq NC$ and all $\subseteq$ inclusions are conjectured to be strict [3].

Theorem 7 *Every query in* $\mathcal{FO}(\mathbb{R},+,*,0,1,<)$

has TC^0 data complexity. $\square$

Since the behavior of generic (invariant under permutations) queries is fully determined by their behavior on an infinite set, we obtain from corollary 4 (this can also be derived from combining the results of [4] and [14]):

Corollary 10 *Every generic query in $\mathcal{FO}(\mathbb{R}, +, *, 0, 1, <)$ has AC^0 data complexity.* $\square$

From corollary 3 and classical descriptive complexity results (cf. [17]) we obtain

Corollary 11 *Every generic query in $\mathcal{SO}^{\emptyset}(\mathbb{U}, \Omega)$ has PH data complexity, and every generic query in $\mathcal{ESO}^{\emptyset}(\mathbb{U}, \Omega)$ has NP data complexity.* $\square$

7 Conclusions

Through the results of this paper, along with recent works such as [4, 20, 24], we have a much better feel for the expressive capabilities of constraint languages. Although the implications of these results for the design of query languages are dependent on many parameters of the application domain, we can draw a few general conclusions for language design.

Our results indicate that it is particularly promising to focus on constraint query languages over o-minimal structures. In particular, it seems that these languages inherit most of the pleasant formal properties of the pure relational calculus, along with uniform versions of the formal properties of first-order logic over the real ordered field. We can use the techniques developed here to get finer information about formal properties of definable sets that are posessed by these languages,

Although many of our main results yield constructive proofs, we have not fully explored all the algorithmic consequences of the theorems. In particular, we are interested in investigating semantic query optimization strategies enabled by theorems 1 and 4 in detail. The effective versions of theorem 4 can be seen as generalizations of the classical Tarski-Seidenberg algorithm for quantifier elimination. We are interested in seeing if algorithms based on these results can be useful in geometric theorem-proving applications that involving large numbers of rational parameters.

The results here can be seen from a mathematical view as yielding interesting information concerning the structure of sets definable from formulae with free second-order variables. As such, they can be seen as extending works such as [28], in showing the 'tame' behavior of important fragments of analytic geometry. In particular, results such as theorem 4 can yield interesting expressivity limits even in cases where there is no known effective procedure.

Many of the results within this paper point to connections between model-theoretic properties of a structure $\mathcal{M}$ and expressibility properties of the constraint query language based on $\mathcal{M}$. In future work, we plan to give more detailed information on this relationship, including results for constraint query languages based on classes appearing in model-theoretic stability theory.

The collapse results indicate that in many cases the class of generic queries expressible in a constraint query language equals a previously studied class. It may be possible, therefore, to classify the complexity classes that can be captured by the set of generic queries in some $\mathcal{FO}(\mathbb{U}, \Omega)$.

We are also interested in the interplay between Ramsey-like theorems and collapse results. We can, for instance, get additional information about collapsing sets by making use of results in Ramsey theory and set theory, cf. [9]. For instance: we extend the arrow notation by $\mathcal{FO}(\mathbb{U}, \Omega) \Rightarrow_\kappa \mathcal{FO}(\mathbb{U}, \Theta)$ if, for every $L(SC, \Omega)$ sentence ϕ, we can find an infinite set $X \subseteq \mathbb{U}$ of cardinality κ and a $L(SC, \Theta)$ sentence ψ such that for any $D \in \mathit{Inst}(SC, X)$, $D \models_X \phi$ iff $D \models_X \psi$. Furthermore, we use the notation $\Rightarrow_\kappa^\lambda$ to mean that the cardinality of $\mathbb{U} - X$ is λ. We can now show that:

Theorem 8 *1) There exists a signature Ω on $\mathbb{R}$ such that $\mathcal{FO}(\mathbb{R}, \Omega) \Rightarrow_\kappa \mathcal{FO}(\mathbb{R}, <)$ implies $\kappa = \aleph_0$.*

2) If Ω is analytic, then $\mathcal{FO}(\mathbb{R},\Omega) \Rightarrow_c \mathcal{FO}(\mathbb{R})$. However, $\mathcal{FO}(\mathbb{R},+,*,\Omega) \not\Rightarrow_c^{\aleph_0} \mathcal{FO}(\mathbb{R})$. $\square$

Results such as these may be useful in analyzing the finer structure of constraint queries (for instance, in the consideration of queries satisfying weaker notions of genericity), and in analyzing constraint queries with cardinality quantifiers. We plan to make a more detailed study of cardinality collapse results in subsequent work.

Acknowledgement: We thank Dan Suciu for helpful discussions of the material presented in section 6.

References

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

[2] A. V. Aho and J. D. Ullman. Universality of data retrieval languages. In *POPL'79*, pages 110–120.

[3] D.A. Barrington, N. Immerman, H. Straubing. On uniformity within NC^1. *JCSS*, 41:274–306,1990.

[4] M. Benedikt, G. Dong, L. Libkin and L. Wong. Relational expressive power of constraint query languages. Manuscript, 1995.

[5] M. Ben-Or, D. Kozen, J. Reif. The complexity of elementary algebra and geometry. *JCSS*, 32:251–264, 1986.

[6] R.B. Boppana and M. Sipser. The Complexity of Finite Functions. In *Handbook of Theoretical Computer Science*, Vol. A, chapter 14, (J. van Leeuwen editor), North-Holland, 1990.

[7] A. Chandra and D. Harel. Computable queries for relational databases. *JCSS*, 21(2):156–178, 1980.

[8] C.C. Chang and H.J. Keisler. *Model Theory.* North Holland, 1990.

[9] P. Erdős, A. Hajnal, A. Máté and R. Rado. *Combinatorial Set Theory: Partition Relations for Cardinals.* North Holland, 1984.

[10] H. Gaifman. On local and non-local properties. In *Proceedings of the Herbrand Symposium, Logic Colloquium '81*, pages 105–135, North Holland, 1982.

[11] R.L. Graham, B.L. Rothschild and J.H. Spencer. *Ramsey Theory.* John Wiley & Sons, 1990.

[12] S. Grumbach and J. Su. Finitely representable databases, In *PODS'94*, pages 289–300.

[13] S. Grumbach and J. Su. First-order definability over constraint databases. *Proc. Conf. Constraint Programming*, 1995.

[14] S. Grumbach, J. Su, and C. Tollu. Linear constraint databases. In *Proceedings of LCC, 1994.*

[15] C.W. Henson. The isomorphism property in nonstandard analysis and its use in the theory of Banach spaces. *Journal of Symbolic Logic* 39 (1974), 717–731.

[16] R. Hull and J. Su. Domain independence and the relational calculus. *Acta Informatica* 31:513–524, 1994.

[17] N. Immerman. Descriptive complexity: A logician's approach to computation. *Notices of the AMS* 42 (1995), 1127–1133.

[18] P. Kanellakis. Constraint programming and database languages: A tutorial. In *PODS'95*, pages 46–53.

[19] P. Kanellakis, G. Kuper, and P. Revesz. Constraint query languages. *JCSS* 51 (1995), 26–52. Extended abstract in *PODS'90.*

[20] M. Otto and J. Van den Bussche. First-order queries on databases embedded in an infinite structure. Technical Report, University of Antwerp, October 1995.

[21] C. Papadimitriou, D. Suciu and V. Vianu. Topological queries in spatial databases. *PODS'96*, to appear.

[22] I. Parberry and G. Schnitger. Parallel computation and threshold functions. *JCSS* 36 (1988), 278–302.

[23] J. Paredaens, J. Van den Bussche, and D. Van Gucht. Towards a theory of spatial database queries. In *PODS'94*, pages 279–288.

[24] J. Paredaens, J. Van den Bussche, and D. Van Gucht. First-order queries on finite structures over the reals. In *LICS'95*, pages 79–87.

[25] A. Pillay, C. Steinhorn. Definable sets in ordered structures. *Bulletin of the AMS* 11 (1984), 159–162.

[26] J. G. Rosenstein. *Linear Orderings.* Academic Press, New York, 1982.

[27] V. Tannen. Languages for collection types: A tutorial. in *PODS'94.*

[28] L. Van den Dries, A. Macintyre and D. Marker. The elementary theory of restricted analytic fields with exponentiation. *Annals of Mathematics* 85 (1994), 19–56.

APPENDIX

Proof of theorem 1

We write $\mathcal{FO}(\mathbb{U},\Omega) \Rrightarrow \mathcal{FO}(\mathbb{U},\Theta)$ if for any schema SC, any $L(SC,\Omega)$ sentence ϕ, we can find an infinite set $X \subseteq \mathbb{U}$ and a $L(SC,\Theta)$ sentence ψ such that for any D with $adom(D) \subseteq X$, and any $Y \subseteq X$, $D \models_Y \phi$ iff $D \models_Y \psi$. Now we have to show that $\mathcal{FO}(\mathbb{U},\Omega) \Rrightarrow \mathcal{FO}(\mathbb{U},<)$. We first need a simple observation.

Lemma 1 *Let Ω' be an expansion of Ω and Θ' be an expansion of Θ. Then $\mathcal{FO}(\mathbb{U},\Omega') \mapsto \mathcal{FO}(\mathbb{U},\Theta)$ implies $\mathcal{FO}(\mathbb{U},\Omega) \mapsto \mathcal{FO}(\mathbb{U},\Theta')$, where $\mapsto$ is either $\Rightarrow$ or $\Rrightarrow$.* $\square$

By lemma 1, we must show that $\mathcal{FO}(\mathbb{U},\Omega,<) \Rrightarrow \mathcal{FO}(\mathbb{U},<)$. Given a schema SC and any signature Θ, we call a formula of the form $R_i(z_1,\ldots,z_{\tau_i})$ or $\neg R_i(z_1,\ldots,z_{\tau_i})$ a *simple relational formula* if all z_is are variables. A formula of the form $T(t_1(\vec{z}_1),\ldots,t_m(\vec{z}_m))$ (or $\neg T(t_1(\vec{z}_1),\ldots,t_m(\vec{z}_m))$), or of the form $t_1(\vec{z}_1) = t_2(\vec{z}_2)$ (or $\neg(t_1(\vec{z}_1) = t_2(\vec{z}_2))$), is called a *simple constraint formula formula* if T is a Θ-predicate, all t_is are Θ-terms and all $\vec{z}_i$s contain only variables. Finally, a *simple formula* of $L(SC,\Theta)$ is a Boolean combination of simple relational and simple constraint formulae in $L(SC,\Theta)$.

The theorem will follow from the next three lemmas.

Lemma 2 *For any Θ and any $L(SC,\Theta)$ sentence Ψ, it is possible to construct another $L(SC,\Theta)$ sentence Φ of the form*

$$\Phi \;=\; \mathrm{Q}_1 x_1 \ldots \mathrm{Q}_n x_n . \phi(x_1,\ldots,x_n),$$

where $\phi(x_1,\ldots,x_n)$ is a simple $L(SC,\Theta)$ formula, and for any $D \in Inst(SC,\mathbb{U})$ and any $X \subseteq \mathbb{U}$,

$$D \models_X \Psi \leftrightarrow \Phi.$$

This simple lemma was proved in [4]. The next lemma gives our main construction. The reader can see that the proof of this lemma is essentially contained in the proof of theorem 1 of [4]; we give it here for the sake of completeness.

Lemma 3 *Let $\mathbb{U}$ be ordered and Θ contain the symbol $<$ for the order on $\mathbb{U}$. Then for any simple formula $\phi(x_1,\ldots,x_n)$ in $L(SC,\Theta)$, it is possible to find a simple $L(SC,<)$ formula $\psi(x_1,\ldots,x_n)$ and an infinite set $X \subseteq \mathbb{U}$ such that for any $D \in Inst(SC,X)$ and any $c_1,\ldots,c_n \in X$,*

$$D \models \phi(c_1,\ldots,c_n) \quad \textit{iff} \quad D \models \psi(c_1,\ldots,c_n)$$

Finally, we need the following:

Lemma 4 *Let $X \subseteq \mathbb{U}$ be any nonempty set. Let $\phi(x_1,\ldots,x_n)$ and $\psi(x_1,\ldots,x_n)$ be any two quantifier-free first-order formulae in any language containing the schema predicates such that for any $c_1,\ldots,c_n \in X$ and any $D \in Inst(SC,X)$, $D \models \phi(c_1,\ldots,c_n)$ iff $D \models \psi(c_1,\ldots,c_n)$. Then, for any quantifier prefix $\mathrm{Q}_1 \ldots \mathrm{Q}_n$ and any $Y \subseteq X$, $D \models_Y \mathrm{Q}_1 x_1 \ldots \mathrm{Q}_n x_n . \phi(x_1,\ldots,x_n)$ iff $D \models_Y \mathrm{Q}_1 x_1 \ldots \mathrm{Q}_n x_n . \psi(x_1,\ldots,x_n)$.*

To see how lemmas 2, 3, 4 imply the theorem, we start with an arbitrary sentence Φ. By lemma 2, we assume that it is in the prenex normal form with the quantifier-free part $\phi(x_1, \ldots, x_n)$ being a simple $L(SC, \Omega, <)$ formula. Applying lemma 3, we get an infinite set $X \subseteq \mathbb{U}$ and a $L(SC, <)$ formula $\psi(x_1, \ldots, x_n)$ equivalent to ϕ on instances over X when variables take values from X. Then applying lemma 4, we get that the sentence obtained by putting the quantifier prefix of Φ on top of $\psi(x_1, \ldots, x_n)$ gives us the desired $L(SC, <)$ sentence.

Thus, it remains to prove lemmas 3 and 4

Proof of lemma 3. Let $\phi(x_1, \ldots, x_n)$ be a simple formula in $L(SC, \Theta)$. We can assume without loss of generality that ϕ is in disjunctive normal form. We let $\phi_r(x_1, \ldots, x_n)$ be the disjunction of those disjuncts of ϕ in which all conjuncts are simple relational formulae; $\phi_c(x_1, \ldots, x_n)$ is defined as the disjunction of those in which at least one conjunct is a simple constraint formula.

We need some new notation. Let $V = \{x_1, \ldots, x_n\}$. An ordered partition on V is a partition $X_1, \ldots, X_s$ with the order on its blocks: $X_1 \prec \ldots \prec X_s$. The set of ordered partitions on V is denoted by $Part_{\prec}(V)$; for an ordered partition $P \in Part_{\prec}(V)$, $\xi(P)$ is the formula with free variables $x_1, \ldots, x_n$ in the language of $<$ that specifies this partition completely. For example, for $P = \{x_1, x_2\} \prec \{x_3, x_4\}$ on $\{x_1, x_2, x_3, x_4\}$, $\xi(P)$ is $(x_1 = x_2) \wedge (x_3 = x_4) \wedge (x_1 < x_3)$.

Now we can see that $\phi = \phi_r \vee \phi_c$ is equivalent to

$$\phi'(x_1, \ldots, x_n) \quad = \quad \phi_r(x_1, \ldots, x_n) \vee \bigvee_{P \in Part_{\prec}(V)} (\xi(P) \wedge \phi_c(x_1, \ldots, x_n)).$$

We now need the following key claim that will allow us to eliminate constraints from ϕ'.

Claim 1 *Let $\mathcal{X}$ be a nonempty subset of $Part_{\prec}(V)$, and let $\psi(x_1, \ldots, x_n)$ be $\psi_r(x_1, \ldots, x_n) \vee \bigvee_{P \in \mathcal{X}} (\xi(P) \wedge \psi_c(x_1, \ldots, x_n))$, where ψ_r is a simple formula in $L(SC, <)$, and ψ_c is a simple formula in $L(SC, \Theta)$ such that ψ_c is in disjunctive normal form. Then, for any infinite set $Y \subseteq \mathbb{U}$, and any $P_0 \in \mathcal{X}$, we can find an infinite subset $Y(P_0) \subseteq Y$ and a simple $L(SC, <)$ formula $\psi_r[P_0](x_1, \ldots, x_n)$ such that, for any $D \in Inst(SC, Y(P_0))$ and $\vec{c} = (c_1, \ldots, c_n) \in Y(P_0)^n$,*

$$D \models \psi_r(\vec{c}) \vee \bigvee_{P \in \mathcal{X}} (\xi(P)[\vec{c}/\vec{x}] \wedge \psi_c(\vec{c})) \quad \textit{iff} \quad D \models \psi_r(\vec{c}) \vee \psi_r[P_0](\vec{c}) \vee \bigvee_{P \in \mathcal{X} - \{P_0\}} (\xi(P)[\vec{c}/\vec{x}] \wedge \psi_c(\vec{c}))$$

where $\xi(P)[\vec{c}/\vec{x}]$ denotes $\xi(P)$ in which c_i is substituted for x_i, $i = 1, \ldots, n$.

Proof of claim 1. Let P_0 be $X_1 \prec \ldots \prec X_s$. Let $y_1, \ldots, y_s$ be representatives of $X_1, \ldots, X_s$. Let $ord(y_1, \ldots, y_s)$ be the conjunction of all $y_i < y_j$ for $1 \leq i < j \leq s$. Let $\psi'_c(\vec{y})$ be the formula obtained from ψ_c by replacing each x_i with y_j, if $x_i \in X_j$. Then, for any $D \in Inst(SC, \mathbb{U})$ and any $\vec{c} \in \mathbb{U}^n$,

$$D \models (\xi(P_0) \wedge ord(y_1, \ldots, y_s) \wedge \psi'_c(\vec{y}))[\vec{c}/\vec{x}] \quad \text{iff} \quad D \models (\xi(P_0) \wedge \psi_c(\vec{x}))[\vec{c}/\vec{x}].$$

Now consider $\psi'_c(\vec{y})$ (which is a disjunctive normal form formula). Let $C_1, \ldots, C_p$ be all the simple constraint formulae in it. Define a partition of s-element ordered tuples of Y into 2^p classes as follows. For every $\vec{\beta} = (\beta_1, \ldots, \beta_p) \in \{0, 1\}^p$, we put an ordered tuple $c_1 < \ldots < c_s$ into the class $Y_{\vec{\beta}}$ if the truth value of $C_i(c_1, \ldots, c_s)$ is β_i for every $i = 1, \ldots, p$. By Ramsey's theorem [11], we can find an infinite $Y(P_0) \subseteq Y$ such that there is a class $Y_{\vec{\beta}}$ such that all s-element ordered tuples of elements of $Y(P_0)$ are in $Y_{\vec{\beta}}$.

Let $\psi_c''(\vec{y})$ be the formula obtain from ψ_c' by replacing each C_i with $\mathbf{T}$ if $\beta_i = 1$ and $\mathbf{F}$ if $\beta_i = 0$. Note that ψ_c'' is a simple $L(SC, <)$ formula. Now it follows from the above that for every D and every $\vec{c} = (c_1, \ldots, c_s) \in Y(P_0)^s$ such that $c_1 < \ldots < c_s$,

$$D \models \psi_c'(\vec{c}) \quad \text{iff} \quad D \models \psi_c''(\vec{c}).$$

Therefore, for any $\vec{c} \in Y(P_0)^n$, $D \models (ord(y_1, \ldots, y_s) \wedge \psi_c'(\vec{c}))[\vec{c}/\vec{x}]$ if and only if $D \models (ord(y_1, \ldots, y_s) \wedge \psi_c''(\vec{y}))[\vec{c}/\vec{x}]$. This shows that we can take $\xi(P_0) \wedge ord(y_1, \ldots, y_s) \wedge \psi_c''(y_1, \ldots, y_s)$ as $\psi_r[P_0]$. Since this is a $L(SC, <)$ formula, the proof of claim 1 is now complete.

Now we show how to use the claim to conclude the proof of lemma 3. We enumerate $Part_{\prec}(V)$ as $P_1, \ldots, P_m$ where m is the cardinality of $Part_{\prec}(V)$. Starting with ϕ', we let $\phi_0 = \phi_r$ and $Y_0 = \mathbb{U}$, and repeatedly apply the claim obtaining infinite subsets of $\mathbb{U}$ and $L(SC, <)$ formulae as follows:

1. $\phi_1 = \phi_0 \vee \phi_0[P_1]$ and $Y_1 = Y_0(P_1)$;
2. $\phi_2 = \phi_1 \vee \phi_1[P_2]$ and $Y_2 = Y_1(P_2)$;

   ...

m. $\phi_m = \phi_{m-1} \vee \phi_{m-1}[P_m]$ and $Y_m = Y_{m-1}(P_m)$.

Now it follows from claim 1 that for every $D \in Inst(SC, Y_m)$, and any $c_1, \ldots, c_n \in Y_m$,

$$D \models \phi'(c_1, \ldots, c_n) \quad \text{iff} \quad D \models \phi_m(c_1, \ldots, c_n).$$

Since $D \models \phi(c_1, \ldots, c_n) \leftrightarrow \phi'(c_1, \ldots, c_n)$, we conclude the proof of lemma 3 by taking X to be Y_m and ψ to be ϕ_m.

Proof of lemma 4. We show the following: let $\chi_1(x, \vec{y})$ and $\chi_2(x, \vec{y})$ be quantifier-free formulae such that for any $\vec{c} \subset X$ and $c_0 \in X$ and any $D \in Inst(SC, X)$, $D \models \chi_1(c_0, \vec{c})$ iff $D \models \chi_2(c_0, \vec{c})$. Let $Y \subseteq X$. Then $D \models_Y \exists x.\chi_1(x, \vec{c})$ iff $D \models_Y \exists x.\chi_2(x, \vec{c})$. Indeed, we have $D \models_Y \exists x.\chi_1(x, \vec{c})$ iff there is $c_0 \in Y \cup adom(D) \subseteq X$ such that $D \models \chi_1(c_0, \vec{c})$ iff $D \models \chi_2(c_0, \vec{c})$ iff $D \models_Y \exists x.\chi_2(x, \vec{c})$. Similarly we obtain that $D \models_Y \forall x.\chi_1(x, \vec{c})$ iff $D \models_Y \forall x.\chi_2(x, \vec{c})$. Now applying this inductively to $\phi(x_1, \ldots, x_n)$ and $\psi(x_1, \ldots, x_n)$ proves lemma 4.

This concludes the proof of the theorem. $\square$

Proof of theorem 2

Let Ω_a be the (infinite) signature that consists of all mutivariate functions on the reals that arise as restrictions of analytic functions to the real arguments. By lemma 1, it is enough to prove $\mathcal{FO}(\mathbb{R}, \Omega_a) \Rrightarrow \mathcal{FO}(\mathbb{R}, \emptyset)$. (See the definition of $\Rrightarrow$ in the proof of theorem 1.)

Consider a $L(SC, \Omega_a)$ sentence Φ. By lemma 2, we can assume without loss of generality that Φ is of the form $\mathrm{Q}_1 x_1 \ldots \mathrm{Q} x_n.\phi(x_1, \ldots, x_n)$, where ϕ is in disjunctive normal form, and its atomic subformulae are either of the form $R_i(z_1, \ldots, z_{\tau_i})$, where all z's are variables, or $t_1(\vec{z}) = t_2(\vec{y})$, where t_1, t_2 are Ω_a-terms and $\vec{y}, \vec{z}$ only contain variables, or negations of the above.

Now we need the following lemma. For countable X this statement was proved in [4]; here we need to generalize it to uncountable subsets of the reals.

Lemma 5 *Let $\mathcal{F} = \{f_1(\vec{x}), \ldots, f_m(\vec{x})\}$ be a finite family of functions from Ω_a, where $\vec{x} = (x_1, \ldots, x_l)$. Assume that none of the functions in $\mathcal{F}$ is identically zero. Then it is possible to find a set $X \subseteq \mathbb{R}$ of cardinality of the continuum such that for any tuple $\vec{c}$ of l distinct elements of X, none of $f_i(\vec{c})$ equals zero.*

Proof of lemma 5. We need some notation first. Let f be a function from Ω_a with $x_{i_1}, \ldots, x_{i_k}$ some of its parameters. Let $c_{i_1}, \ldots, c_{i_k} \in \mathbb{R}$. Then $f[c_{i_1}/x_{i_1}, \ldots, c_{i_k}/x_{i_k}]$ denotes the function obtained from f by substituting c_{i_j} for x_{i_j}; this function has k fewer parameters than f. Since composition of analytic functions is analytic again, this function is still an Ω_a function.

Let $A \subseteq \mathbb{R}$. We say that A is $\mathcal{F}$-nice if A is infinite and for any tuple $\vec{c}$ of l distinct elements of A, none of $f_i(\vec{c})$ equals zero. It is known from [4] that there exists a $\mathcal{F}$-nice set. To recall why, note that if $X \subset \mathbb{R}$ is countable and $\mathcal{H}$ is a countable collection of Ω_a functions with at most l parameters $x_1, \ldots, x_l$, and none of the functions in $\mathcal{H}$ is identically zero, then there exists $c \in \mathbb{R} - X$ such that none of the functions of form $h[c/x_i]$, $h \in \mathcal{H}$, is identically zero. Indeed, for any function $h(x_{i_1}, \ldots, x_{i_{k+1}})$, one can find a vector $(c_1, \ldots, c_k)$ such that $h'(x_{i_{k+1}}) = h[c_1/x_{i_1}, \ldots, c_k/x_{i_k}]$ is not identically zero. Since h' is still analytic, its set of zeros $Z(h, k+1)$ is at most countable. Similarly define sets $Z(h, i)$ for every $1 \leq i \leq l$, assuming such set to be empty if x_i is not a parameter of h. Now the set $Z = X \cup \bigcup_{h \in \mathcal{H}, 1 \leq i \leq l} Z(h, i)$ is countable, and thus any $c \in \mathbb{R} - Z$ proves our observation. This observation gives us a simple inductive construction that leads to a countable $\mathcal{F}$-nice set.

It is easy to see that the union of a chain of $\mathcal{F}$-nice sets is $\mathcal{F}$-nice again; thus there exists a maximal $\mathcal{F}$-nice set by Zorn's lemma. Let A be a maximal $\mathcal{F}$-nice set. To conclude the proof of the lemma, we will show that A has the cardinality of the continuum.

Let κ be the cardinality of A; assume $\kappa < c$. Consider the set $\mathcal{G}$ of all functions of the form $g(x_i) = f[c_1/x_1, \ldots, c_l/x_l]$ where $f \in \mathcal{F}$ and substitution occurs in all positions except the ith one, and all c_js are distinct elements of A. Then the cardinality of $\mathcal{G}$ is still κ.

Furthermore, each function in $\mathcal{G}$ is analytic, and not identically zero, since no f from $\mathcal{F}$ is zero on any vector of distinct elements from A. Let Z_g be the the set of zeros of $g \in \mathcal{G}$. Since each $g \in \mathcal{G}$ is analytic, Z_g is at most countable. Furthermore, the cardinality of $Z = \bigcup_{g \in \mathcal{G}} Z_g$ is at most κ. Hence, there exists $a \in \mathbb{R} - (A \cup Z)$. But now it follows from our construction that $A \cup \{a\}$ is $\mathcal{F}$-nice, thus contradicting the maximality of A. This shows that A has the cardinality of the continuum and completes the proof of lemma 5.

The proof of theorem 2 now follows from the lemma below and lemmas 2 and 4 in exactly the same way as proved theorem 1.

Lemma 6 *Let $\phi(x_1, \ldots, x_n)$ be a simple formula in $L(SC, \Omega_a)$. Then there exists an uncountable set $X \subseteq \mathbb{R}$ and a $L(SC)$ formula $\psi(x_1, \ldots, x_n)$ such that for any $D \in Inst(SC, X)$ and any $c_1, \ldots, c_n \in X$,*

$$D \models \phi(c_1, \ldots, c_n) \quad \textit{iff} \quad D \models \psi(c_1, \ldots, c_n).$$

Proof of lemma 6. Let $V = \{x_1, \ldots, x_n\}$. Let $Part(V)$ be the familty of all partitions on V. We let $\chi(P)$ be a formula in the language of equality that represents the partition $P \in Part(V)$; for example, $(x_1 = x_2) \wedge (x_3 = x_4) \wedge \neg(x_1 = x_2)$ represents the partition $\{\{x_1, x_2\}, \{x_3, x_4\}\}$.

Rewrite ϕ to an equivalent formula

$$\phi'(x_1, \ldots, x_n) \quad = \quad \bigvee_{P \in Part(V)} (\chi(P) \wedge \phi_P)$$

where ϕ_P is obtained from ϕ by choosing one representative of each block of the partition P and replacing each x_i by the representative of the block x_i is in. Next, we look at all atomic subformulae $t_1(\vec{z}) = t_2(\vec{y})$, and for those that are not identically true, define a new function $t(\vec{z}, \vec{y}) = t_1(\vec{z}) - t_2(\vec{y})$. Let $\mathcal{F}$ be the collection of all functions obtained in this way. According to lemma 5, we can find an uncountable set X such that assigning distinct values from X to the variables from V simultaneously invalidates all equations of form $f(\vec{x}) = 0$, $f \in \mathcal{F}$. Let ϕ'_P be the formula obtained from ϕ_P by replacing each subformula $t_1(\vec{z}) = t_2(\vec{y})$ by $\mathbf{T}$ if it represents an identically true condition, and by $\mathbf{F}$ otherwise. Now it follows from lemma 5 that for every $D \in \mathit{Inst}(SC, X)$ and every tuple $\vec{c}$ of s distinct elements of X, where P has s blocks, $D \models \phi_P(\vec{c})$ iff $D \models \phi'_P(\vec{c})$. Hence, for every n-vector $\vec{c}$ of elements of X,

$$D \models (\chi(P) \wedge \phi_P)[\vec{c}/\vec{x}] \quad \text{iff} \quad D \models (\chi(P) \wedge \phi'_P)[\vec{c}/\vec{x}].$$

Thus, ψ can be taken to be $\bigvee_{P \in Part(V)} (\chi(P) \wedge \phi'_P)$. This completes the proof of lemma 6 and the theorem. □

Proof of theorem 3

Define $SC_{\vec{P}}$ as $SC \cup \langle P_1, \ldots, P_m \rangle$. Note that in the course of proving theorem 1 we actually established the following: for any $L(SC, \Omega)$ formula $\phi(x_1, \ldots, x_n)$, we can find a $L(SC, <)$ formula $\psi(x_1, \ldots, x_n)$ and an infinite set $X \subseteq \mathbb{U}$ such that for every $\vec{c} \in X^n$ and every $D \in \mathit{Inst}(SC, X)$, $D \models \phi(\vec{c})$ iff $D \models \psi(\vec{c})$. Exactly the same proof will give us a similar result that constructs a $L(SC_{\vec{P}}, <)$ formula from a $L(SC_{\vec{P}}, \Omega)$ formula, and similarly a $L(SC_{\vec{P}})$ formula for the case of an analytic signature on the reals. Thus, to prove part 1 of the theorem, we need an analog of lemma 4 for adding both first-order and second-order quantifiers. Such a result is proved by a straightforward induction, thus giving us part 1 of the theorem.

To prove part 2, let P be a binary predicate and let $\mathrm{ORD}(P)$ be the first-order sentence expressing the fact that P defines a linear ordering. Suppose that we are given a second-order sentence $\Phi = \mathrm{Q}'_1 P_1 \ldots \mathrm{Q}'_m P_m \mathrm{Q}_1 x_1 \ldots \mathrm{Q}_n x_n . \phi(P_1, \ldots, P_m, x_1, \ldots, x_n)$. We can rewrite the proof of theorem 1 using $P(x, y)$ insetad of $x < y$ every time we introduce the order relation into the formula. Then it is not hard to see that we can find an infinite set X and a $L(SC_{\vec{P}}, P)$ formula $\psi(x_1, \ldots, x_n)$ such that $D \models \phi(\vec{c})$ iff $D \models \psi(\vec{c})$ for every $D \in \mathit{Inst}(SC, X)$ and $\vec{c} \in X^n$, provided that P defines a linear ordering on X. That is, $D \models \mathrm{ORD}(P) \wedge \phi(\vec{c})$ iff $D \models \mathrm{ORD}(P) \wedge \psi(\vec{c})$.

Let $\Psi(P)$ be $\mathrm{ORD}(P) \wedge \psi(x_1, \ldots, x_n)$ preceded by the quantifier prefix (both first- and second-order) of Φ, and similarly let $\Phi(P)$ be $\mathrm{ORD}(P) \wedge \phi(x_1, \ldots, x_n)$ preceded by the quantifier prefix of Φ. Define $\Phi' = \exists P . \Phi(P)$ and $\Psi' = \exists P . \Psi(P)$. By the analog of lemma 4 for both first- and second-order quantifiers, we get $D \models_X \Phi'$ iff $D \models_X \Psi'$ for any $D \in \mathit{Inst}(SC, X)$. Since Φ does not mention P and a linear order can be defined on any set, we have $D \models_X \Phi'$ iff $D \models_X \Phi$. Thus, $D \models_X \Phi$ iff $D \models_X \Psi'$, and Ψ' does not mention any symbol in Ω, nor the order relation $\leq$. Since F is an orderable fragment, Ψ' belongs to that fragment. The proof is now complete. □

Proof of theorem 4

As in [4], we will give here a nonconstructive proof using the techniques of nonstandard universes. A more detailed analysis, available in the full paper, yields a constructive proof. We start with the definitions; see [8] for more information.

For any set S, the *superstructure* $V(S)$ over S is defined as $V(S) = \bigcup_{n<\omega} V_n(S)$ where $V_1(S) = S$, and $V_{n+1}(S) = V_n(S) \cup \{X \mid X \subset V_n(S)\}$. We will work with the structure $\langle V(S), \in\rangle$ considered as a structure for the first-order language for the epsilon relation. A *bounded-quantifier formula* in this language is a formula built up from atomic formulas by the logical connectives and the quantification: $\forall X \in Y$, $\exists X \in Y$, where X and Y are variables.

A *nonstandard universe* consists of a pair of superstructures $V(S)$ and $V(Y)$ and a mapping $* : V(S) \to V(Y)$ which is the identity on S, such that $Y = {}^*S$ and the following holds:

(Transfer Principle) For any bounded quantifier formula $\phi(v_1, \ldots, v_n)$ and any list $a_1, \ldots, a_n$ of elements from $V(S)$, $\phi(a_1, \ldots, a_n)$ is true in $V(S)$ if and only if $\phi({}^*a_1, \ldots, {}^*a_n)$ is true in $V(Y)$.

Elements of $V(Y)$ in the image of the $*$-map are called *standard*. Element of $V(Y)$ in the downward transitive closure of the standard sets under $\in$ are called *internal*; elements that are not internal are called *external*. A map is called internal if its graph is an internal set. We also assume the following

(Countable Saturation Principle) For every standard A, and every countable collection $\Sigma(x, \vec{v})$ of bounded-quantifier formulas, and for every vector $\vec{c}$ of internal sets, if every finite subset of $\Sigma(x, \vec{c}/\vec{v})$ is satisfied in $V(Y)$ by an element of A, then $\Sigma(x, \vec{c}/\vec{v})$ is satisfied by an element of A.

Finally, we work in a nonstandard universe that satisfies the following axiom:

(Isomorphism Property) For any finite first-order language L, and any two L-structures M_1 and M_2 that are internal, if M_1 and M_2 are L-elementary equivalent (i.e., agree on all standard sentences of L), then there is an (not necessarily internal) L-isomorphism between M_1 and M_2.

For basic facts about the isomorphism property, and a proof that saturated models with the isomorphism property exist, the reader is referred to [15].

We omit the $*$ when convenient: for example, if $<$ is an ordering on S, x_1 and x_2 are elements of ${}^*S$, then we will write $x_1 < x_2$ rather than $x_1 {}^*\!< x_2$. Similarly, if q is a query on schema SC, and D is a *-database (an element of the set ${}^*DB$, where DB is the set of databases for SC), then we will refer to $q(D)$ rather than ${}^*q(D)$.

Let $\mathcal{M} = \langle \mathbb{U}, \Omega\rangle$. For the proof we take S to be the disjoint union of $\mathbb{U}$ and $\mathbb{N}$. We need the following proposition from [4].

Proposition 3 *Let L be a language for which each symbol is internal, and let M be an L-structure such that the domain of M is internal and the interpretation of each symbol in L is internal. Let $\phi(\vec{x})$ be a formula of L that has standard finite cardinality (i.e. number of symbols). Then the internal satisfaction predicate ${}^*\!\models$ agrees with the external satifaction predicate $\models$ on ϕ. That is, if $\vec{c}$ is a finite sequence of parameters from M, then $M\, {}^*\!\models \phi(\vec{c})$ iff $M \models \phi(\vec{c})$.*

Given this proposition, we will not distinguish the two kinds of satisfaction predicates dealing with finitary ϕ's.

A careful look at the constructions used in propositions 10, 11 and 12 of [4] show that the following is proved.

Lemma 7 *If $\mathcal{M}$ is an o-minimal structure, D and D' are two hyperfinite database instances over $\mathcal{M}$, and f is a mapping from $adom(D)$ onto $adom(D')$ that preserves the schema relations and also has the property that for every formula $\phi(\vec{x})$ in the language of $\mathcal{M}$, for each tuple $\vec{z}$ from $adom(D)$,*

$$^*\mathcal{M} \models \phi(\vec{z}) \Leftrightarrow {}^*\mathcal{M} \models \phi(f(\vec{z})).$$

Then D and D' satisfy all the same (natural semantics) queries over $^*\mathcal{M}$.

We will first show

Lemma 8 *If we have two hyperfinite instances A and A' that agree on every standard active semantics query, then they agree on every natural semantics query as well.*

Proof of Lemma. Fix a counterexample, that is, fix A and A', and a natural semantics query ϕ on which they disagree. Let $^*\mathcal{M}(A)$ be the expansion of $^*\mathcal{M}$ to $L(SC,\Omega)$ given by interpreting the schema relation sybols as in A, and let $^*\mathcal{M}(A')$ be likewise.

Let Ω' be the language containing (only) predicate symbols for each atomic formula of Ω. Notice that Ω', unlike Ω, is necessarily purely relational. Let $\mathcal{M}'$ be the model for Ω' with domain equal to $\mathbb{U}$, and with the predicates of Ω' interpreted in the obvious way. It is easy to see that $\mathcal{M}'$ also admits elimination of quantifiers. Consider $\Omega'(A)$ and $\Omega'(A')$ as structures for $L(SC,\Omega')$ in which the domains are the active domains of A and A', respectively, the schema relations are unchanged, and each predicate of Ω' is interpreted as the * of the corresponding definable subset of $\mathcal{M}$. Then we have

Claim 2 *$\Omega'(A)$ and $\Omega'(A')$ are elementary equivalent in $L(SC,\Omega')$*

Proof of claim 2. Every active domain query in $L(SC,\Omega)$ is equivalent to $Q_1 \ldots Q_n. \bigvee_i \bigwedge_j A_{ij}$, where the Q_i's are quantifiers, and each A_{ij} is either of the form 1) $R(\vec{x})$ for R in SC and $\vec{x}$ variables, or 2) $P(\tau_1(\vec{x}),\ldots,\tau_n(\vec{x}))$, where P is either an Ω predicate or equality, and the τ_is are Ω terms, or 3) the negation of 1) or 2).

For any $L(SC,\Omega)$ formula $\rho(\vec{x})$ in normal form as above, let $\Gamma[\rho](\vec{x})$ be the $L(SC,\Omega')$ formula obtained by converting each atomic formula of the form 2) to the corresponding formula of $L(SC,\Omega')$, and similarly with negations of 2). That is, we convert an atomic subformula of the form $P(\tau_1(\vec{x}),\ldots,\tau_n(\vec{x}))$ to one of the form $S(\vec{x})$, where S is one of the new symbols of Ω'. Note that the free variables of ρ and of $\Gamma[\rho]$ are the same. We can prove by induction that for D equal to either A or A', any $L(SC,\Omega)$ formula $\phi(\vec{x})$ and any vector $\vec{c}$ in the active domain of D, we get

$$D \models_{\emptyset} \phi(\vec{c}) \quad \text{if and only if} \quad \Omega'(D) \models \Gamma[\phi](\vec{c}).$$

We also note that the map Γ maps onto the set of sentences of $L(SC,\Omega')$. Indeed, Γ has an inverse defined on $L(SC,\Omega')$ by converting every atomic subfomula of the form $P(\vec{x})$, where P is from Ω', to the corresponding atomic formula of Ω. We then see that Γ is an isomorphism of the Lindenbaum algebra of active-semantics formulae of $L(SC,\Omega)$ to the algebra of natural semantics formulae of $L(SC,\Omega')$. The claim follows from this.

We apply the Isomorphism Principle to $\Omega'(A)$ and $\Omega'(A')$, to get a mapping f from the active domain of A onto active domain of A' that preserves schema relations and the predicates of Ω' (i.e. everything in $L(SC, \Omega')$). Since our model $\mathcal{M}'$ has elimination of quantifiers, we get that for each $\phi(\vec{x})$ in Ω' and $\vec{c}$ in the active domain of A,

$$^*\mathcal{M}' \models \phi(\vec{c}) \quad \text{if and only if} \quad ^*\mathcal{M}' \models \phi(f(\vec{c})),$$

since $\phi(\vec{c})$ is equivalent to a boolean combination of atomic formulae, each of which will be preserved by f.

Now Lemma 7 shows that $^*\mathcal{M}'(A)$ and $^*\mathcal{M}'(A')$ (which are defined analagously to $^*\mathcal{M}(A)$ and $^*\mathcal{M}(A')$ but for the language Ω' instead of Ω) satisfy all the same sentences of $L(SC, \Omega')$. We now finish the proof of lemma 8 with the following claim.

Claim 3 *$^*\mathcal{M}(A)$ and $^*\mathcal{M}(A')$ satisfy all the same sentences of $L(SC, \Omega)$.*

Proof of claim 3. The translation Γ from $L(SC, \Omega)$ to $L(SC, \Omega')$ given in the proof of claim 2, also has the property that for D equal to either A or A', any $L(SC, \Omega)$ formula $\phi(\vec{x})$ and any vector $\vec{c}$ from $\mathbb{U}$, we get

$$^*\mathcal{M}(D) \models \phi(\vec{c}) \quad \text{if and only if} \quad ^*\mathcal{M}'(D) \models \Gamma[\phi](\vec{c}).$$

This can be proved by simple induction again, using the definition of Ω' in the base step.

From this, we get that for any sentence ψ in $L(SC, \Omega)$,

$$^*\mathcal{M}(A) \models \psi \;\Leftrightarrow\; ^*\mathcal{M}'(A) \models \Gamma[\psi] \;\Leftrightarrow\; ^*\mathcal{M}'(A') \models \Gamma[\psi] \;\Leftrightarrow\; ^*\mathcal{M}(A') \models \psi,$$

and this gives us the proof of the claim.

Now claim 3 gives us a contradiction, which proves the lemma.

Finally, we show that lemma 8 implies the theorem. Suppose there were a counterexample q to the theorem (that is, q is definable as a natural-semantiics query, but not as an active-semantics query). We first note that for every finite collection F of active semantics queries, there must be two finite instances A_F and A'_F that agree on all queries in F but disagree on q . If the previous sentence were not the case, then the set of instances satisfying q would be composed of equivalence classes of the equivalence relation given by

$$A \equiv_F B \quad \Leftrightarrow \quad \bigwedge_{\phi \in F} A \models \phi \leftrightarrow B \models \phi.$$

Since each equivalence class of $\equiv_F$ is definable by a boolean combination of queries in F, this would make q definable by a boolean combination of queries in F, contradicting the fact that q is not definable by an active-semantics query.

Now by applying saturation, we would get two $*$-finite instances A and A' in the nonstandard universe that agree on every standard active-semantics query, but disagree on q, contradicting lemma 8. This completes the proof of the theorem. □

Proof of proposition 1

Assume $\mathcal{FO}(\mathbb{U}, <) \Rightarrow \mathcal{FO}^{\emptyset}(\mathbb{U}, <)$. Let the schema consist of one unary predicate R and consider the following $L(R, <)$ sentence:

$$\psi \;=\; \exists x.(\neg R(x) \;\wedge\; \exists y \exists z.(R(y) \;\wedge\; R(z) \;\wedge\; y < x \;\wedge\; x < z))$$

Then we can find an infinite set $X \subseteq \mathbb{U}$ and a $L(R,<)$ sentence ϕ such that $D \models_X \psi$ iff $D \models_\emptyset \phi$ for any $D \in \mathit{Inst}(SC, X)$. Note that $D \models_X \psi$ iff there are two elements of $adom(D)$ such that there exists an element of X between them that is not in $adom(D)$. Since ϕ is a $L(SC,<)$ sentence, then for every $D \in \mathit{Inst}(SC, X)$ and every monotone injective partial map f on $adom(D)$ such that $f(adom(D)) \subset X$, $D \models_\emptyset \phi$ iff $f(D) \models_\emptyset \phi$.

Since X is an infinite subset of a scattered ordering, we can find two finite subsets $Y = \{y_1 < \ldots < y_m\}$ and $Z = \{z_1 < \ldots < z_m\}$, $m \geq 2$, such that y_{i+1} is the successor of y_i in X, and between any z_i and z_{i+1} there is another element of X, $i = 1, \ldots, m-1$. Let $f_{Y \to Z}$ be a partial map on X that sends each y_i to z_i, $i = 1, \ldots, m$. Define $D_Y \in \mathit{Inst}(SC, X)$ by interpreting R as Y and $D_Z \in \mathit{Inst}(SC, X)$ by interpreting R as Z. Then

$$D_Y \models_X \psi \ \text{ iff } \ D_Y \models_\emptyset \phi \ \text{ iff } \ f_{Y \to Z}(D_Y) \models_\emptyset \phi \ \text{ iff } \ D_Z \models_\emptyset \phi \ \text{ iff } \ D_Z \models_X \psi.$$

However, $D_Z \models \psi$ and $D_Y \models \neg\psi$. This contradiction proves the proposition. □

Proof of proposition 2

The set of natural numbers can be defined by a second-order formula with one unary existential second order quantifier in the language of $+, 0, <$:

$$\phi(n) \;=\; \exists P.[P(0) \wedge (\forall x.P(x) \to (x = 0 \vee P(x-1))) \wedge (\forall x \forall y.P(x) \wedge x < y < x+1 \to \neg P(y))] \;\wedge\; P(n).$$

Then it follows from [13] that any total recursive query on databases whose active domain consists only of natural numbers is in $\mathcal{SO}_F(\mathbb{R}, +, *, 0, 1, <)$. On the other hand, every query in $\mathcal{SO}^\emptyset(\mathbb{R}, +, *, 0, 1, <)$ (and thus in $\mathcal{SO}_F^\emptyset(\mathbb{R}, +, *, 0, 1, <)$) has *PSPACE* data complexity, which proves the proposition. □

Proof of theorem 5

Let $\phi(x_1, \ldots, x_n)$ be a $L(SC_1, \Omega)$ formula defining a nonboolean query for some n-ary relational symbol in SC_2. We now extend SC_1 to SC' by n unary predicate symbols $S_1, \ldots, S_n$ which are not present in $SC_1 \cup SC_2$. Define the following $L(SC', \Omega)$ sentence:

$$\begin{array}{rcl} \Phi & = & \bigwedge_{i=1}^n ((\exists x.S_i(x)) \;\wedge\; (\forall x \forall y.(S_i(x) \wedge S_i(y) \to x = y))) \\ & \wedge & (\forall x_1 \ldots \forall x_n.(S_1(x_1) \wedge \ldots \wedge S_n(x_n)) \to \phi(x_1, \ldots, x_n)) \end{array}$$

It says that S_is are interpreted as singletons, and ϕ is satisfied on the elements of S_is.

Assume that $\mathcal{FO}(\mathbb{U}, SC, \Omega) \Rightarrow \mathcal{FO}(\mathbb{U}, SC, \Theta)$ for any schema SC. Applying this to SC', we get a $L(SC', \Theta)$ sentence Ψ and an infinite set X such that for every $D \in \mathit{Inst}(SC', X)$, $D \models_X \Phi$ iff $D \models_X \Psi$. Let $\psi(x_1, \ldots, x_n)$ be a $L(SC_1, \Theta)$ formula obtained from Ψ be replacing each $S_i(z)$ with $z = x_i$ (we assume here that Ψ does not mention any of the x_is).

To prove $\mathcal{FO}_{SC_2}(\mathbb{U}, SC_1, \Omega) \Rightarrow \mathcal{FO}_{SC_2}(\mathbb{U}, SC_1, \Theta)$, it will be enough to show that $D \models_X \phi(\vec{a})$ iff $D \models_X \psi(\vec{a})$ for every $D \in \mathit{Inst}(SC_1, X)$ and every $\vec{a} \in (adom(D) \cup X)^n$. For every $D \in \mathit{Inst}(SC_1, X)$ and $\vec{a} = (a_1, \ldots, a_n)$, define $D(\vec{a}) \in \mathit{Inst}(SC', X)$ by interpreting each S_i as $\{a_i\}$.

From the definition of Φ it follows that $D \models_X \phi(\vec{a})$ iff $D(\vec{a}) \models_X \Phi$. We know that $D(\vec{a}) \models_X \Phi$ iff $D(\vec{a}) \models_X \Psi$. From the definition of ψ, $D(\vec{a}) \models_X \Psi$ iff $D(\vec{a}) \models_X \psi(\vec{a})$. Finally, since ψ does not mention any S_is and $\vec{a} \in (adom(D) \cup X)^n$ and the range of quantification is the same on both sides, we have $D(\vec{a}) \models_X \psi(\vec{a})$ iff $D \models_X \psi(\vec{a})$. Thus, $D \models_X \phi(\vec{a})$ iff $D \models_X \psi(\vec{a})$, proving the first statement of the theorem.

To prove the second statement of the theorem, we start again with a $L(SC_1, \Omega)$ formula $\phi(x_1, \ldots, x_n)$ and define SC' and Φ as above. Then we find a $L(SC', \Theta)$ sentence Ψ and an infinite X such that $D \models_{A \cap X} \Phi$ iff $D \models_{B \cap X} \Psi$ for every $D \in \mathit{Inst}(SC', X)$. Define $\psi(x_1, \ldots, x_n)$ as before. Since we are now interested in domain-preserving queries only, to conclude the proof we must show that for any $D \in \mathit{Inst}(SC_1, X)$ and any $\vec{a} \in \mathit{adom}(D)^n$, $D \models_{A \cap X} \phi(\vec{a})$ iff $D \models_{B \cap X} \psi(\vec{a})$.

The proof proceeds by establishing the same chain of equivalences as above. First, $D \models_{A \cap X} \phi(\vec{a})$ iff $D(\vec{a}) \models_{A \cap X} \Phi$ since $\vec{a} \in \mathit{adom}(D)^n$ and does not extend the range of quantification. Next, $D(\vec{a}) \models_{A \cap X} \Phi$ iff $D(\vec{a}) \models_{B \cap X} \Psi$ iff $D(\vec{a}) \models_{B \cap X} \psi(\vec{a})$. Finally, since ψ does not mention any of the S_is and $\mathit{adom}(D) = \mathit{adom}(D(\vec{a}))$ and thus the range of quantification is the same on both sides, $D(\vec{a}) \models_{B \cap X} \psi(\vec{a})$ iff $D \models_{B \cap X} \psi(\vec{a})$. Thus, $D \models_{A \cap X} \phi(\vec{a})$ iff $D \models_{B \cap X} \psi(\vec{a})$, which completes the proof. □

Proof of theorem 6

As in the proof of theorem 5, it is enough to consider one relational symbol in SC_2. Assume that it is n-ary, and a query is given by a $L(SC_1, \Omega)$ formula $\phi(x_1, \ldots, x_n)$. As in the proof of theorem 5, extend SC_1 to SC' by unary predicates $S_1, \ldots, S_n$, and define the $L(SC', \Omega)$ sentence Φ as before. Using the assumption of the theorem, we get a $L(SC', \Omega)$ sentence Ψ such that $D \models \Phi$ iff $D \models_\emptyset \Psi$ for any $D \in \mathit{Inst}(SC', \mathbb{U})$.

Next, we define a translation of $L(SC', \Omega)$ formulae into $L(SC_1, \Omega)$ formulae as below. We assume that the variables $x_1, \ldots, x_n$ are not used by the input formula; R_is are SC_1 predicates and t_is are Ω-terms; $\vec{y}_i$ is a vector of variables, P is an Ω-predicate.

$$\begin{array}{rcl}
R_i(t_1(\vec{y}_1), \ldots, t_m(\vec{y}_m))^\circ & = & R_i(t_1(\vec{y}_1), \ldots, t_m(\vec{y}_m)) \\
(t_1(\vec{y}_1) = t_2(\vec{y}_2))^\circ & = & t_1(\vec{y}_1) = t_2(\vec{y}_2) \\
P(t_1(\vec{y}_1), \ldots, t_m(\vec{y}_m))^\circ & = & P(t_1(\vec{y}_1), \ldots, t_m(\vec{y}_m)) \\
S_i(t(\vec{y}))^\circ & = & x_i = t(\vec{y}) \\
(\chi \vee \xi)^\circ & = & \chi^\circ \vee \xi^\circ \\
(\neg\chi)^\circ & = & \neg\chi^\circ \\
(\exists x.\chi(x, \vec{y}))^\circ & = & (\exists x.\chi^\circ(x, \vec{y})) \vee \chi^\circ(x_1, \vec{y}) \vee \ldots \vee \chi^\circ(x_n, \vec{y})
\end{array}$$

Note that for a formula $\chi(y_1, \ldots, y_m)$, the free variables of χ° are $y_1, \ldots, y_m, x_1, \ldots, x_n$. Thus, we shall write χ° as $\chi^\circ(y_1, \ldots, y_m, x_1, \ldots, x_n)$ or $\chi^\circ(\vec{y}, \vec{x})$ in this case.

Let $\chi(y_1, \ldots, y_m)$ be a $L(SC', \Omega)$ formula, and $D \in \mathit{Inst}(SC_1, \mathbb{U})$. We now prove that for $\chi^\circ(y_1, \ldots, y_m, x_1, \ldots, x_n)$, for any $\vec{c} \in \mathbb{U}^m$ and any $\vec{a} \in \mathbb{U}^n$, $D \models_\emptyset \chi^\circ(\vec{c}, \vec{a})$ iff $D(\vec{a}) \models_\emptyset \chi(\vec{c})$. Here, as before, $D(\vec{a})$ is a SC' instance in which S_i is interpreted as $\{a_i\}$.

We prove this by induction on the structure of χ. The cases of atomic formulae and logical connectives are immediate, so we only consider the existential quantification case. For any $L(SC', \Omega)$ formula $\chi(x, \vec{y})$, we have $D(\vec{a}) \models_\emptyset (\exists x.\chi(x, \vec{y}))[\vec{c}/\vec{y}]$ iff there exists $c_0 \in \mathit{adom}(D(\vec{a}))$ such that $D(\vec{a}) \models_\emptyset \chi(c_0, \vec{c})$. This happens iff either (1) $D(\vec{a}) \models_\emptyset \chi(c_0, \vec{c})$ for some $c_0 \in \mathit{adom}(D)$ or (2) $D(\vec{a}) \models_\emptyset \chi(a_i, \vec{c})$ for some a_i from $\vec{a}$. By the induction hypothesis, (1) is equivalent to $D \models_\emptyset \chi^\circ(c_0, \vec{c}, \vec{a})$ for some $c_0 \in \mathit{adom}(D)$; that is, $D \models_\emptyset \exists x.\chi^\circ(x, \vec{c}, \vec{a})$. Similarly, by the induction hypothesis, $D(\vec{a}) \models_\emptyset \chi(a_i, \vec{c})$ iff $D \models_\emptyset \chi^\circ(a_i, \vec{c}, \vec{a})$. Thus, $D(\vec{a}) \models_\emptyset (\exists x.\chi(x, \vec{y}))[\vec{c}/\vec{y}]$ iff $D \models_\emptyset (\exists x.\chi^\circ(x, \vec{c}, \vec{a})) \vee \bigvee_{i=1}^n \chi^\circ(a_i, \vec{c}, \vec{a})$ iff $D \models_\emptyset (\exists x.\chi(x, \vec{y}))^\circ[\vec{c}/\vec{y}, \vec{a}/\vec{x}]$, thus proving the case and our claim.

Now we define $\psi(x_1, \ldots, x_n)$ as Ψ°. To finish the proof, we must show that $D \models \phi(\vec{a})$ iff $D \models_\emptyset \psi(\vec{a})$ for any $D \in Inst(SC_1, \mathbb{U})$ and any $\vec{a} \in \mathbb{U}^n$. According to the property of the translation $(\cdot)^\circ$ we just proved, $D(\vec{a}) \models_\emptyset \Psi$ iff $D \models_\emptyset \psi(\vec{a})$. Thus, $D \models \phi(\vec{a})$ iff $D(\vec{a}) \models \Phi$ iff $D(\vec{a}) \models_\emptyset \Psi$ iff $D \models_\emptyset \psi(\vec{a})$, as required. The proof is complete. $\square$

Proof of theorem 7

Let RA_{poly} be relational algebra in which polynomial inequality constraints are allowed as selection predicates. For instance, $\sigma_{\#1>\#2^3+4}(R)$ is a RA_{poly} expression that, when applied to a set of pairs, selects those pairs (x, y) for which $x > y^3+4$. Now we show that every $\mathcal{FO}(\mathbb{R}, +, *, 0, 1, <)$ is equivalent to some RA_{poly} expression. That is, for each $L(SC, +, *, 0, 1, <)$ sentence ϕ, there is a RA_{poly} expression e_ϕ such that $D \models \phi$ if $e_\phi(D) = \{()\}$ (empty tuple) and $D \models \neg\phi$ if $e_\phi(D) = \{\}$.

To prove this, we first use theorem 4 and assume that ϕ is evaluated under the active semantics. Then we follow the classical proof of the equivalence of relational calculus and relational algebra, cf. [1]. As noticed in the proof of theorem 1, we can assume without loss of generality that all atomic subformulae of ϕ are of form $R_i(z_1, \ldots, z_{\tau_i})$, where z_is are variables, or $t_1(\vec{z})\{=,<\}t_2(\vec{y})$, where t_1, t_2 are terms, and $\vec{y}$, $\vec{z}$ only contain variables, or negations of the above. Thus, we only need to augment the proof of [1] by one case to account for $t_1(\vec{z})\{=,<\}t_2(\vec{y})$ atomic subformulae.

Let ADOM be a relational algebra expression such that for every D, $\mathrm{ADOM}(D)$ evaluates to $adom(D)$. Let $x_1, \ldots, x_n$ be all the variables of $\vec{y}$ and $\vec{z}$. Then $t_1(\vec{z})\theta t_2(\vec{y})$, where $\theta \in \{=,<\}$, is translated into

$$\sigma_{t_1(\vec{z})\theta t_2(\vec{y})}(\mathrm{ADOM}(D) \times \ldots \times \mathrm{ADOM}(D))$$

where the selection is applied to the product of n copies of $\mathrm{ADOM}(D)$. It is routine to verify that the proof of [1] goes through when it is augmented in such a way. Just to provide evidence for that, recall that the idea of translating atomic formulae not involving relational symbols is to construct all tuples of the elements from the active domain that satisfy certain condition. For instance, $x = y$ is translated as $\sigma_{\#1=\#2}(\mathrm{ADOM}(D) \times \mathrm{ADOM}(D))$. We are doing precisely that for polynomial constraints.

Now it remains to be proved that any RA_{poly} query has TC^0 data complexity. This proof proceeds exactly as the proof of AC^0 data complexity for relational algebra (see [1]) with one exception: every time the σ_p operator is encountered, we have to compute the condition p. If p is of form $t_1(\vec{y})\{=,<,\neq,\not<\}t_2(\vec{z})$ where t_1, t_2 are terms (that is, polynomials), we construct circuits that compute t_1 and t_2 first and then make the comparison. Unlike the equality and ineqaulity predicates, we need here circuits that are capable of computing polynomial functions. Since addition and multiplication are in TC^0 [3, 22], we can insert a threshold circuit that computes σ_p. The proof that all other operations can be computed by TC^0 circuits (in fact, by AC^0 circuits) is exactly the same as in [1]. The theorem is proved. $\square$

Proof of theorem 8, 1)

Let Ω contain one binary predicate symbol $\lhd$ to be interpreted as a well-ordering. We shall write $x \lhd y$ instead of $\lhd(x, y)$. Assume that $\mathcal{FO}(\mathbb{R}, \Omega) \Rightarrow_\kappa \mathcal{FO}(\mathbb{R}, <)$ for $\kappa > \aleph_0$.

We consider the schema $SC = \{R_1, R_2\}$ where both relational symbols are unary, and let

$$\Phi \;=\; \forall x \forall y.(R_1(x) \wedge R_2(y)) \rightarrow x \lhd y.$$

Then we can find an uncountable X, and $L(SC,<)$ sentence Ψ such that for every $D \in \mathit{Inst}(SC,X)$, $D \models_X \Phi$ iff $D \models_X \Psi$. We now transform both Φ and Ψ into formulae $\phi(z,v)$ and $\psi(z,v)$ by replacing $R_1(x)$ with $x = z$ and $R_2(x)$ with $x = v$. Now for any $c_1, c_2 \in X$ we have $c_1 \lhd c_2$ iff $D \models_X \phi(c_1, c_2)$ iff $D \models_X \psi(c_1, c_2)$. Since ψ does not mention relational symbols, it is a formula in the language of $<$ that defines well-ordering on the model $(X,<)$. In other words, for any $c_1, c_2 \in X$, $(X,<) \models \psi(c_1, c_2)$ iff $c_1 \lhd c_2$.

Now we will show that no such formula exists. Since X is uncountable and $\mathbb{R}$ does not contain ω_1 or ω_1^* as suborderings, it follows from [26, 5.32] that X contains a dense subordering. Without loss of generality, we assume this dense subordering to be an open interval $Y = (a,b) = \{x \mid a < x < b\}$ (simply by extending it to an interval first and then picking any two distinct points of that interval). According to [26, 13.38], there exists a formula $\chi(x,y)$ in the language of $<$ such that for any $c_1, c_2 \in Y$, $(X,<) \models \psi(c_1,c_2)$ iff $(Y,<) \models \chi(c_1,c_2)$. Since Y is an open dense interval, by quantifier elimination for dense orders without endpoints we find a quantifier-free fromula $\xi(x,y)$ equivalent to $\chi(x,y)$. That is, for any $c_1, c_2 \in Y$, $(Y,<) \models \xi(c_1,c_2)$ iff $c_1 \lhd c_2$. However, it can be seen by a straightforward analysis of quantifier-free formulae in two variables that none of them defines a well-ordering on a dense ordered set. This contradiction proves the first part of the theorem. □

Proof of theorem 8, 2)

The first result is immediately implied by the proof of theorem 2 since we showed that the set X can always be chosen to be of size continuum. To prove the second result, we show the following: there exists a formula $\psi(x,y)$ in the language of $+,*$ such that for any set $X \subseteq \mathbb{R}$ with countable complement, and every $a,b \in X$, $\models_X \psi(a,b)$ iff $a < b$ (where $<$ is the usual ordering on the reals). We define this formula as

$$\psi(x,y) \;=\; \exists r_1 \exists r_2.[(\exists v_1 \exists v_2.\neg(r_1 + v_1 = v_1) \wedge \neg(r_2 + v_2 = v_2)) \;\wedge\; (y = x + r_1 * r_1 + r_2 * r_2)]$$

In other words, $\psi(x,y)$ says that we can find two nonzero elements of X such that $y - x = r_1^2 + r_2^2$. It is easy to show that this happens iff $x < y$. Indeed, the only if part is obvious, and to see the if part, let x be less than y and define the set $Y = \{\sqrt{y - x - r^2} \mid 0 < r < \sqrt{y - x}, r \in X\}$. Since this set is uncountable, it must contain an element of X, say s. Let $t = \sqrt{s}$ (by definition, $s > 0$). Then $y - x = r^2 + t^2$, finishing the proof that ψ defines the ordering.

Now assume that $\mathcal{FO}(\mathbb{R},+,*,\Omega) \Rightarrow_c^{\aleph_0} \mathcal{FO}(\mathbb{R})$. Then, for any schema SC and any $L(SC,+,*,\Omega)$ sentence ϕ, we can find a set X with countable complement such that any $D \in \mathit{Inst}(SC,X)$ satisfies ϕ iff it satisfies some sentence φ in the language of the schema relations. By [16] this means that $D \models_\emptyset \varphi$. Since ordering is definable as a $L(SC,+,*)$ formula, this means that for every $L(SC,<)$ sentence ϕ' we can find a $L(SC)$ sentence φ' such that $D \models \phi'$ iff $D \models_\emptyset \varphi'$. However, this is known not to be the case, see [1]. This contradiction proves the second part of the theorem. □