On Preservation under Homomorphisms and Unions of Conjunctive Queries

ALBERT ATSERIAS

Universitat Politècnica de Catalunya, Barcelona, Spain

ANUJ DAWAR

University of Cambridge Computer Laboratory, Cambridge, U.K.

AND

PHOKION G. KOLAITIS

IBM Almaden Research Center, San Jose, California

Abstract. Unions of conjunctive queries, also known as select-project-join-union queries, are the most frequently asked queries in relational database systems. These queries are definable by existential positive first-order formulas and are preserved under homomorphisms. A classical result of mathematical logic asserts that the existential positive formulas are the only first-order formulas (up to logical equivalence) that are preserved under homomorphisms on all structures, finite and infinite. The question of whether the homomorphism-preservation theorem holds for the class of all finite structures resisted solution for a long time. It was eventually shown that, unlike other classical preservation theorems, the homomorphism-preservation theorem does hold in the finite. In this article, we show that the homomorphism-preservation theorem holds also for several restricted classes of finite structures of interest in graph theory and database theory. Specifically, we show that this result holds for all classes of finite structures of bounded degree, all classes of finite structures of bounded treewidth, and, more generally, all classes of finite structures whose cores exclude at least one minor.

Categories and Subject Descriptors: F.4.1 [**Mathematical Logic and Formal Languages**]: Mathematical Logic—*Model theory*; H.2.3 [**Database Management**]: Languages—*Query languages*

A. Atserias was partially supported by CICYT TIN2004-0434 and by the European Commission through the RTN COMBSTRU HPRN-CT2002-00278.

A. Dawar was supported in part by EPSRC grant GR/S06721.

P. Kolaitis is on leave from UC Santa Cruz.

This research was partially supported by National Science Foundation Grant No. IIS-9907419.

Authors' addresses: A. Atserias, Universitat Politècnica de Catalunya, Barcelona, Spain, e-mail: atserias@lsi.upc.edu; A. Dawar, University of Cambridge Computer Laboratory, Cambridge, U.K., e-mail: anuj.dawar@cl.cam.ac.uk; P. G. Kolaitis, IBM Almaden Research Center, San Jose, CA, e-mail: kolaitis@almaden.ibm.com.

General Terms: Theory

Additional Key Words and Phrases: Conjunctive queries, Datalog, finite model theory, first-order logic, graph minors, homomorphisms, infinitary logic, preservation

1. *Introduction*

It is well known that the most frequently asked queries in databases are expressible in the *select-project-join-union* (SPJU) fragment of relational algebra (see Abiteboul et al. [1995]). From the point of view of relational calculus or first-order logic, the class of SPJU queries corresponds to the class of queries definable by *existential positive* formulas of first-order logic, that is, formulas built from atomic formulas using conjunction, disjunction, and existential quantification only. By distributing conjunctions and existential quantifiers over disjunctions, every existential positive formula can be written as a disjunction of existential formulas in which the quantifier-free part is a conjunction of atomic formulas. It is for this reason that SPJU queries are also known as *unions of conjunctive queries*. Starting with the work of Chandra and Merlin [1977], the study of conjunctive queries and their unions has occupied a central place in database theory; in particular, researchers have investigated in depth certain fundamental algorithmic problems about (unions of) conjunctive queries, such as the containment and the evaluation problem for these queries.

Let $\mathbf{A} = (A, R_1^{\mathbf{A}}, \ldots, R_m^{\mathbf{A}})$ and $\mathbf{B} = (B, R_1^{\mathbf{B}}, \ldots, R_m^{\mathbf{B}})$ be two relational structures over the same vocabulary (database schema) $R_1, \ldots, R_m$. Recall that a *homomorphism from* $\mathbf{A}$ *to* $\mathbf{B}$ is a map $h : A \to B$ such that for every relation symbol R_i and every tuple $\mathbf{a} = (a_1, \ldots, a_r)$ from A, if $\mathbf{a} \in R_i^{\mathbf{A}}$ then $h(\mathbf{a}) = (h(a_1), \ldots, h(a_r)) \in R_i^{\mathbf{B}}$. As already realized by Chandra and Merlin [1977], the study of conjunctive queries is intimately connected to homomorphisms. In particular, unions of conjunctive queries are preserved under homomorphisms, where a query q is said to be *preserved under homomorphisms* if whenever $\mathbf{a} \in q(\mathbf{A})$ and h is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$, then $h(\mathbf{a}) \in q(\mathbf{B})$. Note that if a query q is preserved under homomorphisms, then it is also preserved under *extensions*, which means that whenever $\mathbf{A}$ is an induced substructure of $\mathbf{B}$ and $\mathbf{a} \in q(\mathbf{A})$, then $\mathbf{a} \in q(\mathbf{B})$. In addition, such a query q is *monotone*, which means that whenever $\mathbf{a} \in q(\mathbf{A})$ and $\mathbf{B}$ is obtained from $\mathbf{A}$ by adding tuples to some of the relations of $\mathbf{A}$, then $\mathbf{a} \in q(\mathbf{B})$. These preservation properties can be thought of as asserting that the query satisfies a strong form of the open world assumption, in that a tuple in the result of the query will remain so under the addition of new facts to the databases, such as the introduction of new elements and new tuples in the relations.

Classical *preservation theorems* of model theory are results that match semantic properties of first-order formulas with syntactic properties of first-order formulas. Specifically, the Łoś-Tarski Theorem asserts that a first-order formula is preserved under extensions on all structures (finite and infinite) if and only if it is logically equivalent to an existential formula (see Hodges [1993]). Another classical preservation theorem in model theory, known as Lyndon's Positivity Theorem, states that a first-order formula is monotone on all structures (finite and infinite) if and only if it is logically equivalent to a positive first-order formula. The non-trivial part in these results is to show that if a first-order formula has the semantic property stated, then it is logically equivalent to a first-order formula that has the corresponding

syntactic property. The proofs make an essential use of the Compactness Theorem of first-order logic (and, hence, of infinite structures). The same technique can also be used to show that the following *homomorphism-preservation* theorem holds: a first-order formula is preserved under homomorphisms on all structures (finite and infinite) if and only if it is logically equivalent to an existential positive first-order formula.

The aforementioned classical preservation theorems are about the class of all structures (finite and infinite) over some fixed vocabulary. It is natural to ask whether these preservation theorems *relativize*, that is, whether they hold for restricted classes of structures. Note that if a preservation theorem holds for a class $\mathcal{C}$ of structures, then restricting the statement of the theorem to a subclass $\mathcal{C}'$ of $\mathcal{C}$ weakens both the hypothesis and the conclusion of the theorem. Thus, unlike many other results of model theory, a preservation theorem may hold for a class $\mathcal{C}$ of structures, but may fail for some subclass $\mathcal{C}'$ of $\mathcal{C}$.

Research in finite model theory addressed the question of whether classical preservation theorems about the class of all structures hold for the class of all finite structures. As it turned out, classical preservation theorems tend to fail when we restrict ourselves to finite structures. In particular, the Łoś–Tarski Theorem fails in the finite, that is, there is a first-order formula that is preserved under extensions on the class of all finite structures, but is not equivalent to any existential formula [Tait 1959; Gurevich 1984]. Similarly, Lyndon's Positivity Theorem is also known to fail in the finite [Ajtai and Gurevich 1987; Stolboushkin 1995]. As for the homomorphism-preservation theorem, its status in the finite had remained unsettled for quite a long time. In fact, the finite version of the homomorphism-preservation theorem had received considerable attention by the finite model theory community and had been singled out as a central problem (Problem 5.9 on the finite model theory website at `http://www-mgi.informatik.rwth-aachen.de/FMT/`). Moreover, it motivated a lot of related research in this area, including Alechina and Gurevich [1997], Feder and Vardi [2003], Gurevich [1990], and Rosen [2002]. Eventually, in an important breakthrough, Rossman [2005] proved that the homomorphism-preservation theorem *does* hold in the finite. In other words, Rossman proved that if a first-order formula is preserved under homomorphisms on the class of all finite structures, then it is equivalent, on finite structures, to an existential positive first-order formula. In particular, suppose that some arbitrary relational algebra query which may also involve the *set-theoretic difference* operator is preserved under homomorphisms on all finite structures; Rossman's result shows that this query can be transformed to an equivalent SPJU query.

In this article, we show that the homomorphism-preservation theorem holds for numerous restricted classes of finite structures of interest in graph theory and database theory. It should be noted that our results were established and published in preliminary form [Atserias et al. 2004] before Rossman proved that the homomorphism-preservation theorem holds for the class of all finite structures. It should also be pointed out that our results are not implied by Rossman's theorem, since, as explained earlier, preservation theorems about a class of structures need not relativize to a subclass of that class. In its full generality, our main result asserts that the homomorphism-preservation theorem holds for every class $\mathcal{C}$ of finite structures that is closed under substructures and disjoint unions, and has the property that the Gaifman graphs of the structures in $\mathcal{C}$ exclude at least one minor. This result contains as special cases the homomorphism-preservation

theorem for the classes of all structures of bounded treewidth, and the classes of all structures that exclude at least one minor; in particular, the homomorphism-preservation theorem holds for the class of all planar graphs. If we restrict attention to Boolean queries, we are able to further extend the classes of structures on which the homomorphism preservation theorem holds. In particular, we can show that the theorem for Boolean queries holds on every class $\mathcal{C}$ of finite structures that is closed under substructures and disjoint unions, and such that the *cores* of the structures in $\mathcal{C}$ exclude at least one minor. To put these results in perspective, let us briefly comment on some of the key notions. *Treewidth* is a measure of how tree-like a graph (or, more generally, a relational structure) is. It has played a key role in Robertson and Seymour's celebrated work on graph minors (see Downey and Fellows [1999]). Moreover, classes of structures of bounded treewidth have turned out to possess good algorithmic properties, in the sense that various NP-complete problems, including constraint satisfaction problems and database query evaluation problems, are solvable in polynomial-time when restricted to inputs of bounded treewidth [Dechter and Pearl 1989; Downey and Fellows 1999, Grohe et al. 2001, 2002]. The *core* of a structure $\mathbf{A}$ is a substructure $\mathbf{B}$ of $\mathbf{A}$ such that there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$, but there is no homomorphism from $\mathbf{A}$ to a proper substructure $\mathbf{B}'$ of $\mathbf{B}$. This concept originated in graph theory (see Hell and Nesetril [1992]), but has found applications in conjunctive query processing and optimization [Chandra and Merlin 1977] and, more recently, in data exchange [Fagin et al. 2003].

The proofs of our results combine earlier work about preservation properties in the finite with some heavy combinatorial machinery. Ajtai and Gurevich [1994] showed that if a query q on the class of all finite structures is expressible in both Datalog and first-order logic, then it is also definable by an existential positive formula; furthermore, every Datalog program defining q must be bounded. This is an important result about Datalog programs in its own right, but it is also a partial result towards the homomorphism-preservation theorem in the finite because all Datalog queries are preserved under homomorphisms (since such queries are infinitary unions of conjunctive queries). At a high level, the proof of the Ajtai–Gurevich theorem can be decomposed into two modular parts. The first is a combinatorial lemma to the effect that if q is a first-order query that is preserved under homomorphisms on finite structures, then the *minimal* models of q satisfy a certain "density" condition (incidentally, the minimal models of a Boolean query that is preserved under homomorphisms are cores). The second part shows that if all minimal models of a Datalog query satisfy the "density" condition, then there are only finitely many of them. This means that q has finitely many minimal models, which easily implies that q is definable by a union of conjunctive queries. To obtain our main theorem, we use the same architecture in the proof, but, in place of the second part, we essentially show that if $\mathcal{C}$ is a class of finite structures satisfying the hypothesis of the theorem (such as having bounded treewidth or excluding a minor), then every collection of structures in $\mathcal{C}$ that satisfies the "density" condition must be finite. In turn, this requires the use of the Sunflower Lemma of Erdös and Rado, as well as Ramsey's Theorem.

Furthermore, equipped with this new machinery, we obtain a different and perhaps more transparent proof of the Ajtai–Gurevich Theorem. Actually, we show that the Ajtai–Gurevich Theorem can be extended to a family of finite-variable infinitary logics that taken together are strictly more expressive than Datalog. This

is obtained by using tight connections between number of variables, treewidth, and minimal models.

In Section 2, we review some basic notions from logic and graph theory that we will need in the sequel. Section 3 contains certain combinatorial facts about the minimal models of a first-order query that is preserved under homomorphisms. In Sections 4 and 5, we establish the main results regarding classes of bounded treewidth and classes with excluded minors respectively. In Section 6, we discuss the relationship between peservation for Boolean and non-Boolean queries. We show that the preservation results for Boolean queries can be established for larger classes of structures. Finally, in Section 7, we obtain the aforementioned extension of the Ajtai–Gurevich Theorem.

2. *Preliminaries*

This section contains the definitions of some basic notions and a minimum amount of background material.

2.1. RELATIONAL STRUCTURES AND GRAPHS. A *relational vocabulary* σ is a finite set of *relation symbols*, each with a specified *arity*. A σ*-structure* $\mathbf{A}$ consists of a *universe* A, or *domain*, and an *interpretation* which associates to each relation symbol $R \in \sigma$ of some arity r, a relation $R^{\mathbf{A}} \subseteq A^r$. A *graph* is a structure $\mathbf{G} = (V, E)$, where E is a binary relation that is symmetric and irreflexive. Thus, our graphs are undirected, loopless, and without parallel edges.

A σ-structure $\mathbf{B}$ is called a *substructure* of $\mathbf{A}$ if $B \subseteq A$ and $R^{\mathbf{B}} \subseteq R^{\mathbf{A}}$ for every $R \in \sigma$. It is called an *induced substructure* if $R^{\mathbf{B}} = R^{\mathbf{A}} \cap B^r$ for every $R \in \sigma$ of arity r. Notice the analogy with the graph-theoretical concept of *subgraph* and *induced subgraph*. A substructure $\mathbf{B}$ of $\mathbf{A}$ is proper if $\mathbf{A} \neq \mathbf{B}$.

A *homomorphism* from $\mathbf{A}$ to $\mathbf{B}$ is a mapping $h : A \to B$ from the universe of $\mathbf{A}$ to the universe of $\mathbf{B}$ that preserves the relations, that is if $(a_1, \ldots, a_r) \in R^{\mathbf{A}}$, then $(h(a_1), \ldots, h(a_r)) \in R^{\mathbf{B}}$. We say that two structures $\mathbf{A}$ and $\mathbf{B}$ are *homomorphically equivalent* if there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$ and a homomorphism from $\mathbf{B}$ to $\mathbf{A}$. Note that, if $\mathbf{A}$ is a substructure of $\mathbf{B}$, then the injection mapping is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$.

The *Gaifman graph* of a σ-structure $\mathbf{A}$, denoted by $\mathcal{G}(\mathbf{A})$, is the (undirected) graph whose set of nodes is the universe of $\mathbf{A}$, and whose set of edges consists of all pairs (a, a') of distinct elements of A such that a and a' appear together in some tuple of a relation in $\mathbf{A}$. The *degree* of a structure is the degree of its Gaifman graph, that is, the maximum number of neighbors of nodes of the Gaifman graph.

Let $\mathbf{G} = (V, E)$ be a graph. Moreover, let $u \in V$ be a vertex and let $d \geq 0$ be an integer. The *d-neighborhood* of u in $\mathbf{G}$, denoted by $N_d^{\mathbf{G}}(u)$, is defined inductively as follows:

(1) $N_0^{\mathbf{G}}(u) = \{u\}$;

(2) $N_{d+1}^{\mathbf{G}}(u) = N_d^{\mathbf{G}}(u) \cup \{v \in V : (v, w) \in E \text{ for some } w \in N_d^{\mathbf{G}}(u)\}$.

A *tree* is an acyclic connected graph. A *tree-decomposition* of $\mathbf{G}$ is a labeled tree $\mathbf{T}$ such that

(1) each node of $\mathbf{T}$ is labeled by a non-empty subset of V;

(2) for every edge $\{u, v\} \in E$, there is a node of $\mathbf{T}$ whose label contains $\{u, v\}$;

(3) for every $u \in V$, the set X of nodes of $\mathbf{T}$ whose labels include u forms a connected subtree of $\mathbf{T}$.

The *width* of a tree-decomposition is the maximum cardinality of a label in $\mathbf{T}$ minus one. The *treewidth* of $\mathbf{G}$ is the smallest k for which $\mathbf{G}$ has a tree-decomposition of width k. The *treewidth* of a σ-structure is the treewidth of its Gaifman graph. Note that trees have treewidth one.

For every positive integer $k \geq 2$, we write $\mathcal{T}(k)$ to denote the class of all σ-structures of treewidth less than k. In the sequel, whenever we say that a collection $\mathcal{C}$ of σ-structures has *bounded treewidth*, we mean that there is a positive integer k such that $\mathcal{C} \subseteq \mathcal{T}(k)$.

We say that a graph $\mathbf{G}$ is a *minor* of $\mathbf{H}$ if $\mathbf{G}$ can be obtained from a subgraph of $\mathbf{H}$ by contracting edges. The contraction of an edge consists in identifying its two endpoints into a single node, and removing the resulting loop. An equivalent characterization (see Diestel [1997]) states that $\mathbf{G}$ is a minor of $\mathbf{H}$ if there is a map that associates to each vertex v of $\mathbf{G}$ a nonempty *connected* subgraph $\mathbf{H}_v$ of $\mathbf{H}$ such that $\mathbf{H}_u$ and $\mathbf{H}_v$ are disjoint for $u \neq v$ and if there is an edge between u and v in $\mathbf{G}$ then there is an edge in $\mathbf{H}$ between some node in $\mathbf{H}_u$ and some node in $\mathbf{H}_v$. We will sometimes refer to the subgraphs $\mathbf{H}_v$ as the *connected patches* that witness that $\mathbf{G}$ is a minor of $\mathbf{H}$.

It is not hard to see that $\mathcal{T}(k)$ is closed under taking minors, that is, if $\mathbf{G}$ is a minor of $\mathbf{H}$ and the treewidth of $\mathbf{H}$ is less than k, then the treewidth of $\mathbf{G}$ is also less than k. Since the treewidth of $\mathbf{K}_k$, the complete graph on k vertices, is $k-1$, it follows that $\mathbf{K}_{k+1}$ is not a minor of any graph in $\mathcal{T}(k)$. Finally, we will make use of the fact that $\mathbf{K}_k$ is a minor of $\mathbf{K}_{k-1,k-1}$, the complete bipartite graph on two sets of $k-1$ nodes. To see this, contract the edges of a perfect matching of size $k-2$ sitting inside $\mathbf{K}_{k-1,k-1}$. The result is a complete graph on $k-2$ nodes, which, together with the remaining two nodes of $\mathbf{K}_{k-1,k-1}$ and all remaining edges, gives a $\mathbf{K}_k$.

2.2. First-Order Logic and Conjunctive Queries. Let σ be a relational vocabulary. The *atomic formulas* of σ are those of the form $R(x_1, \ldots, x_r)$, where $R \in \sigma$ is a relation symbol of arity r, and $x_1, \ldots, x_r$ are first-order variables that are not necessarily distinct. Formulas of the form $x = y$ are also atomic formulas, and we refer to them as *equalities*. The collection of *first-order formulas* is obtained by closing the atomic formulas under negation, conjunction, disjunction, universal and existential first-order quantification. The semantics of first-order logic is standard. If $\mathbf{A}$ is a σ-structure and φ is a first-order formula, we use the notation $\mathbf{A} \models \varphi$ to denote the fact that φ is true in $\mathbf{A}$. The collection of *existential-positive* first-order formulas is obtained by closing the atomic formulas under conjunction, disjunction, and existential quantification. By substituting variables, it is easy to see that equalities can be eliminated from existential-positive formulas.

An important fragment of existential-positive formulas is formed by the collection of sentences of the form $\exists x_1 \cdots \exists x_n \theta$, where θ is a conjunction of atomic formulas with variables among $x_1, \ldots, x_n$. These formulas define the class of Boolean *conjunctive* queries (also known as *select-project-join* queries or, in short, SPJ-queries). In the sequel, we will occasionally use the term *conjunctive query* to denote both a formula $\exists x_1 \cdots \exists x_n \theta$ as above and the query defined by that formula. Every finite structure $\mathbf{A}$ with n elements gives rise to a *canonical conjunctive query* $\varphi_{\mathbf{A}}$, which is obtained by first associating a different variable x_i with every element a_i of $\mathbf{A}$, $1 \leq i \leq n$, then forming the conjunction of all atomic facts true

in **A**, and finally existentially quantifying all variables x_i, $1 \leq i \leq n$. In other words, the formula $\varphi_{\mathbf{A}}$ is the existential closure of the *positive diagram* of **A** (see Hodges [1993]). Conversely, every conjunctive query $\exists x_1 \cdots \exists x_n \theta$ gives rise to a *canonical structure* **A** with n elements, where the elements of **A** are the variables $x_1, \ldots, x_n$ and the relations of **A** consist of the tuples of variables in the conjuncts of θ. Chandra and Merlin [1977] showed the following basic result, which has found many uses in database theory and the theory of constraint satisfaction problems.

THEOREM 2.1 (CHANDRA–MERLIN THEOREM). *Let* **A** *and* **B** *be two finite structures. The following statements are equivalent.*

(1) *There is a homomorphism from* **A** *to* **B**.

(2) $\mathbf{B} \models \varphi_{\mathbf{A}}$.

(3) $\varphi_{\mathbf{B}}$ *logically implies* $\varphi_{\mathbf{A}}$.

2.3. INDUCTIVE DEFINITIONS AND DATALOG. Let σ be a relational vocabulary. An *inductive system* of first-order formulas is a finite sequence

$$\varphi_1(x_1, \ldots, x_{k_1}, S_1, \ldots, S_r), \ldots, \varphi_r(x_1, \ldots, x_{k_r}, S_1, \ldots, S_r)$$

of first-order formulas such that each S_i is a relation symbol of arity k_i, not already in σ. Every such system gives rise to an operator Φ on sequences of relations of a σ-structure. More precisely, if **A** is a σ-structure with universe A and $R_i \subseteq A^{k_i}$ is a relation for every $i \in \{1, \ldots, r\}$, we define

$$\Phi_i(R_1, \ldots, R_r) = \{(a_1, \ldots, a_{k_i}) \in A^{k_i} : \mathbf{A} \models \varphi_i(a_1, \ldots, a_{k_i}, R_1, \ldots, R_r)\},$$

and $\Phi(R_1, \ldots, R_r) = (\Phi_1(R_1, \ldots, R_r), \ldots, \Phi_r(R_1, \ldots, R_r))$. The stages $\Phi^m = (\Phi_1^m, \ldots, \Phi_r^m)$ of Φ are defined by the induction $\Phi_i^0 = (\emptyset, \ldots, \emptyset)$, and $\Phi_i^{m+1} = \Phi_i(\Phi_1^m, \ldots, \Phi_r^m)$. If each formula φ_i is positive in the relation symbols $S_1, \ldots, S_r$, then the associated operator Φ is monotone in each of its arguments. In such a case, the sequence of stages $\Phi^0, \Phi^1, \ldots$ converges to the least fixed-point $\Phi^\infty = (\Phi_1^\infty, \ldots, \Phi_r^\infty)$ of the operator Φ. Moreover, if **A** is finite, then there exists a finite m_0 such that $\Phi^\infty = \Phi^{m_0}$.

A *Datalog program* is a finite set of rules of the form $T_0 \leftarrow T_1, \ldots, T_m$, where each T_i is an atomic formula. The left-hand side of each rule is called the *head* of the rule, while the right-hand side is called the *body*. The relation symbols that occur in the heads are the *intensional* database predicates (IDBs), while all others are the *extensional* database predicates (EDBs). Note that IDBs may occur in the bodies too, thus, a Datalog program is a recursive specification of the IDBs with semantics obtained via least fixed-points of monotone operators (see Ullman [1989]). For example, the following Datalog program defines the *transitive closure* of the edge relation E of a graph $\mathbf{G} = (V, E)$:

$$\begin{aligned} T(x, y) &\leftarrow E(x, y); \\ T(x, y) &\leftarrow E(x, z), T(z, y). \end{aligned}$$

A key parameter in analyzing Datalog programs is the number of variables used. We write k-Datalog for the collection of all Datalog programs with at most k variables in total. For instance, the above is a 3-Datalog program. A Datalog program can be

read as an inductive system of first-order formulas (as above) where each formula is existential positive.

Let $\mathcal{C}$ be a class of σ-structures. A query q on $\mathcal{C}$ of arity n is a map that associates to each structure $\mathbf{A}$ in $\mathcal{C}$ an n-ary relation $q(\mathbf{A})$ on the domain of $\mathbf{A}$ that is preserved under isomorphisms between structures. Let L be some logic. We say that q is *L-definable on $\mathcal{C}$* if there exists a formula φ of L such that if $\mathbf{A}$ is in $\mathcal{C}$, then $\mathbf{a} \in q(\mathbf{A})$ if and only if $\mathbf{A}, \mathbf{a} \models \varphi$. A Boolean query is a query of arity 0, which can be identified with an isomorphism-closed subclass of $\mathcal{C}$. Equivalently, a Boolean query is a mapping q from $\mathcal{C}$ to $\{0, 1\}$ that is invariant under isomorphisms. We say that a Boolean query q is *L-definable on $\mathcal{C}$* if there is a sentence ψ of L such that for every $\mathbf{A} \in \mathcal{C}$, we have that $q(\mathbf{A}) = 1$ if and only if $\mathbf{A} \models \psi$.

3. *Preservation under Homomorphisms and Minimal Models*

For the purpose of the constructions in this article, we shall restrict our attention specifically to Boolean queries. The reason for restricting ourselves to Boolean queries is that the notion of *minimal model*, which we rely on, is more naturally defined for Boolean queries. In Section 6, we return to non-Boolean queries and explain why the results apply equally well to these.

For a Boolean query q, we say that a σ-structure $\mathbf{A}$ in $\mathcal{C}$ is a *minimal model of q in $\mathcal{C}$* if $q(\mathbf{A}) = 1$ and there is no proper substructure $\mathbf{B}$ of $\mathbf{A}$ in $\mathcal{C}$ such that $q(\mathbf{B}) = 1$. Recall from Section 2 that substructures are not necessarily induced.

The following characterization is part of the folklore, a proof for the class of all finite σ-structures can be found in Alechina and Gurevich [1997]. Here, we state it in a more general form for classes of finite σ-structures that are closed under substructures, and sketch a proof.

THEOREM 3.1. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures, and let q be a Boolean query on $\mathcal{C}$ that is preserved under homomorphisms on $\mathcal{C}$. The following are equivalent:*

(1) *q has finitely many minimal models in $\mathcal{C}$.*

(2) *q is definable on $\mathcal{C}$ by an existential-positive first-order sentence.*

PROOF. The direction (1)⇒(2) is established by constructing, for each finite structure $\mathbf{A}$, a *canonical* conjunctive query $\varphi_{\mathbf{A}}$, as described earlier. The required existential positive formula defining q is now obtained as the disjunction of $\varphi_{\mathbf{A}}$ over all minimal models $\mathbf{A}$ of q. This follows from the preservation of q under homomorphisms and the fact that, by Theorem 2.1, a structure $\mathbf{B}$ satisfies $\varphi_{\mathbf{A}}$ if and only if there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$.

For the direction (2)⇒(1), we first use the fact that every existential positive formula is equivalent to a finite disjunction $\bigvee_{i=1}^{m} \psi_i$, where each ψ_i is a conjunctive query. For each such conjunctive query ψ_i, let $\mathbf{A}_i$ be the *canonical* finite structure associated with ψ_i, $1 \leq i \leq m$. Note that such a canonical structure $\mathbf{A}_i$ need not be a member of $\mathcal{C}$. Nonetheless, it is not hard to show that every minimal model $\mathbf{B}$ of q in $\mathcal{C}$ is equal to a homomorphic image $h(\mathbf{A}_i)$ of one of the canonical finite structures $\mathbf{A}_i$, $1 \leq i \leq m$. Thus, the cardinality of every minimal model of q in $\mathcal{C}$ is less than or equal to the maximum cardinality of the canonical finite structures $\mathbf{A}_i$, $1 \leq i \leq m$, which implies that q has finitely many minimal models in $\mathcal{C}$. □

By Theorem 3.1, to establish the homomorphism-preservation theorem for the class of all finite structures, we would need to show that any first-order definable query preserved under homomorphisms has only finitely many minimal models. Equivalently, it would suffice to show that for any such query there is a bound on the size of the minimal models. Ajtai and Gurevich [1994], in comparing the expressive power of Datalog and first-order logic, showed that the minimal models of every first-order sentence preserved under homomorphisms satisfy an interesting combinatorial property. Intuitively speaking, they are *dense*. More precisely, if there are arbitrarily large minimal models, then they cannot be very thinly spread out, which means that they do not contain a large set of elements all far away from each other. Furthermore, one cannot remove a small number of elements from a large minimal model to create such a scattered set.

The Ajtai-Gurevich proof of this property is based on Gaifman's Locality Theorem for first-order logic [Gaifman 1982]. Before we state the precise result, we need a definition and a piece of notation. Let $\mathbf{G} = (V, E)$ be a graph. Recall the definition of d-neighborhood $N_d^{\mathbf{G}}(u)$ in Section 2. We say that a subset $A \subseteq V$ of the nodes is *d-scattered* if $N_d^{\mathbf{G}}(u) \cap N_d^{\mathbf{G}}(v) = \emptyset$ for every two distinct $u, v \in A$. For a graph $\mathbf{G} = (V, E)$ and a set $B \subseteq V$, we write $\mathbf{G} - B$ for the graph obtained from $\mathbf{G}$ by removing all nodes in B and the edges to which they are incident. This is a notation we will use repeatedly in the sequel. We are ready for the result of Ajtai and Gurevich. While they proved this for the class of all finite structures, it is easy to see that the proof relativizes to classes satisfying some simple restrictions. This observation follows from the fact that disjoint union and taking a substructure are the only constructions used in the proof in Ajtai and Gurevich [1994].

THEOREM 3.2. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions. Let q be a Boolean query that is first-order definable and preserved under homomorphisms on $\mathcal{C}$. For every $s \geq 0$, there exist integers $d \geq 0$ and $m \geq 0$ such that if $\mathbf{A}$ is a minimal model of q, then there is no $B \subseteq A$ of size at most s such that $\mathcal{G}(\mathbf{A}) - B$ has a d-scattered set of size m. In particular, there exist integers $d \geq 0$ and $m \geq 0$ such that if $\mathbf{A}$ is a minimal model of q, then $\mathcal{G}(\mathbf{A})$ does not have a d-scattered set of size m.*

Now, let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions. With Theorems 3.1 and 3.2 in hand, in order to establish that the homomorphism-preservation theorem holds on $\mathcal{C}$, it suffices to show that for some s and every d and m, all sufficiently large structures in $\mathcal{C}$ have d-scattered sets of size m after removing at most s elements. We formulate this observation as the following corollary, which we will use repeatedly in what follows.

COROLLARY 3.3. *Let $\mathcal{C}$ be a class of finite σ-structures having the following properties:*

(1) *$\mathcal{C}$ is closed under substructures and disjoint unions;*
(2) *for some s and for all d and m, there is an N so that if $\mathbf{A} \in \mathcal{C}$ has more than N elements, then there is a set B of at most s elements such that $\mathcal{G}(\mathbf{A}) - B$ has a d-scattered set of size m.*

On the class $\mathcal{C}$, every Boolean query that is first-order definable and preserved under homomorphisms is definable by an existential positive first-order formula.

There is a case that is particularly easy in which we can take $s = 0$.

LEMMA 3.4. *For every $k \geq 0$, $d \geq 0$, and $m \geq 0$, there exists an $N \geq 0$ such that for all graphs $\mathbf{G} = (V, E^{\mathbf{G}})$ with $|V| > N$ and degree at most k, the graph $\mathbf{G}$ has a d-scattered set of size m.*

PROOF. Fix $d \geq 0$ and $m \geq 0$, let $N = mk^d$, and let $\mathbf{G} = (V, E^{\mathbf{G}})$ be a graph with $|V| > N$. The size of the d-neighborhood of every node in $\mathbf{G}$ is bounded by k^d. Therefore, there are at least m nodes in $\mathbf{G}$ with disjoint d-neighborhoods. □

As an immediate corollary, we obtain the homomorphism-preservation result for classes of structures of bounded-degree.

THEOREM 3.5. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the structures in $\mathcal{C}$ have bounded degree. On the class $\mathcal{C}$, every query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

4. *Classes of Bounded Treewidth*

In this section, we establish the homomorphism-preservation theorem for classes of bounded treewidth. Our aim is to show a combinatorial result to the effect that if we have a bound on the treewidth of structures in a class, then every sufficiently large structure will contain a large scattered set, after we have removed a small number of elements. The results in this section are subsumed by those in Section 5, since a class of structures of bounded treewidth excludes at least one minor (namely, some clique). However, the proof method for classes of bounded treewidth is simpler than the one presented in Section 5 and also yields better bounds on the maximum size of minimal models, so we present it separately.

Unlike for Lemma 3.4, it is no longer sufficient to take $s = 0$. To gain some intuition, consider the tree $\mathbf{S}_n$ which consists of a single root with n children. Since every pair of nodes is at most at distance 2, it is clear that $\mathbf{S}_n$ does not contain a d-scattered set for $d > 1$, yet the tree can be arbitrarily large. However, removing the root leaves a graph where the remaining nodes are scattered as no edges are left. This idea generalizes to arbitrary trees, in the sense that in every sufficiently large tree, we need to remove *at most one* node in order to create a large scattered set. For, either the tree has a node of large degree or a long path. In the first case, we remove a node of large degree and get a large number of disconnected components, hence a scattered set. In the second case, along the long path, we can select a set of elements that are pairwise far away from each other and thus form a scattered set. We generalize this idea to graphs of small treewidth. It turns out that the maximum number of nodes we need to remove to create any desired scattered set is bounded by the treewidth. This is proved using the Sunflower Lemma of Erdös and Rado [1960].

THEOREM 4.1 (SUNFLOWER LEMMA). *Let F be a collection of k-element subsets of a set A. If $|F| > k!(p-1)^k$, then F contains a sunflower with p petals, that is, a subcollection $F' \subseteq F$ of size p for which there exists a set B such that every pair of distinct sets X and Y in F' satisfy $B = X \cap Y$.*

Here is the promised combinatorial result:

LEMMA 4.2. *For every $k \geq 1$, $d \geq 0$, and $m \geq 0$, there exists an $N \geq 0$ such that for all graphs $\mathbf{G} = (V, E^{\mathbf{G}})$ with $|V| > N$ and treewidth less than k, there exists $B \subseteq V$ of size at most k such that $\mathbf{G} - B$ has a d-scattered set of size m.*

PROOF. Let $k \geq 1, d \geq 0$, and $m \geq 0$ be fixed. Define $p = (m-1)(2d+1)+1$, $M = k!(p-1)^k$, and $N = k(m-1)^M$. Let $\mathbf{G} = (V, E^{\mathbf{G}})$ be a graph with $|V| > N$, and let us assume its treewidth is less than k. Let $(\mathbf{T}, \{S_v : v \in T\})$ be a tree-decomposition of $\mathbf{G}$ with sets $S_v \subseteq V$ of size at most k. By standard manipulation on tree-decompositions, we may assume that for every pair of distinct nodes $u, v \in T$, both $S_u - S_v$ and $S_v - S_u$ are nonempty. Observe that the size of T is at least $N/k + 1$. We distinguish two cases:

Case 1. There is a node in $\mathbf{T}$ of degree at least m. Let v be such a node and $B = S_v$. Note that $|B| \leq k$. By our assumption on the tree-decomposition, we know that $S_u - S_v$ is non-empty for every neighbor u of v. Therefore, the graph $\mathbf{G} - B$ contains at least m disconnected components, so a d-scattered set of size m.

Case 2. There is no node in $\mathbf{T}$ of degree at least m. In this case, since the size of T is more than $N/k = (m-1)^M$, there must exist a path in $\mathbf{T}$ of length at least M. Since each S_v on this path has size at most k, and since the length of the path is at least $M = k!(p-1)^k$, by the Sunflower Lemma, there must exist $p = (m-1)(2d+1)+1$ sets $S_{u_1}, \ldots, S_{u_p}$ on this path with a common intersection B. Clearly, $|B| \leq k$, and all $T_i = S_{u_i} - B$ are pairwise disjoint and non-empty by our assumption on the tree-decomposition. We claim that choosing an arbitrary element in $T_{1+i(2d+1)}$ for each $i \in \{0, \ldots, m-1\}$ produces a d-scattered subset in $\mathbf{G} - B$. To see this, we need some notation. Let $R = \bigcup_{i=1}^{p} T_i$ be the union of petals. For $a, b \in R$, let $d(a, b)$ denote the distance between a and b in $\mathbf{G} - B$. For every point $a \in R$, let $P(a) = \{v \in T : a \in S_v\}$. Note that every $P(a)$ is a connected subtree of $\mathbf{T}$ by the third clause of the definition of tree-decomposition. Moreover, since the T_i's are pairwise disjoint, each $P(a)$ contains at most one of the nodes $u_1, \ldots, u_p$ of the sunflower. Consider the shortest path in $\mathbf{T}$ going from a node in $P(a)$ to a node in $P(b)$. We let $m(a, b)$ denote the number of nodes of the sunflower that appear in this path.

CLAIM 4.3. *If a and b belong to R, then $m(a, b) \leq d(a, b)$.*

Suppose a and b are points in R. We proceed by induction on the length n of the shortest path between a and b in $\mathbf{G} - B$. The base case $n = 0$ is obvious since then $m(a, b) = d(a, b) = 0$. We are ready for the inductive case. Let $a = a_0, a_1, \ldots, a_{n+1} = b$ be a shortest path of length $n+1$ in $\mathbf{G} - B$ and assume the claim is true for shorter path-lengths. We need to prove that $m(a, b) \leq n+1$. If $m(a, b) = 0$, there is nothing to prove. Suppose then that $m(a, b) > 0$ and let u_j be a node of the sunflower that appears in the shortest path of the tree between $P(a)$ and $P(b)$ and is closest to $P(b)$. By the second property of tree-decomposition, any path in $\mathbf{G} - B$ from a to b must go through some point in T_j. So let $k \in \{1, \ldots, n\}$ be such that a_k belongs to T_j. Let $c = a_k$, note that $c \in R$, and that the length of the shortest path between a and c in $\mathbf{G} - B$ is $k \leq n$. By induction hypothesis, $m(a, c) \leq d(a, c)$. But also $m(c, b) = 0$ by the choice of j and c in T_j. Thus $m(a, b) \leq m(a, c) + 1$ because $P(c)$ contains at most one node of the sunflower. It follows that $m(a, b) \leq d(a, c) + 1 \leq d(a, b)$, which completes the proof of the claim.

For each $i \in \{0, \ldots, m-1\}$, choose an element a_i in $T_{1+i(2d+1)}$. Then, we have $m(a_i, a_j) > 2d$ for $i \neq j$. The lemma follows from the claim. □

We obtain the homomorphism-preservation theorem for classes of structures of bounded treewidth as an immediate consequence of Lemma 4.2 and Corollary 3.3.

THEOREM 4.4. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the structures in $\mathcal{C}$ have bounded treewidth. On the class $\mathcal{C}$, every query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

Many interesting classes have bounded treewidth. Among others, we find the class of all trees, the class of all unicyclic graphs, and the class of all outerplanar graphs.

5. *Classes with Excluded Minors*

In this section, we extend the combinatorial results from the previous section to classes of graphs, which exclude a minor. We say a class of graphs $\mathcal{C}$ *excludes a graph* $\mathbf{G}$ *as a minor* if no graph in $\mathcal{C}$ has $\mathbf{G}$ as a minor. Note that, every graph $\mathbf{G}$ is a minor of $\mathbf{K}_k$, where k is the number of nodes in $\mathbf{G}$. Thus, if $\mathcal{C}$ excludes $\mathbf{G}$ as a minor, it also excludes $\mathbf{K}_k$ because the graph minor relation is transitive. It therefore suffices to establish our result for classes of structures that exclude $\mathbf{K}_k$ as a minor for some k.

We aim to show that in the class of graphs that exclude $\mathbf{K}_k$ as a minor, every sufficiently large graph will contain large scattered sets after the removal of a small number of elements. Intuitively, if a graph does not contain such a scattered set, then there is a large number of elements with short paths between each pair. Either various paths must pass through a small number of elements or they are nearly disjoint. In the former case, we can remove the elements to get a scattered set; in the latter, we can find $\mathbf{K}_k$ as a minor in the graph. It turns out, again, that k provides a bound on the number of elements we need to remove.

The formal proof of this intuitive idea is inspired by a construction due to Kreidler and Seese [1999], which establishes a result closely related to Theorem 5.3 below (see also Kreidler [1999]). Before the main result, we establish a lemma on bipartite graphs. The proof relies on Ramsey's Theorem (see Graham et al. [1980]).

THEOREM 5.1 (RAMSEY'S THEOREM). *For every $l \geq 0$, $k \geq 0$ and $m \geq 0$, there is an $N \geq 0$ such that if A is a set with $|A| > N$ and $f : [A]^k \to \{1, \ldots, l\}$ a function on the k-element subsets of A, there is a set $I \subseteq A$ with $|I| > m$ such that f is constant on $[I]^k$, the k-element subsets of I.*

For later use, we write $r(l, k, m)$ for the bound N obtained in Ramsey's Theorem. Although we will need it in its full generality, let us briefly comment on the particular case $r(2, 2, m)$. This is a bound for the graph version of Ramsey's Theorem: every graph with more than $r(2, 2, m)$ vertices contains either an independent set with more than m elements or a clique with more than m elements.

The following lemma will be a key stepping stone towards the main result. The lemma says, roughly, that every large bipartite graph $\mathbf{H} = (A \cup B, E \subseteq A \times B)$ that excludes $\mathbf{K}_k$ as a minor contains a large set of points $A' \subseteq A$ without common

neighbors in B, except for a small set of exceptional points $B' \subseteq B$ that are indeed common neighbors of all points in A'. The fact that $\mathbf{H}$ excludes $\mathbf{K}_k$ as a minor guarantees that the set of exceptional points B' is kept small.

LEMMA 5.2. *For every $k \geq 1$ and $m \geq 0$, there is an $N \geq 0$ such that if $\mathbf{H} = (A \cup B, E \subseteq A \times B)$ is a bipartite graph such that $\mathbf{K}_k$ is not a minor of $\mathbf{H}$ and $|A| > N$, then there are sets $A' \subseteq A$ and $B' \subseteq B$ with $|A'| > m$ and $|B'| < k-1$ such that $A' \times B' \subseteq E$ and A' is 1-scattered in $\mathbf{H} - B'$.*

PROOF. The case $k \leq 2$ is trivial as, if $\mathbf{K}_2$ is not a minor of $\mathbf{H}$, then $\mathbf{H}$ contains no edges and taking $N = m$ suffices. We will therefore assume that $k \geq 3$ below. Furthermore, if the lemma is true for some value of m it is also true for all $m' \leq m$. Thus, it suffices to prove it for all large enough m. In what follows, we assume that $m \geq k^2$. Define the function

$$b(n) = r(k+1, k, (k-2)n + k - 2),$$

where r is the Ramsey function. Define $b^0(m) = m$ and $b^{i+1}(m) = b(b^i(m))$, and let $N = b^{k-2}(m)$. We construct the sets A' and B' in a series of stages:

$$A_0 \supseteq A_1 \supseteq \cdots \supseteq A'$$
$$B_0 \subseteq B_1 \subseteq \cdots \subseteq B'.$$

The number of stages of this construction will be less than $k-1$. Begin with $A_0 = A$ and $B_0 = \emptyset$. Now, suppose at stage $r < k-2$ we have sets $A_r \subseteq A$ and $B_r \subseteq B$, with $|B_r| \leq r$ and $|A_r| > b^{k-2-r}(m)$, and such that $A_r \times B_r \subseteq E$. We define A_{r+1} and B_{r+1}. Let $<$ be an arbitrary linear ordering of A_r. Let $f : [A_r]^k \to \{0, \ldots, k\}$ be the function that assigns to each k-element subset $x_1 < x_2 < \cdots < x_k$ of A_r the maximum $j \in \{0, \ldots, k\}$ such that all $x_1, \ldots, x_j$ have a common neighbor in $B - B_r$. By Ramsey's Theorem, there is a set $I \subseteq A_r$, with

$$|I| > (k-2)b^{k-2-(r+1)}(m) + k - 2$$

such that f is constant on $[I]^k$. We consider three cases:

Case 1. $f([I]^k) \leq 1$. Let C denote the last $k-2$ elements of I under the order $<$. Then, $I - C$ is 1-scattered in $\mathbf{H} - B_r$ as every pair of elements in $I - C$ forms the first two elements of some ordered k-element subset of I and therefore cannot have a common neighbor. Note also, that

$$|I - C| \geq (k-2)b^{k-2-(r+1)}(m).$$

Since $r < k-2$, this means $|I - C| \geq (k-2)m \geq m$ as $k \geq 3$. Thus, taking $A' = A_{r+1} = I - C$ and $B' = B_{r+1} = B_r$, we are done.

Case 2. $1 < f([I]^k) < k$. We will argue that, indeed, this case cannot occur. Let $f([I]^k) = t$. If C denotes the last $k-t$ elements of I under the order $<$, then every t-element subset of $I - C$ has a common neighbor in $B - B_r$, as it is the initial segment of size t of some k-element subset of I. Furthermore, no $(t+1)$-element subset of $I - C$ has a common neighbor in $B - B_r$, from which we conclude that the maximal degree of any element in $B - B_r$ (with respect to $I - C$) is t. Now, let $X_1, \ldots, X_k \subseteq I - C$ be a collection of k pairwise disjoint sets, each with exactly t elements. Such a collection exists, since

$$|I - C| > (k-2)b^{k-2-(r+1)}(m) \geq m \geq k^2.$$

Then, by the argument above, for each X_i, there is a $u_i \in B - B_r$ which is a common neighbor of all elements in X_i, and u_i has no other neighbors. Thus, the set $X_i \cup \{u_i\}$ forms a connected patch in the graph $\mathbf{H} - B_r$. Similarly, for each i and j with $1 \leq i < j \leq k$, we can find an element $u_{ij} \in B - B_r$ such that, if $N(u_{ij})$ denotes the set of neighbors of u_{ij} in I, then:

(1) $N(u_{ij}) \subseteq X_i \cup X_j$
(2) $N(u_{ij}) \cap X_i \neq \emptyset$
(3) $N(u_{ij}) \cap X_j \neq \emptyset$.

This is possible as X_i and X_j are disjoint and each has $t > 1$ elements. Thus, we can choose a subset of $X_i \cup X_j$ that meets both sets and has exactly t elements. Any common neighbor of this subset would serve as u_{ij}. Again, u_{ij} cannot have any other neighbors in $I - C$, as no $(t+1)$-element subset of $I - C$ has a common neighbor. Thus, in particular, u_{ij} has no neighbors in any X_l for l different from i and j. We have thus found k distinct connected patches $X_i \cup \{u_i\}$ and pairwise disjoint paths (of length 2) between any pair of them. Thus, $\mathbf{K}_k$ is a minor of $\mathbf{H}$, a contradiction.

Case 3. $f([I]^k) = k$. This means that every k-element subset of I has a common neighbor in $B - B_r$. Let $X = \{x_1, \ldots, x_{k-1}\}$ be a collection of $k-1$ distinct vertices in I. As every k-element subset of I has a common neighbor, there is a function $h : (I - X) \to (B - B_r)$ such that $h(y)$ is a common neighbor of $X \cup \{y\}$. If the range of h contains $k-1$ distinct elements, $\mathbf{H}$ contains $\mathbf{K}_{k-1,k-1}$ as a subgraph and therefore $\mathbf{K}_k$ as a minor. We may, therefore, assume that the range of h has fewer than $k-1$ elements. Thus, there is a $J \subseteq I - X$ with $|J| \geq |I - X|/(k-2)$ on which h is constant. Let $z \in B$ be the element to which h maps J. We let $A_{r+1} = J \cup X$ and $B_{r+1} = B_r \cup \{z\}$. Observe that z is a common neighbor of all elements in A_{r+1}, and that

$$|A_{r+1}| \geq |X| + |I - X|/(k-2),$$

which is at least

$$(k-1) + b^{k-2-(r+1)}(m) - 1 > b^{k-2-(r+1)}(m)$$

as required.

To complete the proof, we need to verify that the number of iterations does not reach $k-1$. Note that the iteration is repeated only in case 3, and in this case B_{r+1} contains one more element than B_r. If the set were to contain $k-1$ elements, as all these elements are neighbors of all elements in A', which has at least $m \geq k$ elements, we would have that $\mathbf{H}$ contains $\mathbf{K}_{k-1,k-1}$, and therefore $\mathbf{K}_k$ as a minor. This establishes that $|B'| < k - 1$. □

The main combinatorial result of this article can now be proved by a construction that iterates Lemma 5.2. For a fixed large graph $\mathbf{G} = (V, E^{\mathbf{G}})$, we proceed inductively and generate two sequences of sets of vertices

$$V = S_0 \supseteq S_1 \supseteq \cdots \supseteq S_i$$
$$\emptyset = Z_0 \subseteq Z_1 \subseteq \cdots \subseteq Z_i,$$

where S_i is an i-scattered set in $\mathbf{G} - Z_i$. Once we have S_i, we can produce an $(i+1)$-scattered set $S_{i+1} \subseteq S_i$ by viewing the i-neighborhoods of a certain subset of S_i on one side of a bipartite graph, and the vertices of $\mathbf{G} - Z_i$ that are adjacent

to those neighborhoods on the other. Lemma 5.2 guarantees a large enough $(i+1)$-scattered set after removing a few more points which are then added to Z_i to obtain Z_{i+1}. Choosing which points of S_i to put on the bipartite graph requires one more application of Ramsey's Theorem. The technical details follow.

THEOREM 5.3. *For every $k \geq 1$, $d \geq 0$, and $m \geq 0$, there is an $N \geq 0$ such that if $\mathbf{G} = (V, E^{\mathbf{G}})$ is a graph such that $\mathbf{K}_k$ is not a minor of $\mathbf{G}$ and $|V| > N$, then there are sets $S \subseteq V$ and $Z \subseteq V$ with $|S| > m$ and $|Z| < k-1$ such that S is d-scattered in $\mathbf{G} - Z$.*

PROOF. Once again, we prove the statement for $k \geq 2$, as the case $k = 1$ is trivial. Define the function

$$c(n) = r(2, 2, b^{k-2}(n)),$$

where b is the function defined in the proof of Lemma 5.2 and r is the Ramsey function. Let $N = c^d(m)$. We construct Z and S in d stages:

$$S_0 \supseteq S_1 \supseteq \cdots \supseteq S$$
$$Z_0 \subseteq Z_1 \subseteq \cdots \subseteq Z,$$

The sets Z_i and S_i at stage i will be such that $|Z_i| < k-1$ and S_i is i-scattered in $\mathbf{G} - Z_i$. Moreover, $|S_i| > c^{d-i}(m)$. Start with $S_0 = V$ and $Z_0 = \emptyset$.

Suppose that Z_i and S_i have already been constructed. We construct Z_{i+1} and S_{i+1}. For every $u \in S_i$, let $N_i(u)$ be the i-neighborhood of u in $\mathbf{G} - Z_i$. Consider the graph whose set of vertices is the set of neighborhoods $\{N_i(u) : u \in S_i\}$, and whose edges connect two different neighborhoods $N_i(u)$ and $N_i(v)$ if there exist $u' \in N_i(u)$ and $v' \in N_i(v)$ such that $\{u', v'\}$ is an edge in $\mathbf{G} - Z_i$. The number of vertices of this graph is

$$|S_i| > c^{d-i}(m) = r(2, 2, b^{k-2}(c^{d-i-1}(m))).$$

By the graph version of Ramsey's Theorem discussed before, this graph contains either an independent set or a clique of more than $b^{k-2}(c^{d-i-1}(m))$ elements. The existence of such a clique implies a $\mathbf{K}_k$ minor in $\mathbf{G}$ since the i-neighborhoods of elements in S_i are disjoint and connected in $\mathbf{G} - Z_i$. Therefore, there must be an independent set, say $\{N_i(u) : u \in I\}$, where $I \subseteq S_i$ and

$$|I| > b^{k-2}(c^{d-i-1}(m)).$$

We define a bipartite graph $\mathbf{H} = (A \cup B, E \subseteq A \times B)$ on which to apply Lemma 5.2. Let $A = I$, and let B be the set of vertices of $\mathbf{G} - Z_i$ that are adjacent to some vertex in $\bigcup_{u \in I} N_i(u)$. By the choice of I, the sets A and B are disjoint. The edges of $\mathbf{H}$ connect vertices $u \in A$ with those vertices $v \in B$ that are adjacent to some vertex in $N_i(u)$. Clearly, $\mathbf{H}$ has no $\mathbf{K}_k$ minor; otherwise $\mathbf{G}$ would also have one since the i-neighborhoods of elements in I form disjoint connected patches in $\mathbf{G} - Z_i$. By Lemma 5.2, there exist $A' \subseteq A$ and $B' \subseteq B$ with $|A'| > c^{d-i-1}(m)$ such that $A' \times B' \subseteq E$ and A' is 1-scattered in $\mathbf{H} - B'$. Let $Z_{i+1} = Z_i \cup B'$ and $S_{i+1} = A'$, which is $(i+1)$-scattered in $\mathbf{G} - Z_{i+1}$. The proof will be complete by showing that if $|Z_{i+1}| \geq k-1$, then $\mathbf{G}$ has a $\mathbf{K}_{k-1,k-1}$ minor, and thus a $\mathbf{K}_k$ minor.

Suppose that $|Z_{i+1}| \geq k-1$. By construction, $A' \times B' \subseteq E$, which means that, in $\mathbf{G}$, each $b \in B'$ is adjacent to some vertex in $N_i(a)$ for every $a \in A'$. In fact, the inductive construction guarantees that each $b \in Z_i$ is also adjacent, in $\mathbf{G}$,

to some vertex $N_i(a)$ for every $a \in A'$. Consider each $N_i(u)$, with $u \in A'$, as a connected patch in the subgraph of $\mathbf{G}$ induced by $\bigcup_{u \in A'} N_i(u)$ and Z_{i+1}. Note that these patches are disjoint. The $\mathbf{K}_{k-1,k-1}$ minor is now clear since $|A'| \geq k-1$ and $|Z_{i+1}| \geq k-1$. □

Combining this with Corollary 3.3, we get the following result.

THEOREM 5.4. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the class of Gaifman graphs of structures in $\mathcal{C}$ excludes at least one minor. On the class $\mathcal{C}$, every query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

We now comment on the relationship between Theorem 5.4 and the earlier Theorems 4.4 and 3.5.

As noted earlier, the class $\mathcal{T}(k)$ of graphs of treewidth less than k excludes $\mathbf{K}_{k+1}$ as a minor. Thus, the homomorphism-preservation theorem for these classes (Theorem 4.4) is a special case of Theorem 5.4. Furthermore, there are many classes characterized by excluded minors that do not have bounded treewidth. An example is the collection of planar graphs, which, by Kuratowski's Theorem, exclude $\mathbf{K}_5$ and $\mathbf{K}_{3,3}$ as minor, but have unbounded treewidth. Another example of a class of graphs that exclude some minor are the graphs of bounded genus. Indeed, any class of graphs closed under taking minors and different from the class of all finite graphs must exclude some minor; consequently, the preservation-under-homomorphisms property holds for all these classes.

A more precise relationship between Theorems 4.4 and 5.4 can be obtained using certain deep results by Robertson and Seymour [1986] about classes of graphs excluding a minor. Specifically, Robertson and Seymour [1986] showed that for every graph $\mathbf{H}$, the class of graphs excluding $\mathbf{H}$ as a minor is of bounded treewidth if and only if $\mathbf{H}$ is planar (this result is a consequence of the the Excluded Grid Theorem of Robertson and Seymour [1986]—see also Diestel [1997, Theorem 12.4.3]). Consequently, for every graph $\mathbf{H}$, the preservation-under-homomorphisms property for the class of graphs excluding $\mathbf{H}$ as a minor can be derived from Theorem 5.4, but not from Theorem 4.4, precisely when $\mathbf{H}$ is a nonplanar graph.

It should also be noted that a class of graphs of bounded degree need not exclude any minor. This can be seen by replacing every node of a $\mathbf{K}_k$ by a binary tree with $k-1$ leaves and connecting different pairs of trees through disjoint pairs of leaves. The resulting graph has degree 3, but has $\mathbf{K}_k$ as a minor. Therefore, Theorem 3.5 can not be derived as a consequence of Theorem 5.4.

6. *Boolean Queries and Cores*

We stated Theorems 3.5, 4.4 and 5.4 for queries of arbitrary arity even though the proofs were based on notions of minimal models defined for Boolean queries. In this section we explain why the results extend to non-Boolean queries. We then show that, if we consider Boolean queries only, the preservation property can be shown for wider classes of structures than those considered in Theorems 3.5, 4.4 and 5.4.

6.1. NON-BOOLEAN QUERIES. Suppose $\mathcal{C}$ is a class of finite σ-structures and q is an n-ary query on $\mathcal{C}$. We say that q is preserved under homomorphisms on $\mathcal{C}$ if,

for any $\mathbf{A}, \mathbf{B} \in \mathcal{C}$ and any n-tuple $\mathbf{a}$ of elements from $\mathbf{A}$ if $\mathbf{a} \in q(\mathbf{A})$ and $h : \mathbf{A} \to \mathbf{B}$ is a homomorphism, then $h(\mathbf{a}) \in q(\mathbf{B})$. In particular, if q is a Boolean query on $\mathcal{C}$, q is preserved under homomorphisms on $\mathcal{C}$ if for every pair of structures $\mathbf{A}$ and $\mathbf{B}$ in $\mathcal{C}$, if there is a homomorphism h from $\mathbf{A}$ to $\mathbf{B}$ and $q(\mathbf{A}) = 1$, then $q(\mathbf{B}) = 1$.

There is a natural way to turn a non-Boolean query into a Boolean query in a vocabulary expanded with constants. Let σ' be the vocabulary obtained by extending σ with n new constant symbols $c_1, \ldots, c_n$ and $\mathcal{C}'$ be the class of all σ'-structures $\mathbf{A}$ whose restriction $\mathbf{A}|_\sigma$ to the vocabulary σ is in $\mathcal{C}$. Similarly, let q' be the *Boolean* query on $\mathcal{C}'$ defined by $q'(\mathbf{A}) = 1$ if and only if $\mathbf{c}^{\mathbf{A}} \in q(\mathbf{A}|_\sigma)$ where $\mathbf{c}^{\mathbf{A}}$ is the n-tuple of elements in $\mathbf{A}$ interpreting the constants $c_1, \ldots, c_n$.

It is easily verified that q is preserved under homomorphisms on $\mathcal{C}$ if, and only if, q' is preserved under homomorphisms on $\mathcal{C}'$ (a homomorphism on structures interpreting constant symbols is also required to preserve the interpretation of constants, that is, if $h : \mathbf{A} \to \mathbf{B}$ is a homomorphism, then $h(c^{\mathbf{A}}) = c^{\mathbf{B}}$). Moreover, for a σ' structure $\mathbf{A}$, the Gaifman graph $\mathcal{G}(\mathbf{A}|_\sigma)$ is identical to $\mathcal{G}(\mathbf{A})$. Thus, $\mathcal{C}$ has bounded degree or bounded treewidth or excludes a given minor if and only if $\mathcal{C}'$ does. Moreover, if q' is definable on $\mathcal{C}'$ by an existential positive sentence ψ, then there is an existential positive formula defining q on $\mathcal{C}$. This formula is obtained by replacing the constants $c_1, \ldots, c_n$ by new variables $x_1, \ldots, x_n$. Thus, if the homomorphism preservation theorem holds for Boolean queries on $\mathcal{C}'$, it holds for n-ary queries on $\mathcal{C}$. However, in our proofs above we also require that the classes of structures we consider are closed under taking substructures and disjoint unions. Unfortunately, these are properties that do not transfer from $\mathcal{C}$ to $\mathcal{C}'$. Due to the additional constants, the latter may fail to have these closure properties even when the former has them.

To get around this problem, we use the notion of a *plebian companion* of a structure introduced by Ajtai and Gurevich [1994]. We give a brief description of their construction. Suppose σ' is a vocabulary including the constant symbols $c_1, \ldots, c_n$ and let $\mathbf{A}$ be a σ'-structure. The plebian companion of $\mathbf{A}$ is a structure $p\mathbf{A}$ in a vocabulary ρ obtained from σ' as follows. Every relation symbol R in σ' is also in ρ but ρ does not contain any of the constants. In addition, for each relation symbol R of arity r and each non-empty partial function $m : \{1, \ldots, r\} \rightharpoonup \{c_1, \ldots, c_n\}$, ρ contains a new relation symbol R_m whose arity is $r - j$ where j is the number of elements of $\{1, \ldots, r\}$ on which m is defined. In particular, if m is total, $r = j$ and R_m is then a 0-ary relation symbol. That is to say, it is a Boolean symbol that is interpreted as either true or false in any ρ-structure.

The plebian companion $p\mathbf{A}$ of $\mathbf{A}$ is a ρ-structure whose universe is obtained from that of $\mathbf{A}$ by excluding the interpretation of the constants. For each relation symbol R in σ', the interpretation of R in $p\mathbf{A}$ is the restriction of $R^{\mathbf{A}}$ to the universe of $p\mathbf{A}$. To define the interpretation of R_m, let $\mathbf{a}$ be an $r - j$ tuple of elements from $p\mathbf{A}$. Let $\mathbf{a}'$ be the r-tuple of elements of $\mathbf{A}$ obtained from $\mathbf{a}$ by inserting in position i the element interpreting the constant $m(i)$. We say that $\mathbf{a} \in R_m^{p\mathbf{A}}$ if and only if $\mathbf{a}' \in R^{\mathbf{A}}$. In the special case that R_m is 0-ary, we say that it is interpreted as true if and only if the unique empty tuple is in R_m by the above rule.

It is straightforward to show that for any σ'-formula φ there is a ρ-formula ψ such that $p\mathbf{A} \models \psi$ if and only if $\mathbf{A} \models \varphi$. Indeed, ψ is obtained by φ by replacing each atomic formula $R(\bar{t})$ in which the tuple of terms $\bar{t}$ contains constants, by the formula $R_m(\bar{x})$ where $\bar{x}$ is obtained from $\bar{t}$ by removing the constants and m is the

partial function that maps i to the constant occurring in position i in $\bar{t}$. It is easily seen that if φ is existential positive, then so is ψ. There is a similarly straightforward translation in the other direction, which also preserves existential positive formulas. We can now make three useful observations about plebian companions.

OBSERVATION 6.1. *The Gaifman graph $\mathcal{G}(p\mathbf{A})$ is a subgraph of $\mathcal{G}(\mathbf{A})$.*

Indeed, $\mathcal{G}(p\mathbf{A})$ is the subgraph of $\mathcal{G}(\mathbf{A})$ induced by the elements that are not named by a constant. Writing $p\mathcal{C}'$ for the collection of plebian companions of the structures in $\mathcal{C}'$, we see that one consequence of the above observation is that $p\mathcal{C}'$ has bounded degree or bounded treewidth or excludes some minor if $\mathcal{C}$ does.

OBSERVATION 6.2. *There is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$ if, and only if, there is a homomorphism from $p\mathbf{A}$ to $p\mathbf{B}$.*

To see that this holds, let h be a homomorphism from $p\mathbf{A}$ to $p\mathbf{B}$. We can extend h to a map $\hat{h}$ from $\mathbf{A}$ to $\mathbf{B}$ by letting $\hat{h}(c^{\mathbf{A}}) = c^{\mathbf{B}}$ for all constants c. Clearly, if $\mathbf{a}$ is a tuple from $\mathbf{A}$ that does not include the interpretation of any of the constants, then for any relation R in σ, $R^{\mathbf{A}}(\mathbf{a}) \Rightarrow R^{p\mathbf{A}}(\mathbf{a}) \Rightarrow R^{p\mathbf{B}}(h(\mathbf{a})) \Rightarrow R^{\mathbf{B}}(\hat{h}(\mathbf{a}))$, since $\hat{h}(\mathbf{a}) = h(\mathbf{a})$. On the other hand, if $\mathbf{a}$ contains constants, let m be the partial function that maps i to the constant occurring in position i and $\mathbf{a}'$ be the tuple obtained from $\mathbf{a}$ by removing the elements named by constants. Since $\hat{h}$ maps $c^{\mathbf{A}}$ to $c^{\mathbf{B}}$ for each constant c, it is easily seen that $\hat{h}(\mathbf{a})$ is the tuple obtained from $\hat{h}(\mathbf{a}')$ by inserting in position i the element $(m(i))^{\mathbf{B}}$. Since, furthermore $\hat{h}(\mathbf{a}')$ is the same as $h(\mathbf{a}')$, we have the following implications: $R^{\mathbf{A}}(\mathbf{a}) \Rightarrow R_m^{p\mathbf{A}}(\mathbf{a}') \Rightarrow R_m^{p\mathbf{B}}(h(\mathbf{a}')) \Rightarrow R^{\mathbf{B}}(\hat{h}(\mathbf{a}))$, establishing that $\hat{h}$ is a homomorphism.

For the other direction, suppose g is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$. We wish to show that the restriction of g to the universe of $p\mathbf{A}$ is a homomorphism from $p\mathbf{A}$ to $p\mathbf{B}$. For any relation symbol R in σ, it is obvious that $R^{p\mathbf{A}}(\mathbf{a}) \Rightarrow R^{p\mathbf{B}}(g(\mathbf{a}))$ just by the fact that g is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$. Now, if R_m is a new symbol in ρ and $\mathbf{a}$ is a tuple such that $R_m^{p\mathbf{A}}(\mathbf{a})$, let $\mathbf{a}'$ be the tuple obtained from $\mathbf{a}$ by inserting in position i the element $(m(i))^{\mathbf{A}}$. Then, we have $R_m^{p\mathbf{A}}(\mathbf{a}) \Rightarrow R^{\mathbf{A}}(\mathbf{a}')$ by the definition of $R_m^{p\mathbf{A}}$, $R^{\mathbf{A}}(\mathbf{a}') \Rightarrow R^{\mathbf{B}}(g(\mathbf{a}'))$ by the fact that g is a homomorphism and $R^{\mathbf{B}}(g(\mathbf{a}')) \Rightarrow R_m^{p\mathbf{B}}(g(\mathbf{a}))$ by the definition of $R_m^{p\mathbf{B}}$ and the fact that g preserves the interpretation of constants.

Finally, the following observation is straightforward.

OBSERVATION 6.3. *If $\mathcal{C}$ is closed under disjoint unions and substructures, then so is $p\mathcal{C}'$.*

Together these observations imply that if the preservation theorem is proved only with respect to Boolean queries for all classes $\mathcal{C}$ of bounded degree, of bounded treewidth or for classes excluding some minor, it is also established for all queries over such classes. For instance, let $\mathcal{C}$ be a class of structures of bounded degree and let φ be a formula, with free variables, that is preserved under homomorphisms on $\mathcal{C}$. Let $p\mathcal{C}'$ be the corresponding class of plebian companions of $\mathcal{C}$ (note that the class depends on the number of free variables in φ). Then, $p\mathcal{C}'$ is also of bounded degree and we have a *sentence* ψ such that for any structure $\mathbf{A} \in \mathcal{C}$ and tuple $\mathbf{a}$ of elements from $\mathbf{A}$, $\mathbf{A} \models \varphi[\mathbf{a}]$ if and only if $p\mathbf{A}' \models \psi$ where $\mathbf{A}'$ is the expansion of $\mathbf{A}$ with constants for all elements in $\mathbf{a}$. Thus, ψ is equivalent to an existential positive sentence on $p\mathcal{C}'$ and by the arguments above, this implies that φ is equivalent to an

existential positive formula on $\mathcal{C}$. This justifies the statement of Theorems 3.5, 4.4 and 5.4 for queries of arbitrary arity.

6.2. CORES. Let q be a Boolean query that is preserved under homomorphisms on all finite σ-structures. The key observation we make is that the minimal models of q are *cores*. The concept of core was introduced in the context of graph theory (see Hell and Nesetril [1992]), but it generalizes naturally to relational structures. A substructure $\mathbf{B}$ of $\mathbf{A}$ is called a *core* of $\mathbf{A}$ if there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$, but, for every proper substructure $\mathbf{B}'$ of $\mathbf{B}$, there is no homomorphism from $\mathbf{A}$ to $\mathbf{B}'$. It can be seen that every finite structure $\mathbf{A}$ has a unique core up to isomorphism, denoted by core($\mathbf{A}$), and that $\mathbf{A}$ is homomorphically equivalent to core($\mathbf{A}$). If a structure $\mathbf{A}$ is its own core, we say that $\mathbf{A}$ is a core. It is now clear from the definitions that if q is a query that is preserved under homomorphisms on all finite σ-structures, then every minimal model of q is a core. More generally, if $\mathcal{C}$ is a class of finite σ-structures closed under substructures, and q is a query preserved under homomorphisms on $\mathcal{C}$, then every minimal model of q in $\mathcal{C}$ is a core.

Now, combining the above observation with Theorem 3.2, we can strengthen Corollary 3.3 so that it is not the structures in a class $\mathcal{C}$ that are required to have the property of low density. It suffices to show that the collection of Gaifman graphs of cores of the structures in $\mathcal{C}$ has this property.

COROLLARY 6.4. *Let $\mathcal{C}$ be a class of finite σ-structures having the following properties:*

(1) *$\mathcal{C}$ is closed under substructures and disjoint unions;*
(2) *for some s and for all d and m, there is an N so that if $\mathbf{A} \in \mathcal{C}$ and* core($\mathbf{A}$) *has more than N elements, then there is a set B of at most s elements such that $\mathcal{G}(\text{core}(\mathbf{A})) - B$ has a d-scattered set of size m.*

On the class $\mathcal{C}$, every Boolean query that is first-order definable and preserved under homomorphisms is definable by an existential positive first-order formula.

Combining this with Lemma 3.4, we obtain a stronger version of Theorem 3.5 specifically for Boolean queries. That is, the following is stronger than Theorem 3.5 in one direction in that it applies to a wider collection of classes of structures, but weaker in another in that it only applies to Boolean queries.

THEOREM 6.5. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the class of cores of structures in $\mathcal{C}$ has bounded degree. On the class $\mathcal{C}$, every Boolean query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

We are able to similarly generalize Theorems 4.4 and 5.4 for the specific case of Boolean queries. More precisely, for every positive integer $k \geq 2$, let $\mathcal{H}(\mathcal{T}(k))$ be the class of all finite σ-structures $\mathbf{A}$ such that the core of $\mathbf{A}$ has treewidth less than k. These classes have been studied in the context of constraint-satisfaction problems in Dalmau et al. [2002] and Grohe [2003]. It is easy to see that for each $k \geq 2$, the class $\mathcal{H}(\mathcal{T}(k))$ coincides with the class of all finite σ-structures that are homomorphically equivalent to a σ-structure of treewidth less than k. In the following, when we say that the structures in a class $\mathcal{C}$ *have cores of bounded treewidth*, we mean that there is a positive integer k such that $\mathcal{C} \subseteq \mathcal{H}(\mathcal{T}(k))$.

THEOREM 6.6. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the structures in $\mathcal{C}$ have cores of bounded treewidth. On the class $\mathcal{C}$, every Boolean query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

In Section 4, we mentioned several natural examples of classes of structures of bounded treewidth. Classes of structures whose cores have bounded treewidth are even more pervasive. For example, the core of every non-trivial bipartite graph is $\mathbf{K}_2$, the graph consisting of a single edge. Hence, the class of bipartite graphs is contained in $\mathcal{H}(\mathcal{T}(2))$. However, all grids are bipartite and have arbitrarily large treewidth. Thus, $\mathcal{T}(2)$ is properly contained in $\mathcal{H}(\mathcal{T}(2))$; in fact, for every $k \geq 2$, we have that $\mathcal{T}(k)$ is properly contained in $\mathcal{H}(\mathcal{T}(k))$. For another example, consider all planar graphs that contain $\mathbf{K}_4$ as a subgraph. By the Four Color Theorem for planar graphs, every such graph is 4-colorable, hence it is homomorphically equivalent to $\mathbf{K}_4$ and so it is contained in $\mathcal{H}(\mathcal{T}(4))$.

Finally, we state the preservation result for Boolean queries and classes of structures whose cores exclude some minor.

THEOREM 6.7. *Let $\mathcal{C}$ be a class of finite σ-structures that is closed under substructures and disjoint unions, and such that the class of Gaifman graphs of cores of structures in $\mathcal{C}$ excludes at least one minor. On the class $\mathcal{C}$, every Boolean query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

Theorem 6.7 subsumes Theorem 6.6 in the same way as Theorem 5.4 subsumes Theorem 4.4, since the Gaifman graphs of cores of structures in $\mathcal{H}(\mathcal{T}(k))$ exclude $\mathbf{K}_{k+1}$ as a minor. The relationship with Theorem 6.5 is less clear. At the end of Section 5 above, we presented an example of a class of structures that has bounded degree but does not exclude any minors. However, the structures involved are not cores. If we could construct a class of cores of bounded degree which nevertheless do not exclude any minor, this would show that Theorems 6.5 and 6.7 are similarly incomparable.

It is not clear whether Theorems 6.5, 6.6 and 6.7 can be extended to non-Boolean queries. All we can say is that the method of plebian companions (from Section 6.1) does not give the desired outcome. To understand why this is the case, recall that we define for any class $\mathcal{C}$ and any n the class $\mathcal{C}'$ of expansions of structures in $\mathcal{C}$ by n constants and then the class $p\mathcal{C}'$ of plebian companions of structures in $\mathcal{C}'$. Since the Gaifman graphs of structures in $\mathcal{C}'$ are the same as those of the corresponding graphs in $\mathcal{C}$ we know that restrictions on the latter also apply to the former. However, it is not the case that the cores of structures in $\mathcal{C}'$ are cores of structures in $\mathcal{C}$. It is possible that the cores of structures in $\mathcal{C}$ have bounded degree (for instance) while the cores of structures in $\mathcal{C}'$ do not. This is illustrated by the following example.

Let a *wheel* be a graph $\mathbf{W}_n$ (for $n \geq 3$) with vertices $h, c_1, \ldots, c_n$ and edges connecting $c_1, \ldots, c_n$ in a simple cycle along with an edge from h (the hub) to each c_i. It is easily seen that, $\mathbf{W}_n$ is 4-colorable and, if n is odd, $\mathbf{W}_n$ is a core. Let a *bicycle* be a graph of the form $\mathbf{B}_n = \mathbf{W}_n + \mathbf{K}_4$, where $n \geq 3$. That is, $\mathbf{B}_n$ is the disjoint union of $\mathbf{W}_n$ and $\mathbf{K}_4$ (note that, as $\mathbf{K}_4$ is the same as $\mathbf{W}_3$, a bicycle consists of two wheels). From the fact that $\mathbf{W}_n$ is 4-colorable, it is clear that the core of $\mathbf{B}_n$ is $\mathbf{K}_4$. Thus, if $\mathcal{C}$ is the class of all bicycles, the cores of structures in $\mathcal{C}$ have bounded

degree. Consider now $(\mathbf{B}_n, h)$, the expansion of $\mathbf{B}_n$ with a constant naming the hub h of $\mathbf{W}_n$. Since any homomorphism of this structure must fix h and $\mathbf{W}_n$ is itself a core when n is odd, it follows for odd $n \geq 5$, we have that $(\mathbf{B}_n, h)$ is itself a core and it contains a node of degree n. Thus, if $\mathcal{C}'$ is the class of expansions of structures in $\mathcal{C}$ by one constant, the class of cores of structures in $\mathcal{C}'$ has unbounded degree.

7. *Ajtai–Gurevich Theorem Revisited*

The Ajtai–Gurevich Theorem [Ajtai and Gurevich 1994] asserts that every Datalog program that is first-order definable on finite structures is *bounded*, that is, the associated monotone operator reaches its least fixed-point after a uniformly bounded number of iterations on every finite structure. The aim of this section is to present a proof of this theorem that is based on the results about treewidth in Section 4. Our proof of the Ajtai–Gurevich Theorem can be construed as a re-intepreration of the original proof that makes explicit the role of bounded treewidth and exposes the components of the original argument. Moreover, we obtain a stronger result for a family of infinitary logics that taken together are strictly more expressive than Datalog. This stronger result, however, is weaker than the result claimed in the preliminary version of this article [Atserias et al. 2004], which appeared in the PODS 2004 Proceedings. In this section, we will also spell out the precise differences between what was claimed in Atserias et al. [2004] and what is actually established here.

7.1. PROOF OF THE AJTAI–GUREVICH THEOREM. The collection of *infinitary formulas* $\mathrm{L}_{\infty\omega}$ is obtained by closing the atomic formulas under negation, infinitary conjunctions, infinitary disjunctions, universal quantification, and existential quantification. For every positive integer k, the k-variable fragment of $\mathrm{L}_{\infty\omega}$, denoted by $\mathrm{L}^k_{\infty\omega}$, consists of all $\mathrm{L}_{\infty\omega}$ formulas with at most k distinct variables; note that each variable may have an unbounded number of occurrences in a $\mathrm{L}^k_{\infty\omega}$-formula. The collection of *existential positive infinitary formulas* $\exists\mathrm{L}^+_{\infty\omega}$ is obtained by closing the atomic formulas under infinitary conjunctions, infinitary disjunctions, and existential quantification. The k-variable fragment of $\exists\mathrm{L}^+_{\infty\omega}$ is denoted by $\exists\mathrm{L}^{k,+}_{\infty\omega}$. From Section 2, recall that a k-Datalog program is a Datalog program in which every rule has at most k distinct variables. It was shown in Kolaitis and Vardi [2000] that for every positive integer k, every k-Datalog query is expressible in $\exists\mathrm{L}^{k,+}_{\infty\omega}$. As a matter of fact, Theorem 4.3 in Kolaitis and Vardi [2000] asserts that k-Datalog is contained in a certain fragment of the existential positive infinitary logic $\exists\mathrm{L}^+_{\infty\omega}$ that we describe next.

For every positive integer k, let CQ^k be the collection of all first-order formulas that have at most k distinct variables and are obtained from atomic formulas using conjunction and existential quantification only; note that each variable may be reused in a CQ^k-formula, so its number of occurrences may be arbitrarily large. Clearly, every CQ^k-formula ψ defines a conjunctive query, since, by transforming ψ to a formula in prenex normal form, we obtain an expression of the form $\exists x_1 \cdots \exists x_n \theta$, where $n \geq k$ and θ is a conjunction of atomic formulas. As an example, the expression

$$\exists x_1 \exists x_2 (E(x_1, x_2) \wedge (\exists x_1 (E(x_2, x_1) \wedge \exists x_2 E(x_1, x_2))))$$

is a CQ^2-formula that is logically equivalent to the conjunctive query

$$\exists x_1 \exists x_2 \exists x_3 \exists x_4 (E(x_1, x_2) \wedge E(x_2, x_3) \wedge E(x_3, x_4)),$$

which asserts that there is a path of length 4.

Next, let $\exists\mathrm{FO}^{k,+}$ be the first-order fragment of $\exists\mathrm{L}^{k,+}_{\infty\omega}$, that is, $\exists\mathrm{FO}^{k,+}$ is the collection of all first-order formulas that have at most k distinct variables and are obtained from atomic formulas using conjunction, disjunction, and existential quantification. Since conjunctions distribute over disjunctions and since existential quantifiers commute with disjunctions, it is clear that every $\exists\mathrm{FO}^{k,+}$-formula is logically equivalent to a finite disjunction $\bigvee_{i=1}^{m} \psi_m$ of CQ^k-formulas.

Finally, let $\bigvee \mathrm{CQ}^k$ be the collection of all disjunctions (finite and infinite) of CQ^k-formulas, that is, $\bigvee \mathrm{CQ}^k$ consists of all $\exists\mathrm{L}^{k,+}_{\infty\omega}$-formulas of the form $\bigvee \Phi$, where Φ is a (possibly infinite) set of CQ^k-formulas. Thus, $\exists\mathrm{FO}^{k,+}$ has the same expressive power as the fragment of $\bigvee \mathrm{CQ}^k$ consisting of all formulas of the form $\bigvee \Phi$, where Φ is a finite set of CQ^k-formulas.

The connection between k-Datalog and k-variable logics can now be stated as follows (see Kolaitis and Vardi [2000, Theorem 4.3]):

THEOREM 7.1. *Let k be a positive integer and π a k-Datalog program.*

(1) *For each positive integer m, the m-th stage of the monotone operator associated with π is definable by a finite disjunction of CQ^k-formulas.*

(2) *The query expressed by π is $\bigvee \mathrm{CQ}^k$-definable. Specifically, if θ_m is a finite disjunction of CQ^k-formulas defining the m-th stage of the monotone operator associated with π, then the query expressed by π is definable by the $\bigvee \mathrm{CQ}^k$-formula $\bigvee_{m\geq 1} \theta_m$.*

The preceding Theorem 7.1 implies that, as regards expressive power, Datalog is contained in the family of infinitary logics $\bigvee \mathrm{CQ}^k$, $k \geq 1$. It is easy to see that this containment is a proper one, since every Datalog query is polynomial-time computable, while even $\bigvee \mathrm{CQ}^2$ can express non-recursive queries. Specifically, for every $n \geq 2$, let ψ_n be a CQ^2-sentence asserting that "there is a path of length n". Then, if S is a nonrecursive set of positive integers, the $\bigvee \mathrm{CQ}^2$-sentence $\bigvee_{n\in S} \psi_n$ defines a Boolean query that is not expressible in Datalog.

We also need a connection between CQ^k-sentences and structures of treewidth less than k. This was first obtained in Kolaitis and Vardi [2000, Remark 5.3] and further refined in Dalmau et al. [2002, Theorem 12]. We state this connection in the next lemma and include its proof for completeness.

LEMMA 7.2. *If k is a positive integer and φ is an CQ^k-sentence, then there is a structure $\mathbf{D}$ of treewidth less than k such that the canonical conjunctive query $\varphi_{\mathbf{D}}$ of $\mathbf{D}$ is logically equivalent to φ.*

PROOF. Assume that φ is a CQ^k-sentence. Let ψ be the result of renaming all occurrences of variables in φ so that each existential quantifier binds a different variable. Repeatedly apply the following rewriting rules to the subformulas of ψ: replace subformulas of the form $\psi' \wedge (\exists x)(\psi'')$ by $(\exists x)(\psi' \wedge \psi'')$, and subformulas of the form $(\exists x)(\psi') \wedge \psi''$ by $(\exists x)(\psi' \wedge \psi'')$. Note that these rules preserve equivalence because each variable is quantified only once in ψ. The result is a conjunctive query

$(\exists x_1)\cdots(\exists x_n)\theta$ that is equivalent to ψ, where θ is a conjunction of atomic facts. Let $\mathbf{D}$ be the canonical structure associated with the conjunctive query $(\exists x_1)\cdots(\exists x_n)\theta$, which means that the universe of $\mathbf{D}$ is the set $\{x_1, \ldots, x_n\}$, and $(x_{i_1}, \ldots, x_{i_r}) \in R^{\mathbf{D}}$ if, and only if, the atomic formula $R(x_{i_1}, \ldots, x_{i_r})$ appears in θ. By construction, the canonical conjunctive query $\varphi_{\mathbf{D}}$ of $\mathbf{D}$ is $(\exists x_1)\cdots(\exists x_n)\theta$, hence it is logically equivalent to φ.

It remains to show that $\mathbf{D}$ has treewidth less than k. Let $\psi_1, \psi_2, \ldots, \psi_r$ be the collection of all subformulas of ψ. View them as nodes of the parse-tree of ψ. Label each node ψ_i of the tree by the set of free variables of ψ_i. Since φ has k variables in total, each ψ_i has at most k free variables, so each label has size at most k. Using the fact that each variable is quantified exactly once in ψ and that each atomic fact of $\mathbf{D}$ is a subformula of ψ, it is not hard to see that the tree and its labeling form a tree-decomposition of $\mathbf{D}$ of width at most $k-1$. Hence, the treewidth of $\mathbf{D}$ is less than k. □

The next lemma establishes a connection between minimal models of $\bigvee \mathrm{CQ}^k$-sentences and structures of treewidth less than k.

LEMMA 7.3. *Let k be a positive integer, let ψ be a $\bigvee \mathrm{CQ}^k$-sentence, and let $\mathbf{A}$ be a model of ψ. There exists a structure $\mathbf{B}$ having the following properties:*

(1) *$\mathbf{B}$ is a minimal model of ψ;*

(2) *the treewidth of $\mathbf{B}$ is less than k;*

(3) *there is a homomorphism from $\mathbf{B}$ to $\mathbf{A}$.*

Furthermore, if $\mathbf{A}$ is a minimal model of ψ, then there is a surjective homomorphism from $\mathbf{B}$ to $\mathbf{A}$.

PROOF. Let ψ be a $\bigvee \mathrm{CQ}^k$-sentence of the form $\bigvee \Phi$, where Φ is a set of CQ^k-sentences. If $\mathbf{A}$ is a model of ψ, then there is a CQ^k-sentence $\varphi \in \Phi$ such that $\mathbf{A} \models \varphi$. By Lemma 7.2, there is a structure $\mathbf{D}$ of treewidth less than k such that φ is logically equivalent to the canonical conjunctive query $\varphi_{\mathbf{D}}$ of $\mathbf{D}$. Consequently, $\mathbf{A} \models \varphi_{\mathbf{D}}$, which, by Theorem 2.1, implies that there is a homomorphism h from $\mathbf{D}$ to $\mathbf{A}$. Since $\mathbf{D}$ is a model of φ, it is also a model of ψ; consequently, there is a substructure $\mathbf{B}$ of $\mathbf{D}$ that is a minimal model of ψ. The treewidth of $\mathbf{B}$ is less than k, since $\mathbf{B}$ is a substructure of $\mathbf{D}$ and the treewidth of $\mathbf{D}$ is less than k. Moreover, the restriction h' of h on $\mathbf{B}$ is a homomorphism from $\mathbf{B}$ to $\mathbf{A}$.

The image $h'(\mathbf{B})$ of $\mathbf{B}$ under h' is a substructure of $\mathbf{A}$; moreover, it is a model of ψ, since $\bigvee \mathrm{CQ}^k$-formulas are preserved under homomorphisms. It follows that if $\mathbf{A}$ is a minimal model of ψ, then $h'(\mathbf{B}) = \mathbf{A}$, which means that h' is an onto homomorphism from $\mathbf{B}$ to $\mathbf{A}$. □

The preceding Lemma 7.3 shows that every minimal model of a $\bigvee \mathrm{CQ}^k$-sentence ψ is the homomorphic image of a minimal model ψ of treewidth less than k. In the preliminary version of this article [Atserias et al. 2004, Lemma 4], we asserted that *every* minimal model of a $\bigvee \mathrm{CQ}^k$-sentence has treewidth less than k. This, however, is not true. As a matter of fact, there are CQ^k-sentences that have minimal models of treewidth at least k. For example, let ψ be the CQ^2-sentence $\exists x_1 \exists x_2((E(x_1, x_2) \wedge (\exists x_1(E(x_2, x_1) \wedge (\exists x_2 E(x_1, x_2))))))$, which asserts that there is a path of length three. The directed 3-element cycle $\mathbf{C}_3$ is a minimal model of ψ, but has treewidth 2.

We are now ready to state and prove the main result of this section.

THEOREM 7.4. *Let k be a positive integer and let $\bigvee \Phi$ be a $\bigvee \mathrm{CQ}^k$-sentence, where Φ is a (possibly infinite) set of CQ^k-sentences. The following statements are equivalent:*

(1) *There is a finite subset Ψ of Φ such that $\bigvee \Phi$ is equivalent to $\bigvee \Psi$ on all finite structures.*

(2) *$\bigvee \Phi$ is equivalent to some $\exists \mathrm{FO}^{k,+}$-sentence on all finite structures.*

(3) *$\bigvee \Phi$ is equivalent to some first-order sentence on all finite structures.*

PROOF. The implications (1) $\Rightarrow$ (2) and (2) $\Rightarrow$ (3) are quite obvious. Towards establishing the implication (3) $\Rightarrow$ (1), assume that $\bigvee \Phi$ is a $\bigvee \mathrm{CQ}^k$-sentence that is equivalent to some first-order sentence ψ on all finite structures. We claim that $\bigvee \Phi$ has finitely many non-isomorphic minimal models. Indeed, if $\bigvee \Phi$ had arbitrarily large minimal models, then Lemma 7.3 implies that $\bigvee \Phi$ has arbitrarily large minimal models of treewidth less than k. But then, by Lemma 4.2, for every $d \geq 0$ and $m \geq 0$, and for every sufficiently large minimal model $\mathbf{A}$ of treewidth less than k, there exists $B \subseteq A$ of size at most k such that $\mathbf{A} - B$ has a d-scattered set of size m. Theorem 3.2 implies immediately that $\bigvee \Phi$ is not equivalent to any first-order sentence on finite structures. This establishes that $\bigvee \Phi$ has finitely many non-isomorphic minimal models.

Let $\mathbf{D}_1, \ldots, \mathbf{D}_m$ be a list of all pairwise non-isomorphic minimal models of $\bigvee \Phi$, and, for each $i \leq m$, let $\varphi_{\mathbf{D}_i}$ be the canonical conjunctive query of $\mathbf{D}_i$. Since $\bigvee \Phi$ is preserved under homomorphisms, we have that $\bigvee \Phi$ is equivalent to $\bigvee_{i=1}^m \varphi_{\mathbf{D}_i}$ on finite structures. In particular, we have that $\bigvee_{i=1}^m \varphi_{\mathbf{D}_i}$ logically implies $\bigvee \Phi$ on finite structures. Since every CQ^k-sentence is logically equivalent to a conjunctive query, the fact that $\bigvee_{i=1}^m \varphi_{\mathbf{D}_i}$ logically implies $\bigvee \Phi$ on finite structures amounts to the union of the conjunctive queries $\varphi_{\mathbf{D}_1}, \ldots, \varphi_{\mathbf{D}_m}$ logically implying the union of the conjunctive queries in Φ. Sagiv and Yannakakis [1981] have shown that a union of conjunctive queries logically implies another union of conjunctive queries if and only if every conjunctive query in the first union logically implies some conjunctive query in the second union. It follows that for every $i \leq m$, there is a CQ^k-sentence θ_i in Φ such that $\varphi_{\mathbf{D}_i}$ logically implies θ_i.[1] This yields that $\bigvee_{i=1}^m \varphi_{\mathbf{D}_i}$ logically implies $\bigvee_{i=1}^m \theta_i$, which, in turn, logically implies $\bigvee \Phi$. At the same time, $\bigvee \Phi$ is logically equivalent to $\bigvee_{i=1}^m \varphi_{\mathbf{D}_i}$; consequently, $\bigvee \Phi$ is also logically equivalent to $\bigvee \Psi$, where $\Psi = \{\theta_i : 1 \leq i \leq m\}$. □

Although the preceding Theorem 7.4 was stated and proved for $\bigvee \mathrm{CQ}^k$-sentences, it holds for $\bigvee \mathrm{CQ}^k$-formulas with free variables. This can be shown using the transformation of non-Boolean queries to Boolean queries, as described in Section 6.

The Ajtai–Gurevich Theorem [Ajtai and Gurevich 1994] can now be obtained easily from Theorems 7.1 and 7.4.

[1]This can also be established directly as follows. Fix some $i \leq m$. Since $\mathbf{D}_i \models \varphi_{\mathbf{D}_i}$, we have that $\mathbf{D}_i \models \bigvee \Phi$. Consequently, there is a CQ^k-sentence θ_i in Φ such that $\mathbf{D}_i \models \theta_i$. Since θ_i is logically equivalent to a conjunctive query, Theorem 2.1 tell us that $\varphi_{\mathbf{D}_i}$ logically implies θ_i.

THEOREM 7.5 (AJTAI–GUREVICH THEOREM). *Let π be a Datalog program. The following statements are equivalent:*

(1) *π is bounded, which means that there is a positive integer s such that, on every finite structure, the query expressed by π can be computed within at most s iterations of the monotone operator associated with π.*
(2) *π is first-order definable, which means that there is a first-order formula such that, on every finite structure, it defines the query expressed by π.*

PROOF. The difficult direction is (2) $\Rightarrow$ (1). Let k be the number of variables of the Datalog program π. By Theorem 7.1, the query expressed by π is definable by a $\bigvee \mathrm{CQ}^k$-formula $\bigvee \Phi$. By Theorem 7.4, if there is a first-order formula that defines this query on all finite structures, then there is a finite subset Ψ of Φ such that $\bigvee \Phi$ is logically equivalent to $\bigvee \Psi$ on finite structures. Consequently, there is a positive integer s such that $\bigvee \Phi$ is logically equivalent to the formula θ_s defining the s-th stage of the monotone operator associated with π. It follows that, on every finite structure, the query expressed by π can be computed within at most s iterations of the monotone operator associated with π. □

Note that Theorem 7.4 is a stronger result than Theorem 7.5, since, as detailed in the remarks following Theorem 7.1, the family of infinitary logics $\bigvee \mathrm{CQ}^k$, $k \geq 1$, has strictly higher expressive power than Datalog.

7.2. ON THE RELATIONSHIP BETWEEN THE INFINITARY LOGICS $\bigvee \mathrm{CQ}^k$ AND $\exists \mathrm{L}^{k,+}_{\infty\omega}$. In the remainder of this section, we will examine the relationship between the full existential positive infinitary logic $\exists \mathrm{L}^{k,+}_{\infty\omega}$ with k variables and its fragment $\bigvee \mathrm{CQ}^k$, $k \geq 1$. In a nutshell, the precise relationship between $\exists \mathrm{L}^{k,+}_{\infty\omega}$ and $\bigvee \mathrm{CQ}^k$ is as follows. On the class of all finite structures, every $\exists \mathrm{L}^{k,+}_{\infty\omega}$-sentence is equivalent to an infinitary disjunction of infinitary conjunctions of CQ^k-sentences; as will be seen below, this normal-form theorem for $\exists \mathrm{L}^{k,+}_{\infty\omega}$ can be obtained easily from results in Kolaitis and Vardi [1995]. In the preliminary version of this article, we claimed that on the class of all finite structures, every $\exists \mathrm{L}^{k,+}_{\infty\omega}$-formula is equivalent to a $\bigvee \mathrm{CQ}^k$-formula. Regrettably, this claim turns out to be false because we will show here that there are infinitary conjunctions $\bigwedge \Phi$ of CQ^2-sentences that are not equivalent to any $\bigvee \mathrm{CQ}^2$-sentence. Thus, the aforementioned normal form of $\exists \mathrm{L}^{k,+}_{\infty\omega}$-sentences as infinitary conjunctions of $\bigvee \mathrm{CQ}^k$-sentences is optimal and cannot be simplified.

The expressive power of $\exists \mathrm{L}^{k,+}_{\infty\omega}$ is captured by the *existential k-pebble game*, introduced in Kolaitis and Vardi [1995] and studied further in Kolaitis and Vardi [2000]. This game is played between two players, the Spoiler and the Duplicator, on two σ-structures $\mathbf{A}$ and $\mathbf{B}$ according to the following rules. Each player has a set of k pebbles $\alpha_1, \ldots, \alpha_k$ and $\beta_1, \ldots, \beta_k$ respectively. In each round of the game, the Spoiler can make one of two different types of moves: either he places a free pebble α_i on an element of the domain of $\mathbf{A}$, or he removes a pebble α_i from a pebbled element of $\mathbf{A}$. To each move of the Spoiler, the Duplicator must respond by placing her corresponding pebble β_i over an element of $\mathbf{B}$, or removing her corresponding pebble β_i from $\mathbf{B}$, respectively. If the Spoiler has a strategy to reach a round in which the set of pairs of pebbled elements is not a partial homomorphism between $\mathbf{A}$ and $\mathbf{B}$, then he wins the game. Otherwise, we say that the Duplicator wins the game. The following link between existential k-pebble games and $\exists \mathrm{L}^{k,+}_{\infty\omega}$ was established in Kolaitis and Vardi [1995, Corollary 4.9 and Remark 4.11].

THEOREM 7.6. *Let k be a positive integer, and let $\mathbf{A}$ and $\mathbf{B}$ be two finite σ-structures. The following statements are equivalent.*

(1) *Every $\exists L^{k,+}_{\infty\omega}$-sentence that is true on $\mathbf{A}$ is also true on $\mathbf{B}$.*

(2) *Every $\exists FO^{k,+}$-sentence that is true on $\mathbf{A}$ is also true on $\mathbf{B}$.*

(3) *The Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$.*

As explained earlier in this section, every $\exists FO^{k,+}$-formula is equivalent to a finite disjunction $\bigvee_{i=1}^{m} \psi_m$ of CQ^k-formulas. Consequently, the second statement in the preceding Theorem 7.6 can be replaced by the seemingly weaker statement

$2'$. *Every CQ^k-sentence that is true on $\mathbf{A}$ is also true on $\mathbf{B}$.*

For every positive integer k and every finite σ-structure $\mathbf{A}$, let $q(\mathbf{A}, k)$ be the query: given a finite σ-structure $\mathbf{B}$, does the Duplicator win the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$?

The next result follows easily from Theorem 7.6 and the preceding observation about statement $2'$.

THEOREM 7.7. *Let k be a positive integer.*

(1) *For every finite σ-structure $\mathbf{A}$, the query $q(\mathbf{A}, k)$ is definable by the following infinitary conjunction of CQ^k-sentences*

$$\bigwedge \{\theta : \theta \text{ is an } CQ^k\text{-sentence and } \mathbf{A} \models \theta\}.$$

(2) *On the class of all finite σ-structures, every $\exists L^{k,+}_{\infty\omega}$-sentence φ is equivalent to the following infinitary disjunction*

$$\bigvee \{q(A, k) : \mathbf{A} \text{ is a finite } \sigma\text{-structure and } \mathbf{A} \models \varphi\}.$$

Consequently, on the class of all finite σ-structures, every $\exists L^{k,+}_{\infty\omega}$-sentence is equivalent to an infinitary disjunction of infinitary conjunctions of CQ^k-sentences.

In what follows, we show that the above normal form for $\exists L^{k,+}_{\infty\omega}$ cannot be improved. For this, we need an auxiliary result concerning the definability of the query $q(\mathbf{A}, k)$.

PROPOSITION 7.8. *Let k be a positive integer and let $\mathbf{A}$ be a finite σ-structure. The following statements are equivalent.*

(1) *The query $q(\mathbf{A}, k)$ is $\bigvee CQ^k$-definable on the class of all finite σ-structures.*

(2) *The query $q(\mathbf{A}, k)$ is CQ^k-definable on the class of all finite structures.*

PROOF. The direction (2) $\Rightarrow$ (1) is obvious. For the direction (1) $\Rightarrow$ (2), let us assume that, on the class of all finite σ-structures, the query $q(\mathbf{A}, k)$ is definable by a sentence $\bigvee \Theta$, where Θ is a (possibly infinite) set of CQ^k-sentences. Since $\mathbf{A}$ satisfies the query $q(\mathbf{A}, k)$, there is a CQ^k-sentence $\theta \in \Theta$ such that $\mathbf{A} \models \theta$. We now claim that θ defines the query $q(\mathbf{A}, k)$ on the class of all finite σ-structures. Indeed, if $\mathbf{B}$ is a finite model of θ, then $\mathbf{B} \models \bigvee \Theta$, hence $\mathbf{B}$ satisfies the query $q(\mathbf{A}, k)$. Conversely, if $\mathbf{B}$ is a finite σ-structure such that the Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$, then, by Theorem 7.6, every $\exists L^{k,+}_{\infty\omega}$-sentence satisfied by $\mathbf{A}$ is also satisfied by $\mathbf{B}$; consequently, $\mathbf{B}$ satisfies θ. □

Assume that k is a positive integer and $\mathbf{A}$ is a finite structure whose core has treewidth less than k. In Dalmau et al. [2002], it was shown that for every finite structure $\mathbf{B}$, the Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$ if and only if there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$. It follows that, in this case, the query $q(\mathbf{A}, k)$ is definable by the canonical conjunctive query $\varphi_{\mathbf{A}}$ of $\mathbf{A}$; furthermore, $\varphi_{\mathbf{A}}$ is equivalent to an CQ^k-sentence, since the core of $\mathbf{A}$ has treewidth less than k. This gives a large collection of structures $\mathbf{A}$ for which the query $q(\mathbf{A}, k)$ is CQ^k-definable, hence it is also $\bigvee \mathrm{CQ}^k$-definable. In contrast, the next proposition shows that this need not always be true.

PROPOSITION 7.9. *Let $\mathbf{C}_3$ be the directed 3-element cycle.*

(1) *The query $q(\mathbf{C}_3, 2)$ is not first-order definable.*
(2) *The query $q(\mathbf{C}_3, 2)$ is $\bigwedge \mathrm{CQ}^2$-definable, but is not $\bigvee \mathrm{CQ}^2$-definable.*

PROOF. Let $\mathbf{B}$ be a finite directed graph. It is easy to verify that the Duplicator wins the existential 2-pebble game on $\mathbf{C}_3$ and $\mathbf{B}$ if and only if $\mathbf{B}$ contains a cycle. Indeed, in the existential 2-pebble game on $\mathbf{C}_3$ and $\mathbf{B}$, the Spoiler can force the Duplicator to play along a path. Since $\mathbf{B}$ is finite, the Duplicator can win the existential 2-pebble game only if $\mathbf{B}$ contains a cycle. Conversely, if $\mathbf{B}$ contains a cycle, then the Duplicator can win the existential 2-pebble game on $\mathbf{C}_3$ and $\mathbf{B}$ by playing along edges of a fixed cycle.

It is well known that the query "given a finite directed graph, is it acyclic?" is not first-order definable (this can be shown using Ehrenfeucht-Fraïssé games). Thus, the query $q(\mathbf{C}_3, 2)$ is not first-order definable.

By Theorem 7.7, the query $q(\mathbf{C}_3, 2)$ is $\bigwedge \mathrm{CQ}^2$-definable. In contrast, Proposition 7.8 implies that $q(\mathbf{C}_3, 2)$ is not $\bigvee \mathrm{CQ}^2$-definable, since, if it were, then it would be CQ^2-definable and, hence, first-order definable. □

COROLLARY 7.10. *On the class of all finite directed graphs, $\bigvee \mathrm{CQ}^2$ is strictly less expressive than $\exists \mathrm{L}^{2,+}_{\infty\omega}$.*

As mentioned earlier, Corollary 7.10 refutes our claim in the preliminary version of this article [Atserias et al. 2004, Lemma 5] to the effect that, on the class of all finite structures, for every positive integer k, every $\exists \mathrm{L}^{k,+}_{\infty\omega}$-sentence is equivalent to a $\bigvee \mathrm{CQ}^k$-sentence.

7.3. EXTENSIONS TO STRONGER INFINITARY LOGICS. Since every $\exists \mathrm{L}^{k,+}_{\infty\omega}$-sentence is preserved under homomorphisms, Rossman's [2005] homomorphism-preservation theorem implies that if a $\exists \mathrm{L}^{k,+}_{\infty\omega}$-sentence is equivalent to a first-order sentence on finite structures, then it is also equivalent to an existential-positive first-order sentence on finite structures [Atserias et al. 2004, Theorem 9]. Moreover, since $\bigvee \mathrm{CQ}^k$ is a fragment of $\exists \mathrm{L}^{k,+}_{\infty\omega}$, Rossman's result also implies that if a $\bigvee \mathrm{CQ}^k$-sentence is equivalent to a first-order sentence on finite structures, then it is also equivalent to an existential-positive first-order sentence. However, Rossman's proof does not yield the stronger result established in Theorem 7.4, namely, that if a $\bigvee \mathrm{CQ}^k$-sentence is equivalent to a first-order sentence on finite structures, then it is equivalent to some $\exists \mathrm{FO}^{k,+}$-sentence (i.e., to some existential-positive first-order sentence with at most k distinct variables). Indeed, Rossman's proof produces an equivalent existential-positive first-order sentence with more than k

distinct variables. In turn, this state of affairs gives rise to the following problem, which is open at present.

Problem. Suppose that a $\exists L^{k,+}_{\infty\omega}$-sentence ψ is equivalent to a first-order sentence on the class of all finite structures. Is it true that ψ is equivalent to some $\exists FO^{k,+}$-sentence on the class of all finite structures?

Finally, it is natural to ask whether the Ajtai–Gurevich Theorem and Theorem 7.4 hold for more expressive logics that allow for some form of negation. Ajtai and Gurevich [1994] showed that their theorem fails both for Datalog programs with negated extensional predicates and for Datalog programs with inequalities $\neq$. It follows that Theorem 7.4 fails for extensions of $\bigvee \mathrm{CQ}^k$ that allow for negated atoms or for inequalities $\neq$. Thus, the results presented in this section are very tightly connected to preservation under homomorphisms, and fail for Datalog extensions and for stronger infinitary logics in which sentences are preserved under two-way homomorphisms or one-to-one homomorphisms.

8. *Concluding Remarks*

We have investigated the homomorphism-preservation theorem for numerous classes of finite structures of interest in graph theory and database theory. As noted earlier, preservation theorems do not always relativize to restricted classes of structures, so our results stand by themselves independently of the fact that the homomorphism-preservation theorem has been shown to hold for the class of all finite structures [Rossman 2005]. Indeed, one can ask the same question for other classes of finite structures. For instance, we could consider classes of bounded local treewidth [Epstein 2000; Frick and Grohe 2000] or of bounded cliquewidth [Courcelle et al. 1993]. The homomorphism-preservation theorem for these classes does not follow from our results, as these classes are not definable by excluded minors. Indeed, the classes of bounded local treewidth generalize both bounded treewidth and bounded degree. Also, the class of all cliques has bounded cliquewidth but does not exclude any minor. However, it is worth investigating whether the kinds of techniques we have developed could yield results about these classes.

Another line of investigation would ask similar questions to those studied here for other classical preservation theorems, and in particular, for those that fail on the class of all finite structures, such as the Łoś-Tarski Theorem and Lyndon's Positivity Theorem. The first results in this direction have been reported in Atserias et al. [2005].

It should also be pointed out that our results are effective. More precisely, for the classes of structures for which we established the homomorphism-preservation theorem, the proofs provide us with a computable bound on the size of the minimal models of a first-order query preserved under homomorphisms. This yields an effective procedure to produce a union of conjunctive queries that is equivalent to a given first-order formula that is preserved under homomorphisms. In turn, for classes of structures whose first-order theory is decidable, such as $\mathcal{T}(k)$, the computable bound can also be used to show that it is decidable whether a first-order formula is preserved under homomorphisms. This should be contrasted with the undecidability of the same problem on the class of all finite structures [Alechina and Gurevich 1997]. The exact complexity of these problems on the class $\mathcal{T}(k)$ could be prohibitive, but this remains to be determined.

REFERENCES

ABITEBOUL, S., HULL, R., AND VIANU, V. 1995. *Foundations of Databases*. Addison-Wesley, Reading, MA.

AJTAI, M., AND GUREVICH, Y. 1987. Monotone versus positive. *J. ACM 34*, 1004–1015.

AJTAI, M., AND GUREVICH, Y. 1994. Datalog vs first-order logic. *J. Comput. Syst. Sci. 49*, 562–588.

ALECHINA, N., AND GUREVICH, Y. 1997. Syntax vs semantics on finite structures. In *Structures in Logic and Computer Science*, J. Mycielski, G. Rozenberg, and A. Salomaa, Eds. Lecture Notes in Computer Science, vol. 1261. Springer-Verlag, New York, 14–33.

ATSERIAS, A., DAWAR, A., AND GROHE, M. 2005. Preservation under extensions on well-behaved finite structures. In *Automata, Languages and Programming, 32nd International Colloquium, ICALP 2005*. Lecture Notes in Computer Science, vol. 3580. Springer-Verlag, New York, 1437–1449.

ATSERIAS, A., DAWAR, A., AND KOLAITIS, P. G. 2004. On preservation under homomorphisms and unions of conjunctive queries. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. ACM, New York, 319–329.

CHANDRA, A., AND MERLIN, P. 1977. Optimal implementation of conjunctive queries in relational databases. In *Proceedings of the 9th ACM Symposium on Theory of Computing*. ACM, New York, 77–90.

COURCELLE, B., ENGELFRIET, J., AND ROZENBERG, G. 1993. Handle rewriting hypergraph grammars. *J. Comput. Syst. Sci. 46*, 218–270.

DALMAU, V., KOLAITIS, P. G., AND VARDI, M. Y. 2002. Constraint satisfaction, bounded treewidth, and finite variable logics. In *Proceedings of the 8th International Conference on Principles and Practice of Constraint Programming—CP 2002*. Lecture Notes in Computer Science, vol. 2470. Springer-Verlag, New York, 310–326.

DECHTER, R., AND PEARL, J. 1989. Tree clustering for constraint networks. *Artif. Intel. 38*, 3, 353–366.

DIESTEL, R. 1997. *Graph Theory*. Springer-Verlag, New York.

DOWNEY, R., AND FELLOWS, M. 1999. *Parametrized Complexity*. Springer-Verlag, New York.

EPSTEIN, D. 2000. Diameter and treewidth in minor-closed graph families. *Algorithmica 27*, 275–291.

ERDÖS, P., AND RADO, R. 1960. Intersection theorems for systems of sets. *J. London Math. Soc. 35*, 85–90.

FAGIN, R., KOLAITIS, P. G., AND POPA, L. 2003. Data exchange: Getting to the core. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*. ACM, New York, 90–101.

FEDER, T., AND VARDI, M. 2003. Homomorphism closed vs existential positive. In *Proceedings of the 18th IEEE Symposium on Logic in Computer Science*. IEEE Computer Society Press, Los Alamitos, CA, 311–320.

FRICK, M., AND GROHE, M. 2000. Deciding first-order properties of locally tree-decomposable functions. *J. ACM 48*, 1184–2006.

GAIFMAN, H. 1982. On local and nonlocal properties. In *Logic Colloquium '81*, J. Stern, Ed. North Holland, Amsterdam, The Netherlands, 105–135.

GRAHAM, R. L., ROTHSCHILD, B. L., AND SPENCER, J. H. 1980. *Ramsey Theory*. Wiley, New York.

GROHE, M. 2003. The complexity of homomorphism and constraint satisfaction problems seen from the other side. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science—FOCS 2003*.

GROHE, M., FLUM, J., AND FRICK, M. 2002. Query evaluation via tree-decompositions. *J. ACM 49*, 716–752.

GROHE, M., SCHWENTICK, T., AND SEGOUFIN, L. 2001. When is the evaluation of conjunctive queries tractable? In *Proceedings of the 32nd ACM Symposium on Theory of Computing*. ACM, New York, 657–666.

GUREVICH, Y. 1984. Toward logic tailored for computational complexity. In *Computation and Proof Theory*, M. Richter et al., Eds. Lecture Notes in Mathematics. Springer-Verlag, New York, 175–216.

GUREVICH, Y. 1990. On finite model theory. In *Feasible Mathematics*, S. Buss and P. Scott, Eds. Birkhäuser, 211–219.

HELL, P., AND NESETRIL, J. 1992. The core of a graph. *Discrete Mathematics 109*, 117–126.

HODGES, W. 1993. *Model Theory*. Cambridge University Press, Cambridge, UK.

KOLAITIS, P. G., AND VARDI, M. Y. 1995. On the expressive power of Datalog: Tools and a case study. *J. Comput. Syst. Sci. 51*, 110–134.

KOLAITIS, P. G., AND VARDI, M. Y. 2000. Conjunctive query containment and constraint satisfaction. *J. Comput. Syst. Sci. 61*, 302–332.

KREIDLER, M. 1999. *Strukturinformation bei der Beschreibung monadischer NP-Probleme*. Shaker-Verlag, New York.

KREIDLER, M., AND SEESE, D. 1999. Monadic NP and graph minors. In *CSL'98: Proceedings of the Annual Conference of the European Association for Computer Science Logic*. Lecture Notes in Computer Science, vol. 1584. Springer-Verlag, New York, 126–141.

ROBERTSON, N., AND SEYMOUR, P. 1986. Graph minors. V. Excluding a planar graph. *J. Combinat. Theory, Ser. B 41*, 92–11.

ROSEN, E. 2002. Some aspects of model theory and finite strucrtures. *Bull. Symb. Logic 8*, 380–403.

ROSSMAN, B. 2005. Existential positive types and preservation under homomorphisisms. In *Proceedings of the 20th IEEE Symposium on Logic in Computer Science*. 467–476.

SAGIV, Y., AND YANNAKAKIS, M. 1981. Equivalence between relational expressions with the union and difference operators. *J. ACM 27*, 4, 633–655.

STOLBOUSHKIN, A. 1995. Finite monotone properties. In *Proceedings of the 10th IEEE Symposium on Logic in Computer Science*. IEEE Computer Society Press, Los Alamitos, CA, 324–330.

TAIT, W. W. 1959. A counterexample to a conjecture of Scott and Suppes. *J. Symb. Logic 24*, 15–16.

ULLMAN, J. 1989. Bottom-up beats top-down for Datalog. In *Proceedings of the 8th ACM Symposium on Principles of Database Systems*. ACM, New York, 140–149.

RECEIVED FEBRUARY 2005; ACCEPTED SEPTEMBER 2005