On Preservation under Homomorphisms and Unions of Conjunctive Queries

Albert Atserias[*]
Universitat Politècnica de
Catalunya, Dept. LSI
Jordi Girona Salgado 1-3
08034 Barcelona, Spain
atserias@lsi.upc.es

Anuj Dawar
University of Cambridge
Computer Laboratory
J.J. Thomson Avenue
Cambridge CB3 0FD, U.K.
anuj.dawar@cl.cam.ac.uk

Phokion G. Kolaitis[†]
University of California
Computer Science Dept.
1156 High Street
Santa Cruz, CA 95064, U.S.A
kolaitis@cs.ucsc.edu

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 existential positive formulas are the only first-order formulas (up to logical equivalence) that are preserved under homomorphisms on all structures, finite and infinite. It is a long-standing open problem in finite model theory, however, to determine whether the same homomorphism-preservation result holds in the finite, that is, whether every first-order formula preserved under homomorphisms on finite structures is logically equivalent to an existential positive formula on finite structures. In this paper, we show that the homomorphism-preservation theorem holds for several large 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.

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 [1]). 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 [5], 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 $\mathbf{A} = (A, R_1^{\mathbf{A}}, \dots, R_m^{\mathbf{A}})$ and $\mathbf{B} = (B, R_1^{\mathbf{B}}, \dots, R_m^{\mathbf{B}})$ be two relational structures over the same vocabulary (database schema) $R_1, \dots, R_m$. Recall that a *homomorphism from* $\mathbf{A}$ *to* $\mathbf{B}$ is a function $h : A \to B$ such that for every relation symbol R_i and every tuple $\mathbf{a} = (a_1, \dots, a_r)$ from A, if $\mathbf{a} \in R_i^{\mathbf{A}}$ then $h(\mathbf{a}) = (h(a_1), \dots, h(a_r)) \in R_i^{\mathbf{B}}$. As already realized by Chandra and Merlin [5], 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 for-

[†] Research partially supported by NSF Grant No. IIS-9907419

[*] Partially supported by CICYT TIC2001-1577-C03-02 and the Future and Emerging Technologies programme of the EU under contract number IST-99-14186 (ALCOM-FT).

PODS 2004 June 14-16, 2004, Paris, France.
Copyright 2004 ACM 1-58113-858-X/04/06 . . . $5.00.

mula (see [24]). 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.

Research in finite model theory has shown that, unfortunately, 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 [32, 21]. Similarly, Lyndon's Positivity Theorem is also known to fail in the finite [2, 31]. Although most classical preservation theorems are by now known to fail in the finite, the status of the homomorphism-preservation theorem in the finite has not been settled thus far. In other words, the following problem remains open: suppose a first-order formula is preserved under homomorphisms on the class of all finite structures; is this formula logically equivalent 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; can this query be transformed to an equivalent SPJU query? This problem has received considerable attention in the finite model theory community, where it has been singled out as a central open problem (Problem 5.9 on the finite model theory website at `http://www-mgi.informatik.rwth-aachen.de/FMT/`). It has motivated a lot of research in this area [4, 14, 22, 29], but, in spite of intensive efforts, it has resisted resolution thus far.

Although the homomorphism-preservation problem is about the class of all finite structures over some fixed vocabulary, it is meaningful to ask the same question for first-order formulas preserved under homomorphisms on restricted classes $\mathcal{C}$ of finite structures. Unlike other results in finite model theory that *relativize*, the homomorphism-preservation theorem, whether it holds or fails in the finite, does not relativize to restricted classes of finite structures. This is because restricting the theorem to a subclass $\mathcal{C}'$ of a class of structures $\mathcal{C}$ weakens both the hypothesis and the consequence of the theorem. This means that the homomorphism-preservation theorem may hold for the class of all finite structures, but it may fail for some restricted class $\mathcal{C}$ of finite structures. It also means that this result may hold for some restricted class $\mathcal{C}$ of finite structures, but it may fail for the class of all finite structures.

In this paper, we show that the homomorphism-preservation theorem holds for numerous large classes of finite structures of interest in graph theory and database theory. 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 cores 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, the classes of all structures whose cores are 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. To put these results in perspective, let us briefly comment on some of the key notions. 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 [23]), but has found applications in conjunctive query processing and optimization [5] and, more recently, in data exchange [13]. The *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 [10]). 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 [8, 10, 19, 20].

The proofs of our results make use of both earlier work about preservation properties in the finite and rather advanced combinatorial machinery. Ajtai and Gurevich [3] 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 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, 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.

Finally, we extend the Ajtai-Gurevich Theorem to a language much richer than Datalog. It is known that Datalog can be viewed as a (proper) fragment of the infinitary logic $\exists L^{\omega;+}_{\infty\omega}$, which is the existential-positive fragment of the finite-variable infinitary logic $L^{\omega}_{\infty\omega}$. Here, we generalize the Ajtai-Gurevich Theorem by establishing that if a query q is both $\exists L^{\omega;+}_{\infty\omega}$-definable and first-order definable on the class of all finite structures, then it is also definable by an existential positive formula. This result is established through a tight connection between the number of variables in a formula of $\exists L^{\omega;+}_{\infty\omega}$ and the treewidth of its 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. Finally, in Section 6 we examine the relationship of these results to definability in Datalog and the infinitary logic $\exists L^{\omega,+}_{\infty\omega}$.

2. Preliminaries

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

Relational structures 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 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.

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

- $N^{\mathbf{G}}_0(u) = \{u\}$;
- $N^{\mathbf{G}}_{d+1}(u) = \{v \in V : (v, w) \in E \text{ for some } w \in N^{\mathbf{G}}_d(u)\}$.

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 [9]) 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 non-empty *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, note that if we contract the edges of a perfect matching between two sets of size $k-2$ in each part, then we obtain a complete graph on $k-2$ nodes, which, together with the remaining two nodes and all remaining edges, gives rise to a $\mathbf{K}_k$.

First-order logic and Datalog 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 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 $\exists x_1 \ldots \exists x_n \theta$, 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 $\exists x_1 \ldots \exists x_n \theta$ 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 $\mathbf{A}$, 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 $\mathbf{A}$ (see [24]). Conversely, every conjunctive query $\exists x_1 \ldots \exists x_n \theta$ gives rise to a *canonical structure* $\mathbf{A}$ with n elements, where the elements of $\mathbf{A}$ are the variables $x_1, \ldots, x_n$ and the relations of $\mathbf{A}$ consist of the tuples of variables in the conjucts of θ. As shown by Chandra and Merlin [5], this basic relationship between conjunctive queries and finite structures plays a key role in database query processing and optimization.

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 [33]). For example, the following Datalog program defines the *transitive closure* of the edge relation E of a graph $\mathbf{G} = (V, E)$:

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

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.

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

Henceforth, we will restrict our attention to Boolean queries. All the results we establish apply equally well to non-Boolean queries. However, some of the definitions (such as that of minimal models) are less intuitive when we consider non-Boolean queries. Thus, in the interest of clarity we present the constructions for Boolean queries only.

Let $\mathcal{C}$ be a class of finite σ-structures and let q be a Boolean query on $\mathcal{C}$. 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. We say that 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 $q(\mathbf{B}) = 1$.

Let q be a query that is preserved under homomorphisms on all finite σ-sturctures. The first observation we make is that the minimal models of a query that is preserved under homomorphisms are *cores*. The concept of core was introduced in the context of graph theory (see [23]), 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.

The following characterization is part of the folklore, a proof for the class of all finite σ-structures can be found in [4]. 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 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 (*sketch*). 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 a structure $\mathbf{B} \models \varphi_{\mathbf{A}}$ if and only if there is a homomorphism from $\mathbf{A}$ to $\mathbf{B}$ (see [5]).

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 $\mathbf{A}_i$ may not be members 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 is $\mathcal{C}$ 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 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 [3], 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 [16]. Before we state the precise result, we need a definition. 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 $u, v \in A$. 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 [3].

Theorem 2 ([3]). *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, where $\mathcal{G}(\mathbf{A}) - B$ is the graph obtained from $\mathcal{G}(\mathbf{A})$ by removing all nodes in B and the edges to which they are incident. 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 1 and 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. Actually, it suffices to show that the collection of Gaifman graphs of cores of the structures in $\mathcal{C}$ has this property. We formulate this observation as the following corollary, which we will use repeatedly in what follows.

Corollary 1. *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 $\mathrm{core}(\mathbf{A})$ has more than N elements, then there is a set B of at most s elements such that $\mathcal{G}(\mathrm{core}(\mathbf{A})) - B$ has a d-scattered set of size m.*

On the class $\mathcal{C}$, every 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 1. *For every $k \geq 0$, $d \geq 0$, and $m \geq 0$, there exists an $N \geq 0$ such that for all graphs $\mathbf{G} = (V, E)$ 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 $\mathbf{G} = (V, E)$ 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 Boolean queries for classes of structures of bounded-degree and, actually, for classes of structures whose cores have bounded degree.

Theorem 3. *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 query that is first-order definable and is preserved under homomorphisms is also definable by an existential-positive first-order formula.*

4. Classes with Cores 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.

Unlike for Lemma 1, 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. 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 [12].

Theorem 4 (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 such that every pair of distinct sets in F' have a common intersection.*

Here is the promised combinatorial result:

Lemma 2. *For every $k \geq 1$, $d \geq 0$, and $m \geq 0$, there exists an $N \geq 0$ such that for all graphs $\mathbf{G} = (V, E)$ 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. Let $p = (m-1)(d+1)+1$, $M = k!(p-1)^k$, and let $N = k(m-1)^M$. Let $\mathbf{G} = (V, E)$ 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. Observe that the size of T is at least $N/k + 1$. 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 non-empty. We distinguish two cases:

Case 1: There is a node in 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 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 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 $p = (m-1)(d+1)+1$ 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 $T_{1+i(d+1)}$ for each $i \in \{0, \ldots, m-1\}$ produces a d-scattered subset in $\mathbf{G} - B$. To see this, it suffices to show that if $a \in T_i$ and $b \in T_{i+2}$, then the distance between a and b in $\mathbf{G} - B$ is at least two. Suppose on the contrary that a and b are adjacent in $\mathbf{G} - B$. Necessarily, a and b must appear together in some S_{u_j}. Then, $j < i+1$ because T_i and T_{i+1} are disjoint. Also, $j > i+1$ because T_{i+2} and T_{i+1} are also disjoint. This is a contradiction, which completes the proof. □

An immediate consequence of Lemma 2 and Corollary 1 is that the homomorphism-preservation theorem holds for classes of structures of bounded treewidth. Indeed, it holds for all classes whose cores have bounded treewidth. 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 in [7, 18]. 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 sequel, whenever 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 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 cores of 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. In particular, among others, they include all trees, all unicyclic graphs, and all outerplanar graphs. Classes of structures whose cores have bounded treewidth are even broader and 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 $\mathcal{H}(\mathcal{T}(4))$.

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 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 [27]. Before the main result, we establish a lemma on bipartite graphs. The proof relies on Ramsey's Theorem (see [17]).

Theorem 6 (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.*

We write $r(l, k, m)$ for the bound N obtained in Ramsey's Theorem. We can now state the lemma on bipartite graphs we require.

Lemma 3. *For every $k \geq 1$ and $m \geq 0$, there is an $N \geq 0$ such that if $\mathbf{G} = (V, U, E \subseteq V \times U)$ is a bipartite graph such that $\mathbf{K}_k$ is not a minor of $\mathbf{G}$ and $|V| > N$, then there are sets $S \subseteq V$ and $Z \subseteq U$ with $|S| > m$ and $|Z| < k - 1$ such that S is 1-scattered in $\mathbf{G} - Z$ and for each $s \in S$ and $z \in Z$ we have $\{s, z\} \in E$.*

Proof. The case $k \leq 2$ is trivial as, if $\mathbf{K}_2$ is not a minor of $\mathbf{G}$, then $\mathbf{G}$ 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. Let $N = b^{k-2}(m)$, that is, the function b iterated $k - 2$ times. We construct the sets S and Z in a series of stages, with at most $k - 1$ iterations. Begin with $S_0 = V$ and $Z_0 = \emptyset$.

Now, suppose at stage r we have sets $S_r \subseteq V$ and $Z_r \subseteq U$, with $|Z_r| \leq r$ and $|S_r| > b^{k-2-r}(m)$ Let $<$ be an arbitrary linear ordering of S_r and let $f : [S_r]^k \to \{0, \ldots, k\}$ be the function that assigns to each k-element subset s of S_r the maximal size of an $<$-initial segment of s such that all of its elements have a common neighbour in $U - Z_r$. By Ramsey's Theorem, there is a set $I \subseteq S_r$, 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 B denote the last $k - 2$ elements of I under the order $<$. Then, $I - B$ is a 1-scattered in $\mathbf{G} - Z_r$ as every pair of elements in $I - B$ forms the first two elements of some ordered k-element subset of I and therefore cannot have a common neighbour. Note also, that $|I - B| \geq (k-2)b^{k-2-(r+1)}(m)$. Since $r < k - 2$, this means $|I - B| \geq (k-2)m \geq m$ as $k \geq 3$. Thus, taking $S = I - B$ and $Z = Z_r$, we are done.

Case 2: $1 < f([I]^k) < k$. Let $f([I]^k) = t$. If B denotes the last $k - t$ elements of I under the order $<$, then every t-element subset of $I - B$ has a common neighbour in $U - Z_r$, as it is the initial segment of size t of some k-element subset of I. Furthermore, no $t + 1$ element subset of $I - B$ has a common neighbour in $U - Z_r$, from which we conclude that the maximal degree of any element in $U - Z_r$ (with respect to $I - B$) is t. Now, let $X_1, \ldots, X_k \subseteq I - B$ be a collection of k pairwise disjoint sets, each with exactly t elements. Such a collection exists, since $|I - B| > (k-2)b^{k-2-(r+1)}(m) \geq m \geq k^2$. Then, by the argument above, for each X_i, there is a $u_i \in U - Z_r$ which is a common neighbour of all elements in X_i, and u_i has no other neighbours. Thus, the set $X_i \cup \{u_i\}$ forms a connected patch in the graph $\mathbf{G} - Z_r$. Similarly, for each i and j with $1 \leq i < j \leq k$, we can find an element $u_{ij} \in U - Z_r$ such that, if $N(u_{ij})$ denotes the set of neighbours 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 neighbour of this subset would serve as u_{ij}. Again, u_{ij} cannot have any other neighbours in $I - B$, as no $t+1$-element subset of $I - B$ has a common neighbour. Thus, in particular, u_{ij} has no neighbours 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 $\mathbf{G}$, a contradiction. We conclude that this case cannot occur.

Case 3: $f([I]^k) = k$. This means that every k-element subset of I has a common neighbour in $U - Z_r$. 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 neighbour, there is a function $w : I - X \to U - Z_r$ such that $w(y)$ is a common neighbour of $X \cup \{y\}$. If the range of w contains $k - 1$ distinct elements, $\mathbf{G}$ 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 w has fewer than $k - 1$ elements. Thus, there is a $J \subseteq I - X$ with $|J| \geq |I - X|/(k-2)$ on which w is constant. Let $z \in U$ be the element to which w maps J. We let $S_{r+1} = J \cup X$ and $Z_{r+1} = Z_r \cup \{z\}$. Observe that z is a common neighbour of all elements in S_{r+1}, and that $|S_{r+1}| \geq |X| + |I - X|/(k-2) > (k-1) + b^{k-2-(r+1)}(m) - 1 > b^{k-2-(r+1)}(m)$ 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, an element is always added to the set Z. If the set were to contain $k-1$ elements, as all these elements are neighbours of all elements in S, which has at least $m \geq k$ elements, we would have that $\mathbf{G}$ contains $\mathbf{K}_{k-1,k-1}$, and therefore $\mathbf{K}_k$ as a minor. This establishes that $|Z| < k-1$. □

The main combinatorial result of this paper can now be proved by a construction that iterates Lemma 3.

Theorem 7. *For every $k \geq 1$, $d \geq 0$, and $m \geq 0$, there is an $N \geq 0$ such that if $\mathbf{G} = (V, E)$ 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 3 and r is the Ramsey function. Let $N = c^d(m)$. We construct Z and S in d stages. 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 $\mathbf{G}_i$, where $\mathbf{G}_i = \mathbf{G} - Z_i$. 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 $\mathbf{G}_i$. 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 $\mathbf{G}_i$. By Ramsey's theorem, either this graph contains 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 $\mathbf{G}_i$. 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 $\mathbf{H} = (A, B, E \subseteq A \times B)$ on which to apply Lemma 3. Let $A = I$, and let B be the set of vertices of $\mathbf{G}_i$ 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 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 $\mathbf{G}_i$. By Lemma 3, there exist $S' \subseteq A$ and $Z' \subseteq B$ such that $|S'| > c^{d-i-1}(m)$, S' is 1-scattered in $H - Z'$ and $\{a, b\} \in E$ for each $a \in S'$ and $b \in Z'$. Let $Z_{i+1} = Z_i \cup Z'$ and $S_{i+1} = S'$, which is $(i+1)$-scattered in $\mathbf{G}_{i+1} = \mathbf{G} - Z_{i+1}$. 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, $\{a, b\} \in E$ for each $a \in S'$ and $b \in Z'$, which means that, in $\mathbf{G}$, each $b \in Z'$ is adjacent to some vertex in $N_i(a)$ for every $a \in S'$. 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 $a \in S'$. Consider each $N_i(u)$, with $u \in S'$, as a connected patch in the subgraph of $\mathbf{G}$ induced by $\bigcup_{u \in S'} N_i(u)$ and Z_{i+1}. Note that these patches are disjoint. The $\mathbf{K}_{k-1,k-1}$ minor is now clear since $|S'| \geq k-1$ and $|Z_{i+1}| \geq k-1$. □

Combining this with Corollary 1 we get the following result.

Theorem 8. *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 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 8 and the earlier Theorems 5 and 3.

As noted earlier, the class $\mathcal{T}(k)$ of graphs of treewidth less than k excludes $\mathbf{K}_{k+1}$ as a minor. By the same token, the Gaifman graphs of cores of structures in $\mathcal{H}(\mathcal{T}(k))$ exclude $\mathbf{K}_{k+1}$ as a minor. Thus, the homomorphism-preservation theorem for these classes (Theorem 5) is a special case of Theorem 8. 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 5 and 8 can be obtained using certain deep results by Robertson and Seymour [28] about classes of graphs excluding a minor. Specifically, Robertson and Seymour [28] 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 non-planar (this result is a consequence of the the Excluded Grid Theorem of Robertson and Seymour [28] - see also [9, pages 264–274]). 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 8, but not from Theorem 5, precisely when $\mathbf{H}$ is a non-planar 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 can not be derived as a consequence of Theorem 8, unless one could show that the class $\mathcal{C}$ under consideration has the property that the graphs of bounded degree in $\mathcal{C}$ that are also cores exclude some minor.

6. Ajtai-Gurevich Theorem Revisited

The Ajtai-Gurevich Theorem [3] asserts that every Datalog program that is first-order definable is *bounded*, that is, the associated monotone operator reaches its least fixed-point after a uniformly bounded number of iterations on every structure. The aim of this section is to show that the results on treewidth in Section 4 yield an alternative and perhaps more transparent proof of this result. Actually, we obtain a stronger result for a logic strictly more expressive than Datalog. We proceed in a sequence of lemmas. The first lemma asserts that the minimal models of existential-positive first-order formulas have treewidth bounded by the number of variables. This is a consequence of results in [7], following [26], where it is shown that the core of a structure has treewidth less than k if and only if the canonical conjunctive query of that structure is logically equivalent to an existential-positive first-order formula with k variables. For completeness, we outline a self-contained proof.

Lemma 4. *If φ is an existential-positive first-order formula with k variables, then every minimal model of φ has treewidth less than k.*

Proof. We show that every model of φ contains a submodel with treewidth less than k. Since existential quantifiers commute with disjunctions and since conjunctions distribute over disjunctions, every existential-positive first-order sentence with k variables can be written as a finite disjunction of existential-positive disjunction-free first-order sentences with k variables. Hence, it suffices to show that every model of an existential-positive disjunction-free first-order sentence with k variables has a submodel of treewidth less than k. Let ψ be the result of renaming all occurrences of variables in φ so that each existential quantifier bounds 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 sentence of the form $(\exists x_1)\cdots(\exists x_n)\theta$ that is equivalent to ψ, where θ is a conjunction of atomic facts. Now, suppose $\mathbf{A}$ is a model of ψ. Consider the substructure $\mathbf{B}$ of $\mathbf{A}$ whose universe consists of the (not necessarily distinct) witnesses $a_1, \ldots, a_n$ of the n existential quantifiers in which $R(a_{i_1}, \ldots, a_{i_r})$ holds if, and only if, the atomic formula $R(x_{i_1}, \ldots, x_{i_r})$ appears in θ. Let us now see that this model 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 consisting of the witnesses a_j that interpret the 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 in the diagram of $\mathbf{B}$ is a subformula of ψ, it is not hard to see that the tree and its labeling is a tree-decomposition of $\mathbf{B}$ of width at most $k-1$. Hence, the treewidth of $\mathbf{B}$ is less than k. □

We now turn to infinitary logic as an intermediate step towards the Ajtai-Gurevich Theorem. The collection of *infinitary formulas* $L_{\infty\omega}$ is obtained by closing the atomic formulas under negation, infinitary conjunctions, infinitary disjunctions, universal and existential quantification. The k-variable fragment of $L_{\infty\omega}$ is denoted by $L^k_{\infty\omega}$. The collection of *existential-positive infinitary formulas* $\exists L^+_{\infty\omega}$ is obtained by closing the atomic formulas under infinitary conjunctions, infinitary disjunctions, and existential quantification. The k-variable fragment of $\exists L^+_{\infty\omega}$ is denoted by $\exists L^{k,+}_{\infty\omega}$. It was shown in [26, Theorem 4.1] that for every positive integer k, every k-Datalog program is expressible in $\exists L^{k,+}_{\infty\omega}$. This provides the link between Datalog and infinitary logic.

The expressive power of $\exists L^{k,+}_{\infty\omega}$ is captured by the *existential k-pebble games*, first defined in [25, 26]. These games are 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 numbered $\{1, \ldots, k\}$. In each round of the game, the Spoiler can make one of two different moves: either he places a free pebble on an element of the domain of $\mathbf{A}$, or he removes a pebble from a pebbled element of $\mathbf{A}$. To each move of the Spoiler, the Duplicator must respond by placing her corresponding pebble over an element of $\mathbf{B}$, or removing her corresponding pebble from $\mathbf{B}$, respectively. If the Spoiler reaches 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 $\exists L^{k,+}_{\infty\omega}$ was proved in [25, 26]: let φ be an $\exists L^{k,+}_{\infty\omega}$ sentence; if the Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{A}$ is a model of φ, then $\mathbf{B}$ is also a model φ. We use existential k-pebble games to prove a normal form for $\exists L^{k,+}_{\infty\omega}$

Lemma 5. *On the class of all finite σ-structures, every $\exists L^{k,+}_{\infty\omega}$ sentence is equivalent to an infinitary disjunction of existential-positive first-order sentences with k variables.*

Proof (*sketch*). Let φ be an $\exists L^{k,+}_{\infty\omega}$ sentence. For every model $\mathbf{A}$ of φ, let $\psi_{\mathbf{A}}$ be the statement "the Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$". This is viewed as a query on finite structures $\mathbf{B}$. We first observe that φ is equivalent to the infinitary disjunction $\bigvee \psi_{\mathbf{A}}$, where $\mathbf{A}$ ranges over all finite models of φ. Indeed, if $\mathbf{B}$ is a model of φ, then $\mathbf{B}$ satisfies $\psi_{\mathbf{B}}$, which is one of the disjuncts of the infinitary disjunction. Conversely, if $\mathbf{B}$ satisfies the infinitary disjunction, then it satisfies $\psi_{\mathbf{A}}$ for some model $\mathbf{A}$ of φ. This means that the Duplicator wins the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$; since $\mathbf{A}$ satisfies φ, it follows that $\mathbf{B}$ satisfies φ.

So, it suffices to show that each statement $\psi_{\mathbf{A}}$ is definable by an existential-positive first-order sentence with k variables. Consider the following query: "Given $\mathbf{A}$ and $\mathbf{B}$, does the Spoiler win the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$?" It was shown in [26, Theorem 4.7] that this query is definable in least fixed-point logic. Moreover, for every fixed $\mathbf{B}$, this query is expressible in k-Datalog; this is shown by instantiating the inductive definition to all k-tuples of $\mathbf{B}$. By instantiating it to all tuples from $\mathbf{A}$, instead of $\mathbf{B}$, we can extract a system of universal first-order formulas that are positive in the recursive predicates, but each atomic formula from σ occurs negatively only. More precisely, for every k-

tuple $\mathbf{a} = (a_1, \dots, a_k)$ from $\mathbf{A}$, we get a k-ary relation symbol $T_{\mathbf{a}}(y_1, \dots, y_k)$ that will be used to express the property "the Spoiler wins from the position $(a_1, \dots, a_k, y_1, \dots, y_k)$". The resulting system of universal first-order formulas has the property that all its finite stages are definable by formulas of first-order logic built using negated atomic formulas, universal quantification, and disjunction. Using the techniques in [26, Theorem 4.3], one can show that these formulas can be rewritten so that only k variables are used. Finally, it remains to show that it is enough to iterate the system finitely many times. This follows from the fact that there are no more than $|A|^k$ configurations for the Spoiler in the existential k-pebble game on $\mathbf{A}$ and $\mathbf{B}$. By negating the disjunction of these finitely many stages, we obtain an existential-positive first-order formula with k variables expressing $\psi_{\mathbf{A}}$. □

An immediate corollary of the normal form and Lemma 4 is that the collection of minimal finite models of an $\exists L^{k,+}_{\infty\omega}$ sentence has treewidth less than k. Actually, there is a converse in that every query q that is preserved under homomorphisms and whose minimal models have treewidth less than k can be expressed by a formula of $\exists L^{k,+}_{\infty\omega}$. This shows the tight connection between the results we established in Section 4 and definability in $\exists L^{k,+}_{\infty\omega}$.

Lemmas 4 and 5 together with Theorem 5 gives us the following:

THEOREM 9. *On the class of all finite σ-structures, every query that is definable by a first-order formula and an $\exists L^{\omega,+}_{\infty\omega}$ formula is also definable by an existential-positive first-order formula.*

We now have all the necessary tools to give an alternative proof of the Ajtai-Gurevich Theorem. In fact, at this point we can give two different proofs of the Ajtai-Gurevich Theorem. One of these proofs uses the compactness theorem of first-order logic. Here, we sketch the second proof, which is more in the spirit of database theory as it uses a well known result by Sagiv and Yannakakis [30] about unions of conjunctive queries.

THEOREM 10 ([3]). *If a Datalog program is first-order definable, then it is bounded.*

PROOF. Let P be a Datalog program that is first-order definable. For every $n \geq 1$, let φ_n be the first-order formula defining the n-th stage of the Datalog program. As is well known, each φ_n is a union of conjunctive queries, so P is definable by an infinite disjunction $\bigvee_{m\geq 1} q_m$, where q_m is a conjunctive query. Since P is first-order definable, Theorem 9 implies that P is also definable by a finite union $\bigvee_{i=1}^{s} q'_i$, where each q'_i is a conjunctive query. Sagiv and Yannakakis [30] 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. Consequently, there is a positive integer t such that $\bigvee_{m\geq 1} q_m$ is logically equivalent to $\bigvee_{m=1}^{t} q_m$. Thus, the Datalog program P is bounded. □

7. Concluding Remarks

We have investigated the homomorphism-preservation theorem in the finite and have shown that it holds for numerous classes of finite structures of interest in graph theory and database theory. As noted earlier, preservation theorems do not relativize to restricted classes of structures, so our results stand by themselves independently of whether the homomorphism-preservation theorem holds or fails on the class of all finite structures. Indeed, one can ask the same question for other classes of finite structures. For instance, we could consider classes of bounded local treewidth [11, 15] or of bounded cliquewidth [6]. 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 generalise 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. Moreover, it is perhaps worth mentioning that another consequence of our results is that in order to prove the homomorphism-preservation theorem for the class of all finite structures, it now suffices to show that the collection of minimal models of any first-order query preserved under homomorphisms excludes some minor.

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 compared with the undecidability of the same problem on the class of all finite structures [4]. The exact complexity of these problems on the class $\mathcal{T}(k)$ could be prohibitive, but this remains to be determined.

8. REFERENCES

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

[2] M. Ajtai and Y. Gurevich. Monotone versus positive. *Journal of the ACM*, 34:1004–1015, 1987.

[3] M. Ajtai and Y. Gurevich. Datalog vs first-order logic. *J. of Computer and System Sciences*, 49:562–588, 1994.

[4] N Alechina and Y. Gurevich. Syntax vs semantics on finite structures. In J. Mycielski, G. Rozenberg, and A Salomaa, editors, *Structures in Logic and Computer*

Science, volume 1261 of *LNCS*, pages 14–33. Springer, 1997.

[5] A.K. Chandra and P.M. Merlin. Optimal implementation of conjunctive queries in relational databases. In *Proc. 9th ACM Symp. on Theory of Computing*, pages 77–90, 1977.

[6] B. Courcelle, J. Engelfriet, and G. Rozenberg. Handle rewriting hypergraph grammars. *Journal of Computer and System Sciences*, 46:218–270, 1993.

[7] V. Dalmau, Ph. G. Kolaitis, , and M. Y. Vardi. Constraint satisfaction, bounded treewidth, and finite variable logics. In *8th International Conference on Principles and Practice of Constraint Programming - CP 2002*, volume 2470 of *Lecture Notes in Computer Science*, pages 310–326. Springer, 2002.

[8] R. Dechter and J. Pearl. Tree clustering for constraint networks. *Artificial Intelligence*, pages 353–366, 1989.

[9] R. Diestel. *Graph Theory*. Springer, 1997.

[10] R.G. Downey and M.R. Fellows. *Parametrized Complexity*. Springer-Verlag, 1999.

[11] D. Epstein. Diameter and treewidth in minor-closed graph families. *Algorithmica*, 27:275–291, 2000.

[12] P. Erdos and R. Rado. Intersection theorems for systems of sets. *J. of London Mathematical Society*, 35:85–90, 1960.

[13] R. Fagin, P. G. Kolaitis, and L. Popa. Data Exchange: Getting to the Core. In *Proc. 22nd ACM Symposium on Principles of Database Systems*, pages 90–101, 2003.

[14] T. Feder and M.Y. Vardi. Homomorphism closed vs existential positive. In *Proc. of the 18th IEEE Symp. on Logic in Computer Science*, pages 311–320, 2003.

[15] M. Frick and M. Grohe. Deciding first-order properties of locally tree-decomposable functions. *Journal of the Association for Computing Machinery*, 48:1184–2006, 2000.

[16] H. Gaifman. On local and nonlocal properties. In J. Stern, editor, *Logic Colloquium '81*, pages 105–135. North Holland, 1982.

[17] R. L. Graham, B. L. Rothschild, and J. H. Spencer. *Ramsey Theory*. Wiley, 1980.

[18] M. Grohe. 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*, 2003.

[19] M. Grohe, J. Flum, and M Frick. Query evaluation via tree-decompositions. *Journal of the ACM*, 49:716–752, 2002.

[20] M. Grohe, T. Schwentick, and Segoufin. When is the evaluation of conjunctive queries tractable? In *Proc. 32nd ACM Symp. on Theory of Computing*, pages 657–666, 2001.

[21] Y. Gurevich. Toward logic tailored for computational complexity. In M. Richter et al., editors, *Computation and Proof Theory*, pages 175–216. Springer Lecture Notes in Mathematics, 1984.

[22] Y. Gurevich. On finite model theory. In S.R. Buss and P.J. Scott, editors, *Feasible Mathematics*, pages 211–219. Birkhäuser, 1990.

[23] P. Hell and J. Nesetril. The core of a graph. *Discrete Math.*, 109:117–126, 1992.

[24] W. Hodges. *Model Theory*. Cambridge University Press, 1993.

[25] Ph. G. Kolaitis and M. Y Vardi. On the expressive power of Datalog: Tools and a case study. *Journal of Computer and System Sciences*, 51:110–134, 1995.

[26] Ph. G. Kolaitis and M. Y Vardi. Conjunctive query containment and constraint satisfaction. *Journal of Computer and System Sciences*, 61:302–332, 2000.

[27] M. Kreidler and D. Seese. Monadic NP and graph minors. In *CSL'98: Proc. of the Annual Conference of the European Association for Computer Science Logic*, volume 1584 of *LNCS*, pages 126–141. Springer, 1999.

[28] N. Robertson and P.D. Seymour. Graph minors V. Excluding a planar graph. *Journal of Combinatorial Theory, Series B*, 41:92–11, 1986.

[29] E. Rosen. Some aspects of model theory and finite strucrtures. *Bulletin of Symbolic Logic*, 8:380–403, 2002.

[30] Y. Sagiv and M. Yannakakis. Equivalence between relational expressions with the union and difference operators. *Journal of the Association for Computing Machinery*, 27(4):633–655, 1981.

[31] A. Stolboushkin. Finite monotone properties. In *Proc. 10th IEEE Symp. on Logic in Computer Science*, pages 324–330, 1995.

[32] W. W. Tait. A counterexample to a conjecture of Scott and Suppes. *Journal of Symbolic Logic*, 24:15–16, 1959.

[33] J.D. Ullman. Bottom-up beats top-down for Datalog. In *Proc. 8th ACM Symp. on Principles of Database Systems*, pages 140–149, 1989.