Pattern logics and auxiliary relations

Citation for published version:
Figueira, D & Libkin, L 2014, Pattern logics and auxiliary relations. in Logic in Computer Science: Joint Meeting of the Twenty-Third EACSL Annual Conference on Computer Science Logic (CSL) and the Twenty-Ninth Annual ACM/IEEE Symposium on Logic in Computer Science (LICS), CSL-LICS '14, Vienna, Austria, July 14 - 18, 2014., 40, ACM, pp. 40. DOI: 10.1145/2603088.2603136

Digital Object Identifier (DOI):
10.1145/2603088.2603136

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Logic in Computer Science

Pattern Logics and Auxiliary Relations

Diego Figueira Leonid Libkin
University of Edinburgh

Abstract

A common theme in the study of logics over finite structures is adding auxiliary predicates to enhance expressiveness and convey additional information. Examples include adding an order or arithmetic for capturing complexity classes, or the power of real-life declarative languages. A recent trend is to add a data-value comparison relation to words, trees, and graphs, for capturing modern data models such as XML and graph databases.

Such additions often result in the loss of good properties of the underlying logic. Our goal is to show that such a loss can be avoided if we use pattern-based logics, standard in XML and graph data querying. The essence of such logics is that auxiliary relations are tested locally with respect to other relations in the structure. These logics are shown to admit strong versions of Hanf and Gaifman locality theorems, which are used to prove a homomorphism preservation theorem, and a decidability result for the satisfiability problem. We discuss applications of these results to pattern logics over data forests, and consequently to querying XML data.

Categories and Subject Descriptors F.4.1 [*Mathematical Logic and Formal Languages*]: Mathematical Logic—Computational logic

Keywords First-order Logic, Pattern, Finite Model Theory, Gaifman-locality, Hanf-locality, Homomorphism Preservation Theorem, data-value, Graph Database, XML, Data-Tree

1. Introduction

A common theme in the study of logics over finite structures is adding auxiliary relations and investigating the extra expressiveness they provide. Such auxiliary relations may serve several purposes. They may provide information about presentations of structures that is otherwise not available (for instance, an ordering of graph vertices). They may provide operations on the underlying domain to make logics closer to practical features that they are modeling (for instance, arithmetic operations are a common feature of real-life database query languages). And sometimes they provide extra features as models evolve (an example of this is the study of data trees as a better abstraction of XML data than trees themselves). We now illustrate these in more detail.

- In the study of logics capturing complexity classes (cf. [26, 30, 32]), one usually adds an auxiliary relation that specifies an ordering of the domain. The goal is to bridge the gap between machine models, that can talk about a particular presentation, and logics, that cannot distinguish isomorphic structures.
- In the study of embedded finite models (cf. [26]), one adds arbitrary auxiliary relations, for instance arithmetic operations. The goal is usually to simulate the expressiveness of real-life query languages over finite databases, as those use extra operations on the domain.
- In the study of data models underlying modern database applications – such as XML and graph data – one distinguishes the basic structure (e.g., a tree for XML) and data that such basic structures carry. The latter is often modeled by a predicate expressing the property that two vertices carry the same data items, i.e., a data equality predicate, resulting in data words, data trees, and data graphs [7, 14–16, 34, 41].

With such auxiliary relations, one usually tries to control the power of the resulting logic. In the case of adding orderings, one does this by insisting on *invariance* of definable properties: they need not depend on a particular ordering that is used. Still, the mere presence of an order or a successor relation enhances the power of logic [32, 38]. At the same time, one can establish useful properties of definable queries, for instance, locality [4, 27], or bounds on their expressiveness [10, 22, 37, 40]. For embedded finite models, the spirit of the restrictions is similar: invariance under one-to-one mappings of the domain of structure. A typical result would state that adding arbitrary auxiliary relations yields no more power than adding an order under such invariance, cf. [26].

In the case of adding the data equality predicate, restrictions guaranteeing good behavior of logics are of a different nature. The study of such predicates started with words and trees, on which first-order logic (FO) and monadic second-order logic (MSO) are undecidable, which explains the interest in preserving decidability (and connection with automata) in the presence of data. The standard restriction guaranteeing decidability is to the *two-variable* fragment, FO^2, see [15, 16], or extensions thereof [14, 42].

We are interested in different types of restrictions guaranteeing good properties of logics with auxiliary relations. Our primary motivation comes from studying logics over advanced data models such as XML and graph data, where auxiliary relations capture the notion of adding data values. A concrete example is shown in Fig. 1. In this tree, nodes have labels from a finite alphabet $\{a, b, c\}$, and they also come equipped with data values from an infinite set, in this case, $\mathbb{N}$. One represents such a tree as a structure that has the usual child relation, unary labeling predicates for alphabet letters,

CSL-LICS 2014, July 14–18, 2014, Vienna, Austria.
Copyright © 2014 ACM 978-1-4503-2886-9…$15.00.
http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

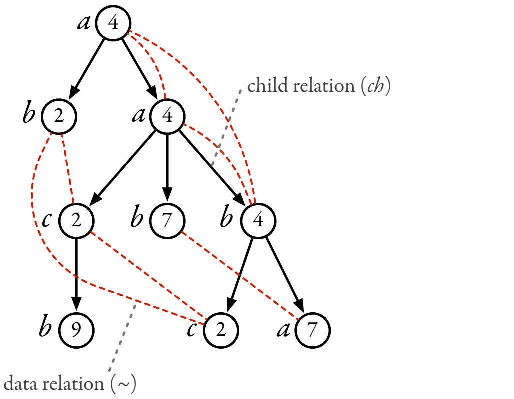

Figure 1. Example of a data tree.

and the *data equality* predicate for data values (instead of infinitely many unary predicates); we write $x \sim y$ if nodes x and y have the same data value (as shown by the dashed lines in the figure).

Restrictions such as invariance are simply not relevant in this setting (and invariant properties cannot say anything interesting about data values). Using FO^2 makes the logic too weak: many languages studied in the context of XML with data [3, 13, 35, 36, 44] or graphs with data [7, 34] need to go beyond the two-variable fragment.

Thus, we would like to find restrictions on logics with auxiliary relations that, while sufficiently expressive, do not impose too severe restrictions. The idea of such restrictions is motivated by practical languages for XML and graph data. A key notion in many of them is that of a *pattern*: essentially a small subgraph, or a subtree, that needs to be matched in a large data set. These often provide the basis for language constructors, for instance, for the navigational language XPath [35, 36], more general XML query languages [3, 13, 17], or path queries over graphs [7, 23]. Patterns also provide a standard abstraction of *incomplete data* in relational, XML, and graph models [1, 2, 8, 9, 29].

An example of a pattern in a tree is a node labeled a that has an a-child with the same data value and a b-child with a different data value; this is clearly matched in the data tree in Fig. 1 by the root and its two children.

Patterns are highly useful as query language constructors for at least two different data models, and at the same time, pattern-based logical languages provide a restriction on the use of auxiliary relations that makes it possible to recover nice properties of the underlying logic. The restriction imposed by patterns is that the use of auxiliary relations becomes *local*: we check those predicates for elements of the domain that are close to each other in the structure. In the above example, we check data equality of a node and its child, which are at distance one in the tree.

Our goal is to show that pattern-based logics with auxiliary relations behave well without imposing very restrictive conditions such as invariance or two-variable limitation. In fact, this good behavior is due to the local nature of auxiliary predicates.

The particular questions we investigate are inspired by applications of logical formalisms with data equality predicate in querying XML and graph data. These questions are split into three groups. The first one concerns the expressiveness of logics, which is a standard subject for investigation. The second one has to do with the homomorphism preservation theorem. This choice is motivated by both the importance of the result in finite model theory ([39] proved that it holds over the class of all finite structures, solving a long-standing open problem; it also holds for several well-behaved classes of finite structures [6, 18]) and its applications in the data management field (*e.g.*, in handling incomplete information in databases [25], in characterizing classes of schema mappings in data exchange [43], and in rewritability of ontology-mediated queries [11]). The third group deals with the basic decidability question, which in fact motivated two-variable restrictions for logics over data words and data trees.

Our main contributions are as follows.

1) Motivated by logics used for querying XML and graph data, we define a pattern logic with auxiliary relations, and an equally expressive logic: FO with local auxiliary relations.

2) When it comes to the expressiveness of the logic, we prove analogs of Hanf and Gaifman locality properties [24, 28] (see also [19, 32]), with neighborhoods defined *without* using the auxiliary relations. These stronger versions of well known locality theorems give us easy tools to prove expressivity bounds for pattern logics.

3) We prove a homomorphism preservation theorem for pattern logics. It works over classes of structures that include, among others, ranked data trees. Crucially, such classes are defined without mentioning auxiliary relations: for instance, we can specify classes of trees so that the homomorphism preservation theorem would hold on them when the data equality predicate is added. Ranked data trees provide just one example of a class of structures for which the result holds; we state the condition in terms of closure and structural properties of the class. We also discuss applications of this result in querying XML data.

Note that such a result does not follow from Rossman's theorem [39], since trees are not FO-definable, nor from other results on homomorphism preservation [6, 18] which impose restrictions that are violated by adding the data equality relation. Homomorphism preservation results in the finite [6, 18, 39] are rather hard to obtain, and our result requires a considerable effort too, crucially relying on the revised form of Gaifman locality theorem.

4) We show a decidability result for pattern logics. It is given for classes of structures which include data words and ranked data trees; we further show that the logic continues to be decidable when one adds the sibling ordering and other relations (for instance, other equivalence relations).

Thus, we retain decidability without such strong restrictions on expressiveness that the two-variable fragment imposes. In the case of unranked data trees, however, we show undecidability of the pattern logic.

Organization. Basic definitions are given in Section 2. In Section 3 we define pattern logics and logics with local auxiliary relations. Section 4 proves locality theorems, and Section 5 shows the homomorphism preservation theorem and explains some of its application. In Section 6 we discuss decidability of pattern logics. Section 7 offers concluding remarks.

2. Preliminaries

Relational Structures A *relational vocabulary* σ is a finite set of relation symbols, each with a specified *arity*. A σ-structure $\mathbb{A}$ consists of a universe A (also called a domain) and an interpretation $R^{\mathbb{A}} \subseteq A^m$ of each $R \in \sigma$ of arity m. We normally use $A, B, \ldots$ to denote the universe of $\mathbb{A}, \mathbb{B}, \ldots$.

Let $\sqcup$ denote the disjoint union of sets (*i.e.*, $A \sqcup B$ is $A \cup B$ if $A \cap B = \emptyset$). Let $\oplus$ denote the disjoint union of σ-structures, that is, $\mathbb{A} \oplus \mathbb{B}$ has universe $A \sqcup B$ and $R^{\mathbb{A}\oplus\mathbb{B}} = R^{\mathbb{A}} \sqcup R^{\mathbb{B}}$ for every $R \in \sigma$. 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 $\mathbb{B}$ is called a *substructure* of another σ-structure $\mathbb{A}$ if $B \subseteq A$ and $R^{\mathbb{B}} \subseteq R^{\mathbb{A}}$ for every $R \in \sigma$. In that case we write $\mathbb{B} \subseteq \mathbb{A}$. A substructure $\mathbb{B}$ is an *induced substructure* if $R^{\mathbb{B}} = R^{\mathbb{A}} \cap B^m$ for every $R \in \sigma$ of arity m. For $A' \subseteq A$ we write $\mathbb{A}|_{A'}$ for the induced substructure of $\mathbb{A}$ with domain A'. For $\sigma' \subseteq \sigma$, the *σ'-induced substructure of* $\mathbb{A}$, denoted by $\mathbb{A}|_{\sigma'}$, is the restriction of $\mathbb{A}$ to the signature σ'.

The *Gaifman graph* of a σ-structure $\mathbb{A}$, denoted by $\mathcal{G}(\mathbb{A})$, is the undirected graph whose set of nodes is the universe of $\mathbb{A}$, and whose edges are pairs (a, a') of distinct elements of A that appear together in some tuple of a relation in σ.

Let $\mathbf{G} = (V, E)$ be a graph. Recall that the *distance* between two vertices u and v is the length of the shortest path from u to v. For a vertex u and an integer $r \geq 0$, the *r-neighborhood* of u in $\mathbf{G}$, denoted by $\mathrm{N}_r^{\mathbf{G}}(u)$, is the set of vertices at distance at most r from u. In particular, $\mathrm{N}_0^{\mathbf{G}}(u) = \{u\}$. For a tuple $\bar{u} = (u_1, \ldots, u_m)$, we let $\mathrm{N}_r^{\mathbf{G}}(\bar{u})$ be $\bigcup\{\mathrm{N}_r^{\mathbf{G}}(u_i) \mid 1 \leq i \leq m\}$. Where this causes no confusion, we also write $\mathrm{N}_r^{\mathbf{G}}(\bar{u})$ for the subgraph of $\mathbf{G}$ induced by this set of vertices.

For a structure $\mathbb{A}$ and a vector $\bar{a}$ of elements of A, we define the r-neighborhood $\mathrm{N}_r^{\mathbb{A}}(\bar{a})$ as the induced substructure with the universe $\mathrm{N}_r^{\mathcal{G}(\mathbb{A})}(\bar{a})$ expanded with $|\bar{a}|$ constant symbols interpreted as the elements of $\bar{a}$. In particular, if two neighborhoods $\mathrm{N}_r^{\mathbb{A}}(\bar{a})$ and $\mathrm{N}_r^{\mathbb{B}}(\bar{b})$ are isomorphic, which is denoted by $\mathrm{N}_r^{\mathbb{A}}(\bar{a}) \cong \mathrm{N}_r^{\mathbb{B}}(\bar{b})$, it means that there is an isomorphism that sends $\bar{a}$ to $\bar{b}$.

Data trees and forests. These structures received much attention as of late, being an abstraction of XML documents with data. The structure of XML documents is normally modeled as a tree; in data trees, we add a data value from an infinite domain to each node. The usual way to represent this as a first-order structure is by means of an equivalence relation $\sim$ on nodes whose meaning is that two nodes carry the same data value, see Fig. 1.

We now give precise definition, generalizing slightly the standard ones from [14–16, 41]. For a vocabulary $\sigma = \tau \sqcup \{ch\}$, we say that a σ-structure $\mathbb{A}$ is a *τ-tree* if $ch^{\mathbb{A}}$ is the child relation of a finite tree. Over τ-trees, we use "root", "parent", "child", "sibling", etc., to make reference to elements of the *ch*-induced tree. For any τ-tree $\mathbb{A}$, we say that *sib* is a *sibling order* for $\mathbb{A}$ if it is the union, over all elements a in $\mathbb{A}$, of total linear orders on sets of a's siblings $\{b \in \mathbb{A} \mid (a, b) \in ch^{\mathbb{A}}\}$. We say that a τ-tree $\mathbb{A}$ is *k-ranked* if every element of $\mathbb{A}$ has at most k children. We write *root*$(\mathbb{A})$ to denote the root element of $\mathbb{A}$. Given a tree $\mathbb{A}$ and an element $a \in A$, by $\mathbb{A}{\upharpoonright}a$ we denote the induced substructure given by the subtree rooted at a.

The class of *data trees* $\mathcal{DT}$ is the class of all τ-trees where τ has monadic relations and one binary equivalence relation $\sim$. The class of *data forests* $\mathcal{DF}$ is the class of disjoint unions of data trees, *i.e.*, $\mathbb{A}_1 \oplus \cdots \oplus \mathbb{A}_m$ for $m \in \mathbb{N}$ such that $\mathbb{A}_1, \ldots, \mathbb{A}_m \in \mathcal{DT}$. By $\mathcal{DF}_k$ we denote the class of data forests in which every data tree is k-ranked.

First-Order Logic (FO) Let σ be a relational vocabulary. Formulae of FO over σ are obtained by closing relational atomic formulae $R(x_1, \ldots, x_m)$, where $R \in \sigma$ is of arity m, and atomic equality formulae $x = y$ under conjunction $\wedge$, disjunction $\vee$, negation $\neg$, existential $\exists$ and universal $\forall$ quantification. The semantics is standard. We write $\mathbb{A} \models \varphi(\bar{a})$ if φ is true in $\mathbb{A}$ when its free variables are interpreted by the tuple of elements $\bar{a}$. The *quantifier rank* of a first-order formula is the maximum depth of quantifier nesting in it.

The class $\exists$Pos of *existential-positive* formulae is the closure of atomic formulae under conjunction, disjunction, and existential quantification. By substituting variables, one can eliminate equalities in $\exists$Pos formulae.

Locality. FO formulae are known to exhibit locality properties, most commonly formulated in terms of theorems by Hanf and Gaifman. Define $(\mathbb{A}, \bar{a}) \leftrightarrows_r (\mathbb{B}, \bar{b})$ for $\bar{a} \in A^n$ and $\bar{b} \in B^n$ if there exists a bijection $f : A \to B$ such that $\mathrm{N}_r^{\mathbb{A}}(\bar{a}c) \cong \mathrm{N}_r^{\mathbb{B}}(\bar{b}f(c))$ for every $c \in A$. Hanf's locality theorem states that for every FO formula $\varphi(\bar{x})$, there exists a number r so that $(\mathbb{A}, \bar{a}) \leftrightarrows_r (\mathbb{B}, \bar{b})$ implies that $\mathbb{A} \models \varphi(\bar{a})$ if and only if $\mathbb{B} \models \varphi(\bar{b})$. In fact r can be taken to be $(3^k - 1)/2$, where k is the quantifier rank of φ. It was originally stated by Hanf in [28] for infinite models, and then restated for sentences over finite models in [22]; the present formulation is from [32].

To state Gaifman's theorem, note that for every integer $r \geq 0$, there is an FO formula $\delta(x, y) \leq r$ stating that the distance between x and y in the Gaifman graph is at most r. Let $\delta(x, y) > r$ be the negation of this formula. Note that $\delta(x, y) \leq r$ could be expressed in $\exists$Pos, by a formula whose quantifier rank is bounded by r plus the maximum arity of a relation of σ (in fact it can even be logarithmic in r).

Define r-local formulae $\psi^r(x)$ as formulae in which all quantifiers are restricted to the r-neighborhood of x, *i.e.*, formulae that are either of the form $\exists y\ (\delta(x, y) \leq r \wedge \ldots)$ or of the form $\forall y\ (\delta(x, y) \leq r \to \ldots)$. A *basic* FO *local sentence* is

$$\exists x_1 \ldots \exists x_n \Big(\bigwedge_{1 \leq i < j \leq n} \delta(x_i, x_j) > 2r \ \wedge \ \bigwedge_{i \leq n} \psi^r(x_i) \Big),$$

where $\psi^r(x)$ is an r-local formula. Gaifman's theorem [24] says that every FO sentence φ is equivalent to a Boolean combination of basic FO local sentences. Here r is exponential in the quantifier rank of φ.

Auxiliary relations. We often deal with a situation where the vocabulary is split into two kinds of relations: main and auxiliary ones. Throughout the paper we shall reserve τ for the vocabulary of auxiliary relations. If σ and τ are two disjoint vocabularies, we write $\sigma\tau$ as an abbreviation of $\sigma \sqcup \tau$. For instance, in data trees, we can treat the equal-value relation τ as the auxiliary relation. In the study of invariance, τ typically contains a single binary relation interpreted as a linear order or a successor.

3. Patterns and local auxiliary relations

As discussed in the introduction, logical languages for XML and graph data are largely based on *patterns* (either tree or graph patterns) [3, 7, 9, 13, 23, 36]. Such a pattern is a small subtree or a subgraph that needs to have a match in the large database. On top of it, one might impose data values [2, 8, 12, 17, 34], for instance by means of the data equality predicate [14–16, 41]. An example of a pattern is the tree in Fig. 1. It has both the child and the data equality relations. Note that the tree (*i.e.*, the reduct of the *ch* relation) is connected, but the data equality relation on it could be arbitrary: in that example, it has three connected components.

We now generalize this to arbitrary structures. Assume that we have, as before, two disjoint vocabularies σ and τ. A *σ-pattern* over $\sigma\tau$ is a $\sigma\tau$-structure $\mathbb{P}$ so that the Gaifman graph of its σ-reduct is connected (*i.e.*, the graph $\mathcal{G}(\mathbb{P}|_{\sigma})$ is connected). We omit 'over $\sigma\tau$' if it is clear from the context.

Note that a σ-pattern over $\sigma\tau$ does include all the relations in τ as well, but it is only the σ-part of it that needs to be connected (like the tree in Fig. 1, where σ is the child relation).

Pattern logic. For an arbitrary structure $\mathbb{A}$, let $\Delta_{\mathbb{A}}$ denote its *positive diagram*, *i.e.*, the conjunction of all relational atomic formulae true in $\mathbb{A}$. If $|A| = n$, then $\Delta_{\mathbb{A}}$ has n free variables. For instance, if $\mathbb{A}$ is a graph with edges (v_1, v_2) and (v_2, v_3), then $\Delta_{\mathbb{A}}(x_1, x_2, x_3) = E(x_1, x_2) \wedge E(x_2, x_3)$.

Fix σ and τ. Formulae of σ-pattern logic $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ are defined as follows:

$$\begin{array}{rcll} \varphi,\psi & := & S(\bar{x}) \ \mid \ x=y & S\in\sigma \\ & \mid & \Delta_{\mathbb{P}}(\bar{x}) & \mathbb{P} \text{ is a } \sigma\text{-pattern} \\ & \mid & \varphi\vee\psi \ \mid \ \varphi\wedge\psi \ \mid \ \neg\varphi & \\ & \mid & \exists x\varphi \ \mid \ \forall x\varphi\,. & \end{array}$$

That is, we restrict $\mathrm{FO}(\sigma\tau)$ by disallowing arbitrary atomic τ-formulae, and only allowing them as part of pattern formulae $\Delta_{\mathbb{P}}$, so that the σ-reduct of $\mathbb{P}$ is connected. The *existential positive* fragment (in which negation and universal quantification are disallowed) is denoted by $\exists\mathrm{Pos}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$.

Note that this is not the minimal definition of the logic: for instance, $S(\bar{x})$ is a particular case of a formula $\Delta_{\mathbb{P}}(\bar{x})$. In fact the logic can be compactly defined as

$$\varphi,\psi := \ (x=y) \mid \Delta_{\mathbb{P}}(\bar{x}) \mid \varphi\wedge\psi \mid \neg\varphi \mid \exists x\varphi.$$

The formula $\Delta_{\mathbb{P}}(\bar{x})$ is a quantifier-free $\mathrm{FO}(\sigma\tau)$ formula; in particular its quantifier rank is zero. Note that if $\mathbb{P}$ has universe $\{v_1,\ldots,v_n\}$, then $\mathbb{A}\models\Delta_{\mathbb{P}}(a_1,\ldots,a_n)$ if and only if the mapping $f : v_i \mapsto a_i$, for $1\le i\le n$, is a homomorphism from $\mathbb{P}$ to $\mathbb{A}$.

Patterns as those used for querying XML documents with data values are indeed such: one specifies a small connected part of the tree, and then adds an arbitrary data equality relation over it, see [8, 12, 17, 36, 41].

τ-local formulae. We now present a slightly different way of looking at the pattern logic, by allowing τ-predicates, but only in a "local" context. The resulting logic will be called $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$. The idea of such local uses of predicates appeared previously not only in connection with patterns in data models, but also in connection with expressivity of logics in the finite [20].

To define it, we need an FO formula $\mathit{conn}_{\sigma}(\bar{z})$ expressing that the σ-substructure induced by $\bar{z}$ is connected. That is, $\mathbb{A}\models\mathit{conn}_{\sigma}(\bar{a})$ if and only if $\mathcal{G}((\mathbb{A}|_{\sigma})|_{\bar{a}})$ has only one connected component. Note that while in general connectivity is not FO-definable, in this case we test for connectivity of a graph of a fixed size $|\bar{z}|$. Thus, $\mathit{conn}_{\sigma}(\bar{z})$ is an FO (in fact, positive, quantifier-free) formula: one simple takes the disjunctions of positive diagrams of all connected graphs with $|\bar{z}|$ vertices.

Now we define $\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ as

$$\begin{array}{rcll} \varphi,\psi & := & S(\bar{x}) \ \mid \ x=y & S\in\sigma \\ & \mid & T(\bar{x})\wedge \mathit{conn}_{\sigma}(\bar{x},\bar{y}) & T\in\tau \\ & \mid & \varphi\vee\psi \ \mid \ \varphi\wedge\psi \ \mid \ \neg\varphi & \\ & \mid & \exists x\varphi \ \mid \ \forall x\varphi & \end{array}$$

We write $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ for the existential positive fragment, without $\neg$ and $\forall$.

Patterns and locality give us two different ways of looking at the same logic:

Proposition 3.1. *For all σ and τ, we have* $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)=\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ *and* $\exists\mathrm{Pos}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)=\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$.

Proof. For any σ-pattern formula $\Delta_{\mathbb{P}}(\bar{x})$, let $\varphi(\bar{x})\in\exists\mathrm{Pos}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ be the formula that results from replacing every occurrence of $T(\bar{y})$ so that $T\in\tau$ with $T(\bar{y})\wedge\mathit{conn}_{\sigma}(\bar{y},\bar{x})$ in $\Delta_{\mathbb{P}}(\bar{x})$. Note that since the σ-reduct of the Gaifman graph of $\mathbb{P}$ is connected, we have that $\Delta_{\mathbb{P}}(\bar{x})$ is equivalent to $\varphi(\bar{x})$. Then, it follows that every $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ formula is equivalent to a $\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ formula; and every $\exists\mathrm{Pos}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ formula is equivalent to a $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ formula.

Conversely, any conjunctive quantifier-free $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ formula $\varphi(\bar{x})$ is equivalent to a finite disjunction of σ-pattern formulae $\bigvee_i\Delta_{\mathbb{P}_i}(\bar{x})$. Indeed, it is easy to see that the *canonical structure* $\mathbb{A}_{\varphi'}$ of $\varphi'=\exists\bar{x}.\varphi$ is a disjoint union $\mathbb{P}_1\oplus\cdots\oplus\mathbb{P}_n$ of structures $\mathbb{P}_i$ where $\mathcal{G}(\mathbb{P}_i|_{\sigma})$ is connected for every $i\in[1,n]$. This is a consequence of the syntactic restriction of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ requiring any two elements in a tuple of a relation $T\in\tau$ to be connected through σ-relations. In other words, $\mathbb{A}_{\varphi'}$ is a disjoint union of *σ-patterns* over $\sigma\tau$. Thus, $\varphi(\bar{x})$ is equivalent to $\bigvee_{i\in[1,n]}\Delta_{\mathbb{P}_i}(\bar{y}_i)$, where $\bar{y}_i$, ranging over $\bar{x}$, are the variables of the domain of $\mathbb{P}_i$. Therefore, every $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ formula is equivalent to a $\exists\mathrm{Pos}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ formula; and every $\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$ formula is equivalent to a $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$. □

For a graph $\mathbf{G}$, let $\mathbf{G}^r$ stand for its r-step transitive closure, i.e., a graph in which we have an edge (u,v) if there is a path from u to v of length at most r in $\mathbf{G}$. Given a $\sigma\tau$-structure $\mathbb{A}$, we say that it is *(τ,r)-local* if $\mathcal{G}(\mathbb{A}|_{\tau})\subseteq\mathcal{G}(\mathbb{A}|_{\sigma})^r$. The *$(\tau,r)$-localization* of $\mathbb{A}$ is its maximal (τ,r)-local substructure. It can be constructed as follows: look at all relations of τ, and remove all tuples $\bar{a}$ from them that have two components a,a' at distance greater than r in $\mathcal{G}(\mathbb{A}|_{\sigma})$.

Proposition 3.2. *Let $\varphi(\bar{x})$ be a formula of $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ (or of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}(\sigma\tau)$). There exists $r\ge 0$ so that $\varphi(\bar{a})$ is true in $\mathbb{A}$ if and only if it is true in the (τ,r)-localization of $\mathbb{A}$, for every tuple $\bar{a}$ of elements of $\mathbb{A}$.*

Proof. Let $\varphi(x_1,\ldots,x_n)\in\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ with quantifier rank k, and let $r=k+n$. Suppose $\mathbb{A}$ is a $\sigma\tau$-structure, $R\in\tau$, and $(b_1,\ldots,b_m)\in R^{\mathbb{A}}$, so that for some $i,j\in[1,m]$, b_i and b_j are at distance $>r$, that is, $\mathrm{N}^{\mathbb{A}|_{\sigma}}_{r}(b_i)\cap\mathrm{N}^{\mathbb{A}|_{\sigma}}_{r}(b_j)=\emptyset$. Let $\mathbb{A}'$ be the structure resulting from removing $(b_1,\ldots,b_m)$ from $R^{\mathbb{A}}$ in $\mathbb{A}$. One can easily show by induction on the quantifier rank of φ that, for every $a_1,\ldots,a_n\in A$ we have $\mathbb{A}\models\varphi(a_1,\ldots,a_n)\Leftrightarrow\mathbb{A}'\models\varphi(a_1,\ldots,a_n)$. This is because, by definition of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$, φ can only test for $R(x_1,\ldots,x_m)$ in conjunction with $\mathit{conn}_{\sigma}(x_1,\ldots,x_m,y_1,\ldots,y_s)$ for some $y_1,\ldots,y_s$, which specifies that the elements denoted by $x_1,\ldots,x_m,y_1,\ldots,y_s$ form a connected component, and thus that every pair x_i,x_j is at distance $\le|\{x_1,\ldots,x_m,y_1,\ldots,y_s\}|\le n+k=r$. By repeating this argument for all tuples $\bar{b}$ in a relation in τ which are not r-local, it follows that, for every structure $\mathbb{A}$ and $a_1,\ldots,a_n$ in $\mathbb{A}$, we have $\mathbb{A}\models\varphi(a_1,\ldots,a_n)$ if and only if $\hat{\mathbb{A}}\models\varphi(a_1,\ldots,a_n)$, where $\hat{\mathbb{A}}$ is the (τ,r)-localization of $\mathbb{A}$. □

It follows that if we define τ-*conn*$(\mathbb{A})$ as $\mathbb{A}|_{A_1}\oplus\cdots\oplus\mathbb{A}|_{A_n}$, where the A_i's are the maximal connected components of $\mathcal{G}(\mathbb{A}|_{\sigma})$, then $\mathbb{A}$ and τ-*conn*$(\mathbb{A})$ agree on all $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ formulae.

Some (τ,r)-local relations always exist: for instance, if τ contains a single binary relation symbol S, then every structure $\mathbb{A}$ with a connected Gaifman graph has a $(S,3)$-local *successor* relation; this follows from the fact that the cube of any connected graph is Hamiltonian [31].

4. Locality of pattern logics

The goal of this section is to show strong locality results for the pattern logic: one can prove analogs of Hanf and Gaifman theorems for $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ where neighborhoods are defined with respect to Gaifman graphs of σ-reducts only, *i.e.*, τ-relations do not count for computing the distance between two elements. This makes neighborhoods smaller, and thus easier to make isomorphic; consequently it makes these versions of locality easier to apply. One particular application is the separation of $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ from $\mathrm{FO}(\sigma\tau)$; others will be crucial for proving homomorphism preservation and decidability results.

For a $\sigma\tau$-structure $\mathbb{A}$, we use $\delta_{\sigma}(a,b)$ to denote the distance between a and b in $\mathcal{G}(\mathbb{A}|_{\sigma})$, the Gaifman graph that takes into account only σ-relations. Note that $\delta_{\sigma}(x,y)\le r$ can be expressed as an existential-positive $\mathrm{FO}(\sigma)$ formula, and $\delta_{\sigma}(x,y)>r$ as its negation. For $\bar{a}=(a_1,\ldots,a_n)$, by $\mathrm{N}^{\mathbb{A}|_{\sigma}}_{r}(\bar{a})$ we mean the substructure of $\mathbb{A}$ induced by the set $\{b \mid \delta_\sigma(b, a_i) \leq r \text{ for some } i \leq n\}$, together with n additional constants, interpreted as the elements of $\bar{a}$. Note that while the elements of $\mathrm{N}_r^{\mathbb{A}|_\sigma}(\bar{a})$ are those at distance $\leq r$ from $\bar{a}$ in $\mathcal{G}(\mathbb{A}|_\sigma)$, the structure $\mathrm{N}_r^{\mathbb{A}|_\sigma}(a)$ contains all the induced relations of σ and τ.

Given two $\sigma\tau$-structures $\mathbb{A}$ and $\mathbb{B}$, we write $(\mathbb{A}, \bar{a}) \leftrightarrows_d^\tau (\mathbb{B}, \bar{b})$ for $\bar{a} \in A^n$ and $\bar{b} \in B^n$ if there exists a bijection $f : A \to B$ such that for every $c \in A$,

$$\mathrm{N}_d^{\mathbb{A}|_\sigma}(\bar{a}c) \cong \mathrm{N}_d^{\mathbb{B}|_\sigma}(\bar{b}f(c)).$$

We say that a formula $\varphi(\bar{x})$ is *Hanf-τ-local* if there exists a number d so that $(\mathbb{A}, \bar{a}) \leftrightarrows_d^\tau (\mathbb{B}, \bar{b})$ implies that $\mathbb{A} \models \varphi(\bar{a})$ if and only if $\mathbb{B} \models \varphi(\bar{b})$.

Theorem 4.1. *Every formula of* $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ *is Hanf-τ-local.*

Proof. This can be seen as a consequence of Proposition 3.2 and Hanf-locality of FO. Given a $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ formula $\varphi(\bar{x})$, let d be so that $(\mathbb{A}, \bar{a}) \leftrightarrows_d (\mathbb{B}, \bar{b})$ implies that $\mathbb{A} \models \varphi(\bar{a})$ if and only if $\mathbb{B} \models \varphi(\bar{b})$. It exists by Hanf-locality of FO. Let r be the radius yielded by Proposition 3.2, so that $\mathbb{A} \models \varphi(\bar{a})$ iff $\mathbb{A}^{\tau,r} \models \varphi(\bar{a})$, where $\mathbb{A}^{\tau,r}$ is the (τ, r)-localization of $\mathbb{A}$. We then have that $d \cdot r$ is the necessary radius. Indeed, if $(\mathbb{A}, \bar{a}) \leftrightarrows_{d\cdot r}^\tau (\mathbb{B}, \bar{b})$, then $(\mathbb{A}^{\tau,r}, \bar{a}) \leftrightarrows_{d\cdot r}^\tau (\mathbb{B}^{\tau,r}, \bar{b})$ and therefore $(\mathbb{A}^{\tau,r}, \bar{a}) \leftrightarrows_d (\mathbb{B}^{\tau,r}, \bar{b})$, which means that $\mathbb{A}^{\tau,r} \models \varphi(\bar{a})$ if and only if $\mathbb{B}^{\tau,r} \models \varphi(\bar{b})$. Since $\mathbb{A}^{\tau,r} \models \varphi(\bar{a}) \Leftrightarrow \mathbb{A} \models \varphi(\bar{a})$ and $\mathbb{B}^{\tau,r} \models \varphi(\bar{a}) \Leftrightarrow \mathbb{B} \models \varphi(\bar{a})$, it then follows that $\mathbb{A} \models \varphi(\bar{a})$ iff $\mathbb{B} \models \varphi(\bar{b})$. □

If τ has only unary predicates, then $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau) = \mathrm{FO}_\tau^{\mathrm{LOC}}(\sigma\tau) = \mathrm{FO}(\sigma\tau)$, since $\mathit{conn}_\sigma(x)$ is true for each x. However, with a single non-unary predicate in τ, locality provides a separation result:

Corollary 4.2. $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau) \subsetneq \mathrm{FO}(\sigma\tau)$ *if and only if τ has at least one relation of arity > 1.*

Proof. Let τ have a relation T of arity $m > 1$. Let φ say that T is nonempty, *i.e.*, $\exists x_1, \ldots, x_m\ T(x_1, \ldots, x_m)$. Assume φ is definable in $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$, and let d witness its Hanf-τ-locality. Take $\mathbb{A}$ to be a structure on an m-element universe $\{a_1, \ldots, a_m\}$ in which all σ and τ relations are empty, and let $\mathbb{B}$ with the universe $\{b_1, \ldots, b_m\}$ be like $\mathbb{A}$ except that $T^{\mathbb{B}}$ contains a single tuple $(b_1, \ldots .b_m)$. Since each neighborhood $\mathrm{N}_r^{\mathbb{A}|_\sigma}(a)$ is a singleton with all relations empty (and likewise for $\mathbb{B}$), the map $f : a_i \mapsto b_i, i \leq m$ witnesses $\mathbb{A} \leftrightarrows_d^\tau \mathbb{B}$. However, $\mathbb{A} \models \neg\varphi$ while $B \models \varphi$. This contradicts Theorem 4.1 and shows that φ is not definable in $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$. □

As a consequence of Hanf-τ-locality, we have the following condition, known as Gaifman locality for formulae with free variables [32] (except that we again only look at neighborhoods defined by σ-relations):

Corollary 4.3. *For every* $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ *formula* $\varphi(\bar{x})$*, there is a number $r \geq 0$ so that $\mathrm{N}_r^{\mathbb{A}|_\sigma}(\bar{a}_1) \cong \mathrm{N}_r^{\mathbb{A}|_\sigma}(\bar{a}_2)$ implies that $\bar{a}_1$ and $\bar{a}_2$ are indistinguishable by φ, i.e., $\mathbb{A} \models \varphi(\bar{a}_1) \leftrightarrow \varphi(\bar{a}_2)$.*

We can also show a strengthening of Gaifman's theorem for sentences. A *basic* $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ *local sentence* is of the form

$$\exists x_1 \ldots x_n \Big(\bigwedge_{1 \leq i < j \leq n} \delta_\sigma(x_i, x_j) > 2r \wedge \bigwedge_{i \leq n} \psi^{r,\sigma}(x_i) \Big), \qquad (\dagger)$$

where ψ is an $\mathrm{FO}(\sigma\tau)$ formula with one free variable, and $\psi^{r,\sigma}(x)$ stands for the relativization of ψ to $\mathrm{N}_r^{\mathbb{A}|_\sigma}(x)$; that is, the result of replacing in ψ every subformula of the form $\exists y\theta$ with $\exists y(\delta_\sigma(x, y) \leq r \wedge \theta)$, and every subformula of the form $\forall y\theta$ with $\forall y(\delta_\sigma(x, y) \leq r \to \theta)$. The *locality radius* of a basic local sentence is r. Its *width* is n. The formula ψ is called the *local condition*.

While syntactically (†) is not an $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ sentence, it can be expressed in this logic. Indeed, it is easy to see that (†) is definable in $\mathrm{FO}_\tau^{\mathrm{LOC}}(\sigma\tau)$: the first conjunct only refers to σ-relations, and in $\psi^{r,\sigma}(x)$, quantification is relativized to the r-neighborhood of x, with respect to σ. Thus, for any occurrence of a τ relation $T(z_1, \ldots, z_k)$, we know that all the z_i's must belong to the same connected component of $\mathcal{G}(\mathbb{A}|_\sigma)$, and in particular be at distance at most $2r$ from each other. Hence, each such atomic formula can be replaced by $\exists \bar{y}\ (T(\bar{z}) \wedge \mathit{conn}_\sigma(\bar{z}, x, \bar{y}))$, where $\bar{y}$ has $k \cdot (r-1)$ variables witnessing paths from each of the z_i's to x of length at most r. This turns (†) into syntactically proper shape of an $\mathrm{FO}_\tau^{\mathrm{LOC}}(\sigma\tau)$ sentence, and thus by Proposition 3.1 it is also an $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ sentence.

We now have an analog of Gaifman's theorem:

Theorem 4.4. *Every* $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ *sentence is equivalent to a Boolean combination of basic* $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ *local sentences (†).*

Proof. We make use again of Proposition 3.2, and of Gaifman's Theorem for FO. Given a $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ sentence φ, by Gaifman's Theorem for FO it is equivalent to a Boolean combination of basic FO local sentences $\psi_1, \ldots, \psi_n$ of the form (†). In particular, it is equivalent to such a boolean combination on the class of (τ, r)-localized structures, where r is as in Propositon 3.2. Over these structures, one can see that each basic FO local sentence ψ_i with locality radius r_i is actually equivalent to a basic $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ sentence with locality radius $\leq r \cdot r_i$. We then have that over (τ, r)-localized structures φ is equivalent to a Boolean combination of basic $\mathrm{FO}_\sigma^{\mathrm{PAT}}(\sigma\tau)$ local sentences. Then, by Propositon 3.2, this generalizes to the class of all structures. □

Note that, like for the analog of Hanf's theorem, we define distance with respect to σ relations only: formulae $\psi^{r,\sigma}$ can mention both σ and τ relations, but the radius of quantification is determined just by σ. This is different from formulae ψ^r used in original Gaifman's theorem, which define neighborhoods with respect to all vocabulary predicates (and thus neighborhoods could be larger). In particular, it is easier to satisfy a basic $\mathrm{FO}_\tau^{\mathrm{LOC}}$ local sentence than a basic FO local sentence, which will make this version of Gaifman's theorem more valuable to us in the next sections.

5. Homomorphism preservation theorem

We now use locality results from the previous section to prove a *homomorphism preservation theorem (hpt)* for a large class of structures including ranked data forests. But first we explain the concept and its applications in computer science.

5.1 HPT and its applications

Preservation theorems relate semantic properties of FO sentences to their syntactic properties: for instance, a sentence is preserved under extensions if and only if it is equivalent to an existential sentence. A sentence φ is *preserved under homomorphisms* on a class $\mathbf{C}$ of structures if, whenever we have a homomorphism $\mathbb{A} \to \mathbb{B}$ between two structures of $\mathbf{C}$, then $\mathbb{A} \models \varphi$ implies $\mathbb{B} \models \varphi$. Then the hpt for $\mathbf{C}$ says that an FO sentence φ is preserved under homomorphisms on $\mathbf{C}$ structures if and only if it is equivalent to an existential positive sentence on $\mathbf{C}$. It is very easy to check that $\exists$Pos sentences are preserved under homomorphisms; the crucial part is the reverse direction.

When $\mathbf{C}$ is the class of all structures (finite and infinite), this is a textbook model theory result. For finite structures, it had remained an open problem for decades, before being confirmed by Rossman [39] a few years ago. The hpt is also known for some classes of well-behaved finite structures, for instance structures of bounded treewidth [6, 18]. These are not consequences of Rossman's theorem (since such classes of structures are not FO-definable), and in fact versions of hpt in the finite tend to be quite hard results to show.

At the same time, these preservation results often find applications in computer science. The reason for this is the special importance of the class of $\exists$Pos, especially in databases and constraint satisfaction [1, 26]. In terms of their expressiveness, $\exists$Pos formulae are equivalent to *unions of conjunctive queries*, or positive relational algebra queries, cf. [1]. These are very common queries, that also play a special role in many contemporary applications of databases, for instance, data integration and exchange. In fact, hpt was used for determining classes of mappings between database schemas that one needs to use in data exchange applications [43] and for characterizing properties of the chase procedure under various types of mappings [21]. In constraint satisfaction, hpt is needed to provide characterization of FO-definable CSP problems [5].

Another area where the hpt proved very useful is handling incomplete data [1, 29]. Suppose we have an FO structure (*i.e.*, a relational database) whose elements come from two sets: $\mathcal{C}$ of *constants* and $\mathcal{V}$ of *variables* (also called *nulls* in the database terminology). For instance, $G = \{(1, x), (x, y), (y, 2)\}$ is a graph whose nodes are $1, 2 \in \mathcal{C}$ and $x, y \in \mathcal{V}$. We view variables as instances of incompleteness: we do not know what these nodes actually are. Another instance of incompleteness is that some edges may be missing altogether. Thus, one defines the *semantics* of an incomplete structure $\mathbb{A}$ as the set $[\![\mathbb{A}]\!]$ of structures $\mathbb{A}'$ over $\mathcal{C}$ so that there is a homomorphism $f : \mathbb{A} \to \mathbb{A}'$ which is the identity on $\mathcal{C}$. For instance, $G' = \{(1, 3), (3, 4), (4, 2), (5, 3)\}$ is in $[\![G]\!]$, as witnessed by the homomorphism $x \mapsto 3,\ y \mapsto 4$.

If we now have an FO sentence (query) φ, the standard way of answering it over an incomplete structure $\mathbb{A}$ is to find *certain answers* $\Box\varphi$: that is, $\mathbb{A} \models \Box\varphi$ iff $\mathbb{A}' \models \varphi$ for each $\mathbb{A}' \in [\![A]\!]$ (see [1, 29]). In general, for arbitrary FO formulae, checking if $\mathbb{A} \models \Box\varphi$ is an undecidable problem, as it is a form of finite validity. However, if φ is an $\exists$Pos sentence, then

$$\mathbb{A} \models \Box\varphi \;\;\Leftrightarrow\;\; \mathbb{A} \models \varphi \qquad (1)$$

(cf. [29]). This says that to evaluate φ with certainty over an incomplete database, we just have to use the standard query evaluation engine and simply evaluate φ itself. This is a very desirable property, and thus it is natural to ask how far it extends. With the help of hpt in the finite one can answer this: if (1) holds for an FO sentence φ, then φ is equivalent to an $\exists$Pos sentence [33]. Thus, unions of conjunctive queries are the maximal class of FO queries for which certain answers can be found by straightforward query evaluation.

The connection between hpt and query evaluation does not stop here: it extends to other relational semantics of incompleteness and other notions of homomorphisms [25]. What is more problematic is extending it to XML with incomplete information [2, 8]. Even though the underlying structure of XML documents is that of trees, adding arbitrary data-equality predicates destroys their small treewidth and all other properties for which the hpt is known [6, 18].

Thus, in addition to the general interest in hpt results in finite model theory, they come with concrete applications, and there are open problems related to them, in particular in the context of XML with data values. One of those will be settled in the next section.

5.2 HPT for the pattern logic

We first state formally what it means for a logic to satisfy the hpt over a class $\mathbf{C}$. Suppose we have a logic $\mathcal{L}$ that has an existential positive fragment $\exists$Pos$\mathcal{L}$ (for instance, FO, or $\text{FO}_\sigma^{\text{PAT}}$, or $\text{FO}_\tau^{\text{LOC}}$). A homomorphism between two structures $\mathbb{A}, \mathbb{B}$ of the same vocabulary is a mapping $h : A \to B$ so that for each relation symbol R in the vocabulary, if $\bar{u} \in R^{\mathbb{A}}$, then $h(\bar{u}) \in R^{\mathbb{B}}$. A sentence φ of $\mathcal{L}$ is preserved under homomorphisms on $\mathbf{C}$ if for every two structures $\mathbb{A}, \mathbb{B} \in \mathbf{C}$ such that there is a homomorphism $h : \mathbb{A} \to \mathbb{B}$, we have that $\mathbb{A} \models \varphi$ implies $\mathbb{B} \models \varphi$. Then the *homomorphism preservation theorem (hpt) holds for* $\mathcal{L}$ *over* $\mathbf{C}$ if the following two conditions are equivalent for a sentence φ of $\mathcal{L}$:

- φ is preserved under homomorphisms on $\mathbf{C}$;
- φ is equivalent to an $\exists$Pos$\mathcal{L}$ sentence on $\mathbf{C}$.

We now state an hpt result for the pattern logic. Let $\mathbf{C}$ be a class of $\sigma\tau$-structures. Given a structure $\mathbb{A}$, we say that a set $X \subseteq A$ is *r-scattered* if $\delta(x, y) > 2r$ for all distinct $x, y \in X$, and it is *σ, r-scattered* if $\delta_\sigma(x, y) > 2r$ for all distinct $x, y \in X$. We say that $\mathbf{C}$ is *σ-wide* if for all $r, m > 0$, there exists a number n so that each $\mathbb{A} \in \mathbf{C}$ with $|A| > n$ has a σ, r-scattered set of size m. An example is the class of structures whose σ-reducts have degree bounded by a fixed number k.

Theorem 5.1. *Let $\mathbf{C}$ be a class of finite $\sigma\tau$-structures that is closed under disjoint unions and induced substructures, and is σ-wide. Then the homomorphism preservation theorem holds for $\text{FO}_\sigma^{\text{PAT}}(\sigma\tau)$ over $\mathbf{C}$.*

Before we outline the proof of this theorem, we mention a few applications. We start with the class of data forests. Recall that in this case, σ contains the child relation *ch*, while τ contains unary labeling predicates and the binary data equality predicate $\sim$ (actually, it makes no difference whether the unary predicates belong to σ or to τ). Note that the class of data forests $\mathcal{DF}$ is closed under disjoint unions and induced substructures. Thus, the hpt holds for every wide class of data forests with respect to the child relation. Clearly, the class of ranked data forests is such, and hence we get:

Corollary 5.2. *The hpt holds for the pattern logic over $\mathcal{DF}_k$, for every $k > 0$.*

We can further extend this using the notion of moderate degree, or degree bounded by $n^{o(1)}$, from [40]. We say that a class $\mathbf{C}$ of σ-structures is of *moderate degree* if all the degrees in their Gaifman graphs are bounded by $n^{o(1)}$ for all sufficiently large structures. That is, for every $\varepsilon > 0$, there exists $N > 0$ so that for every structure $\mathbb{A} \in \mathbf{C}$ of size $n \geq N$, the maximum degree in $\mathcal{G}(\mathbb{A})$ is less than n^ε. A class $\mathbf{C}$ of data forests is of moderate degree if the class of its *ch*-reducts (*i.e.*, the trees themselves) is.

One can show that the class of data forests of moderate degree is *ch*-wide. Thus:

Proposition 5.3. *The hpt holds for the pattern logic over data forests of moderate degree.*

Proof. Let $\mathbf{C}$ be a class of forests of moderate degree; assume that $f(n)$ is the maximum degree in a forest having n nodes. We show that $\mathbf{C}$ is *ch*-wide; that will suffice to conclude the corollary.

Fix $r, m > 0$. Let $d = (2r + 2)m$, and let $\varepsilon = \frac{1}{d+1}$. Then, for all $n \geq m^{d+1}$, we have $\sqrt[d]{n/m} \;\geq\; n^\varepsilon$. Thus, for some N_0 depending on m and r, we have for all $n \geq N_0$:

$$\frac{n}{m} \;\geq\; f(n)^d\,. \qquad (2)$$

Indeed, otherwise we would have $\frac{n}{m} < f(n)^d$ for arbitrarily large n, and thus $f(n) > \left(\frac{n}{m}\right)^{1/d} \geq n^\varepsilon$ for arbitrarily large n, which is impossible since $\mathbf{C}$ is of moderate degree.

Suppose we have a forest of size n for $n > N_0$. Assume that it has m or more connected components. Then it clearly has an r-scattered set of size m. So assume it has fewer than m connected components (trees); then at least one of them, t, is of size at least n/m. We know that if we we have a tree of branching factor at most k and size at least k^d, then it has a branch of length at least d. Since

$|t| \geq \frac{n}{m} \geq f(n)^d$, by (2), we see that t_0 has a branch of length $d = (2r+2)m$. Therefore, this branch alone has an r-scattered set of size m. □

This has direct implications for evaluating queries over XML documents with incomplete information. Such documents are modeled as forests where each node is assigned a data value from either a set $\mathcal{C}$ of constants or a set $\mathcal{V}$ of variables; one can think of them as direct analogs of incomplete relational databases. We refer to them as *XML forests*. These provide the actual model of incomplete XML that is abstracted as data forests; in the latter, instead of having a potentially infinite set of data values, we only have the data equality predicate. Some models of incomplete XML deal with more complex structural incompleteness than the forest structure [2, 8], but we do not consider them here as they quickly lead to intractability of query evaluation [8].

Query languages considered for incomplete XML documents are usually pattern logics [2, 8, 13, 17] or their fragments. As explained in section 5.1, it is crucial for us to know when checking $\mathbb{A} \models \Box\varphi$ is equivalent to checking $\mathbb{A} \models \varphi$, *i.e.*, when finding certain answers can be replaced by straightforward query evaluation. In the database terminology, one says that *naïve evaluation works* for φ. While this is known for relational databases [33], for XML the problem is open. We can now settle it for the class of XML documents whose branching degree is not very large.

Corollary 5.4. *Let φ be a $\mathrm{FO}^{\mathrm{PAT}}_{ch}$ sentence and $\mathbf{C}$ a class of XML incomplete documents of fixed or moderate degree. Then naïve evaluation works for φ over $\mathbf{C}$ if and only if φ is equivalent to an $\exists\mathrm{Pos}^{\mathrm{PAT}}_{ch}$ sentence over $\mathbf{C}$.*

This follows from Corollary 5.2 and Proposition 5.3, and the result from [25] showing the equivalence of naïve evaluation on a class $\mathbf{C}$ and preservation under homomorphisms on the same class.

5.3 Proof of Theorem 5.1

Due to Proposition 3.1, it suffices to prove the result for $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$. As before, we omit the $\sigma\tau$ vocabulary from notations when it is clear from the context. That each formula of $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}$ is preserved under homomorphisms is immediate by a straightforward induction on the structure of the formula. Thus, we prove the opposite direction.

We say that $\mathbb{B}$ is a *segregated substructure* of $\mathbb{A}$ if there are pairwise disjoint subsets $A_1, \dots, A_n$ of A so that $\mathbb{B} = \mathbb{A}|_{A_1} \oplus \cdots \oplus \mathbb{A}|_{A_n}$. We write $\sqsubseteq_{\mathit{seg}}$ for the segregated substructure relation. Note that a class of structures $\mathbf{C}$ is closed under disjoint unions and induced substructures if and only if it is closed under disjoint unions and segregated substructures.

For a sentence φ preserved under homomorphisms on a class of structures $\mathbf{C}$, we say that $\mathbb{A} \in \mathbf{C}$ is a *$\sqsubseteq_{\mathit{seg}}$-minimal model* of φ in $\mathbf{C}$ if $\mathbb{A} \models \varphi$ and for every proper substructure $\mathbb{B} \sqsubseteq_{\mathit{seg}} \mathbb{A}$, $\mathbb{A} \not\sqsubseteq_{\mathit{seg}} \mathbb{B}$ such that $\mathbb{B} \in \mathbf{C}$, we have $\mathbb{B} \not\models \varphi$.

The term *conjunctive query* [1] denotes formulae of the form $\exists x_1, \dots, x_n\ \theta$, where θ is a conjunction of atomic formulae. Every finite structure $\mathbb{A}$ with n elements gives rise to a *canonical conjunctive query* $\varphi_{\mathbb{A}}$, which is the existential closure of its positive diagram, *i.e.*, $\varphi_{\mathbb{A}} = \exists\bar{x}\Delta_{\mathbb{A}}(\bar{x})$, where $\bar{x}$ is the set of all variables used in $\Delta_{\mathbb{A}}$. Conversely, every conjunctive query $\varphi = \exists x_1, \dots, x_n\ \theta$ gives rise to a *canonical structure* $\mathbb{A}_{\varphi}$ with n elements, known as its *tableau* [1], where the elements of $\mathbb{A}$ are the variables $x_1, \dots, x_n$, and the relations of $\mathbb{A}$ consist of the tuples of variables in the conjuncts of θ. The following is well known (cf. [1, 26]). If $\mathbb{A}$ and $\mathbb{B}$ are two finite structures, then

$$\exists \text{ homomorphism } \mathbb{A} \to \mathbb{B} \;\Leftrightarrow\; \mathbb{B} \models \varphi_{\mathbb{A}} \;\Leftrightarrow\; \varphi_{\mathbb{B}} \models \varphi_{\mathbb{A}}. \qquad (3)$$

Not all structures have a canonical conjunctive query that is a formula of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$. Nevertheless, those structures $\mathbb{A}$ so that $\mathbb{A} = \tau\text{-}\mathit{conn}(\mathbb{A})$ do have a canonical $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ conjunctive query.

Lemma 5.5. *For every $\mathbb{A}$ so that $\mathbb{A} = \tau\text{-}\mathit{conn}(\mathbb{A})$, the canonical conjunctive query $\varphi_{\mathbb{A}}$ is definable in $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$.*

Indeed, note that any two elements a, b of A related by a relation $T \in \tau$ belong to the same connected component, and thus $\delta_{\sigma}(a, b) \leq n$ for $n = |A|$. It is then immediate that $\varphi_{\mathbb{A}}$ is definable in $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$.

Also, from the definition of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ we have that for every structure $\mathbb{A}$ and $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ sentence φ,

$$\mathbb{A} \models \varphi \;\Leftrightarrow\; \tau\text{-}\mathit{conn}(\mathbb{A}) \models \varphi. \qquad (4)$$

Lemma 5.6. *Let $\mathbf{C}$ be a class of finite $\sigma\tau$-structures closed under segregated substructures and let φ be a $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ sentence that is preserved under homomorphisms on $\mathbf{C}$. Then the following are equivalent:*

1. *φ has finitely many $\sqsubseteq_{\mathit{seg}}$-minimal models in $\mathbf{C}$ up to isomorphism.*
2. *φ is equivalent on $\mathbf{C}$ to a $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}$ sentence.*

Proof of the lemma. For the 1. ⇒ 2. direction, take any $\sqsubseteq_{\mathit{seg}}$-minimal model $\mathbb{A}$ of φ. Note that $\mathbb{A}$ must be so that $\tau\text{-}\mathit{conn}(\mathbb{A}) = \mathbb{A}$, since otherwise $\tau\text{-}\mathit{conn}(\mathbb{A})$ would be a strict segregated substructure of $\mathbb{A}$ where $\tau\text{-}\mathit{conn}(\mathbb{A}) \models \varphi$, by (4). Indeed, $\tau\text{-}\mathit{conn}(\mathbb{A}) \in \mathbf{C}$ by the hypothesis, and hence $\mathbb{A}$ would not be $\sqsubseteq_{\mathit{seg}}$-minimal. Let $\varphi_{\mathbb{A}}$ be the canonical conjunctive $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ query of $\mathbb{A}$ (it is definable in $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ due to Lemma 5.5). We define the existential positive formula ψ as the disjunction of all $\varphi_{\mathbb{A}}$ for each $\sqsubseteq_{\mathit{seg}}$-minimal model $\mathbb{A}$ of φ. We next show that ψ and φ are equivalent.

Note that, by (3), we have that $\mathbb{B} \models \varphi_{\mathbb{A}}$ iff there is a homomorphism from $\mathbb{A}$ to $\mathbb{B}$. If $\mathbb{B} \models \psi$ then $\mathbb{B} \models \varphi_{\mathbb{A}}$ for some disjunct $\varphi_{\mathbb{A}}$ of ψ, and hence there is a homomorphism from $\mathbb{A}$ to $\mathbb{B}$, by (3). Since $\mathbb{A} \models \varphi$ and φ is closed under homomorphisms, then $\mathbb{B} \models \varphi$. Conversely, if $\mathbb{B} \models \varphi$ there must be some $\sqsubseteq_{\mathit{seg}}$-minimal $\mathbb{A} \sqsubseteq_{\mathit{seg}} \mathbb{B}$ so that $\mathbb{A} \models \varphi$. Hence, ψ contains $\varphi_{\mathbb{A}}$ as a disjunct, meaning that $\mathbb{A} \models \psi$.

For the 2. ⇒ 1. direction, by propagating disjunctions, we can convert every existential-positive $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}$ formula φ 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 $\mathbb{A}_i$ be the canonical finite structure associated with ψ_i, $1 \leq i \leq m$. Note that such a canonical structure $\mathbb{A}_i$ need not be a member of $\mathbf{C}$. Nonetheless, it is easy to see that every $\sqsubseteq_{\mathit{seg}}$-minimal model $\mathbb{B}$ of φ in $\mathbf{C}$ is equal to a homomorphic image $h(\mathbb{A}_i)$ of one of the canonical finite structures $\mathbb{A}_i$, $1 \leq i \leq m$. Thus, the cardinality of every $\sqsubseteq_{\mathit{seg}}$-minimal model of φ in $\mathbf{C}$ is less than or equal to the maximum cardinality of the canonical finite structures $\mathbb{A}_i$, $1 \leq i \leq m$, which implies that φ has finitely many $\sqsubseteq_{\mathit{seg}}$-minimal models in $\mathbf{C}$. This proves the lemma.

The proof is based on the fact that $\exists\mathrm{Pos}^{\mathrm{LOC}}_{\tau}$ sentence is a union of conjunctive queries (*i.e.*, the $\exists, \wedge$ fragment of $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$), and that the canonical conjunctive query $\varphi_{\mathbb{A}}$ of a structure $\mathbb{A}$ (*i.e.*, the existential closure of the positive diagram of $\mathbb{A}$) is definable in $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ as long as $\mathbb{A} = \tau\text{-}\mathit{conn}(\mathbb{A})$.

The main consequence of Lemma 5.6 is that in order to establish that $\mathbf{C}$ has the homomorphism preservation property, it suffices to establish an upper bound on the size of the $\sqsubseteq_{\mathit{seg}}$-minimal models. This is done by:

Proposition 5.7. *Let $\mathbf{C}$ be a class of $\sigma\tau$-structures that is closed under disjoint unions and induced substructures. For every $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ sentence φ that is preserved under homomorphisms, there are $r, m \in \mathbb{N}$ such that if $\mathbb{A}$ is a $\sqsubseteq_{\mathit{seg}}$-minimal model of φ, then $\mathbb{A}$ does not contain a σ, r-scattered set of size m.*

The idea of the proof is as follows. Suppose φ is an $\mathrm{FO}^{\mathrm{LOC}}_{\tau}$ sentence preserved under homomorphisms. Let $\Sigma = \{\varphi_1, \dots, \varphi_s\}$ be a collection of basic local sentences of the form (†) such that

φ is equivalent to a Boolean combination of them. It exists by Theorem 4.4. For each $i \leq s$, let t_i be the locality radius, n_i the width and $\psi_i^{t_i,\sigma}(x)$ the local condition of φ_i. Also, let $t = \max_i t_i$ and $n = \max_i n_i$. We take $r = 2t$ and $m = 2^{s+|\sigma\tau|} + 1$. For each i, we write $\theta_i(y)$ for the formula $\exists x\big(\delta_\sigma(x, y) \leq t_i \wedge \psi_i^{t_i,\sigma}(x)\big)$.

By means of contradiction, suppose that $\mathbb{A}$ is a model of φ so that $\mathbb{A}$ contains an σ, r-scattered set $\{c_1, \ldots, c_m\}$ of size m. Then, by definition $\mathrm{N}_r^{\mathbb{A}|\sigma}(c_i) \cap \mathrm{N}_r^{\mathbb{A}|\sigma}(c_j) = \emptyset$ for $i \neq j$. Furthermore, since $m > 2^{s+|\sigma\tau|}$, there are i and j with $i \neq j$ such that

- for all l, $\mathbb{A} \models \theta_l(c_i)$ if, and only if, $\mathbb{A} \models \theta_l(c_j)$; and
- for all $S \in \sigma\tau$, $(c_i, \ldots, c_i) \in S^{\mathbb{A}}$ if, and only if, $(c_j, \ldots, c_j) \in S^{\mathbb{A}}$.

We then prove that there is some relation $R \in \sigma\tau$ so that there is some tuple $\bar{e} \in R^{\mathbb{A}}$ so that $\bar{e}$ includes c_i and some other element $a \neq c_i$. This implies that $\mathbb{B} = \mathbb{A}|_{A\setminus\{c_i\}} \oplus \mathbb{A}|_{\{c_i\}}$ is a proper segregated substructure of $\mathbb{A}$, and thus $\mathbb{B} \in \mathbf{C}$. Take $\mathbb{B}_n$ to be the disjoint union of n copies of $\mathbb{B}$. We then prove

$$\mathbb{A} \oplus \mathbb{B}_n \models \varphi \quad \text{iff} \quad \mathbb{B}_n \models \varphi. \tag{5}$$

To prove this, we show that no φ_l distinguishes between $\mathbb{A} \oplus \mathbb{B}_n$ and $\mathbb{B}_n$. By symmetry, we restrict our attention to the case $l = 1$.

[$\Leftarrow$] Since $\mathbb{B}_n$ is isomorphic to an induced substructure of $\mathbb{A} \oplus \mathbb{B}_n$, $\mathbb{A} \oplus \mathbb{B}_n$ satisfies φ_1 if $\mathbb{B}_n$ does.

[$\Rightarrow$] We suppose that $(\mathbb{A}\oplus\mathbb{B}_n)|_\sigma$ has a $\sigma, 2t_1$-scattered subset X of cardinality n_1 such that $\mathrm{N}_{t_1}^{(\mathbb{A}\oplus\mathbb{B}_n)|\sigma}(x) \models \psi_1(x)$ for all $x \in X$ and prove that $\mathbb{B}_n$ has such a subset X' as well. The case $n_1 > 1$ is easy. In this case, the t_1-neighborhood of some $x \in X$ does not contain any of the tuples of E_{c_i}. Then, $\mathrm{N}_{t_1}^{\mathbb{B}|\sigma}(x) = \mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(x)$ and therefore each of the n summands of $\mathbb{B}_n$ has a t_1-neighborhood isomorphic to $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(x)$. Since $n \geq n_1$, we can take n_1 distinct elements $\{x_1, \ldots, x_{n_1}\} = X'$ from $\mathbb{B}_n$ so that for every $j \in [1, n_1]$, x_j is in the jth summand of $\mathbb{B}_n$ and $\mathrm{N}_{t_1}^{\mathbb{B}|\sigma}(x_j) = \mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(x)$ (and therefore satisfies ψ_1).

Suppose then that $n_1 = 1$ and let x be the only element of X. It suffices to prove that $\mathbb{A}$ contains an element y such that $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(y)$ does not contain any of the tuples of E_{c_i} and $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(y) \models \psi_1(y)$. If $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(x)$ does not contain any of the tuples of E_{c_i}, we have finished; so suppose that $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(x)$ contains $\bar{e} \in E_{c_i}$. Then $\delta_\sigma(c_i, x) \leq t_1$ and therefore $\mathrm{N}_{2t_1}^{\mathbb{A}|\sigma}(c_i)$ satisfies $\theta_1(c_i)$. Recall that c_i is equivalent to c_j and $\delta_\sigma(c_i, c_j) > 2r \geq 4t_1$. Hence $\mathrm{N}_{2t_1}^{\mathbb{A}|\sigma}(c_j)$ satisfies $\theta_1(c_j)$ and does not contain any of the tuples of E_{c_i}. Hence there exists $y \in A$ such that $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(y) \models \psi_1(y)$ and $\mathrm{N}_{t_1}^{\mathbb{A}|\sigma}(y)$ is included in $\mathrm{N}_{2t_1}^{\mathbb{A}|\sigma}(c_j)$ and therefore does not contain any of the tuples of E_{c_i}. This concludes the proof of (5).

Since φ is preserved under homomorphisms, and by assumption $\mathbb{A} \models \varphi$, the existence of a homomorphism $\mathbb{A} \to \mathbb{A} \oplus \mathbb{B}_n$ implies $\mathbb{A} \oplus \mathbb{B}_n \models \varphi$. By (5), we get $\mathbb{B}_n \models \varphi$, and since we have a homomorphism $\mathbb{B}_n \to \mathbb{B}$, we conclude $\mathbb{B} \models \varphi$. As $\mathbb{B}$ is a proper segregated substructure of $\mathbb{A}$ in $\mathbf{C}$, we have that $\mathbb{A}$ is not a $\sqsubseteq_{\mathit{seg}}$-minimal model of φ. This contradiction proves the proposition.

We now conclude the proof. Let $\varphi \in \mathrm{FO}_\tau^{\mathrm{LOC}}$ be preserved under homomorphisms on $\mathbf{C}$. By Proposition 5.7, there are $r, m > 0$ so that any $\sqsubseteq_{\mathit{seg}}$-minimal model $\mathbb{A}$ of φ does not have a σ, r-scattered set of size m. Since the class is σ-wide, this implies that there is $n > 0$ so that every $\sqsubseteq_{\mathit{seg}}$-minimal model $\mathbb{A}$ of φ has size at most n. Thus, up to isomorphism, there are finitely many $\sqsubseteq_{\mathit{seg}}$-minimal models $\mathbb{A}$ of φ in $\mathbf{C}$. This, by Lemma 5.6, implies that φ is equivalent to a $\exists\mathrm{Pos}_\tau^{\mathrm{LOC}}$ sentence.

6. Decidability of pattern logics

The main initial direction in the study of the data equality predicate for tree-like structures was to extend decidability results of logics for words and trees to those that admit data values as well [14–16, 41]. The motivation for this is static analysis of XML queries. The well known correspondence of XML structural properties and tree automata (cf. [44]) gives us many decidable static analysis tasks, but the connection with automata fails when data values from an infinite set are added. Indeed, FO is undecidable over data trees (even data words). Hence, it was necessary to impose restrictions. The usual way of doing so is by restricting the number of variables: [16] showed that FO^2, first-order logic with two variables, is decidable over data trees, while FO^3 is not. Similar results hold for data words even in the presence of an order [15], and some extensions, of more automata-theoretic flavor, have been proposed as well [14, 42].

The problem with FO^2 is that it is quite restricted for XML querying. While some useful queries, in particular some fragments of XPath, can be captured by it [35], more expressive fragments of XPath, and crucially, more expressive queries that can use more complex patterns [3, 17, 35, 36, 44] are not captured at all by restricting the number of variables. At the same time, many such querying mechanisms are based on patterns, so perhaps one can recover decidability using pattern logics?

6.1 Decidability of pattern logics

By decidability of course we mean decidability of the satisfiability problem. A logic $\mathcal{L}$ is decidable on a class $\mathbf{C}$ if the following problem is decidable: given a sentence φ of $\mathcal{L}$, check whether there exists $\mathbb{A} \in \mathbf{C}$ so that $\mathbb{A} \models \varphi$.

Our results show that in particular, the pattern logic is decidable on the class of ranked data trees. In fact we prove a more general result that allows other auxiliary relations on trees, not just the data equality predicate (*e.g.*, other equivalence relations or linear orders).

Recall that by τ-trees we mean $\{\mathit{ch}\} \sqcup \tau$-structures where *ch* is interpreted as a child relation of a tree. Thus, we deal with the pattern logic $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}(\{\mathit{ch}\} \sqcup \tau)$ over τ-trees. We now define conditions on τ that make it decidable. Those will go beyond having a single data-equality predicate $\sim$ in τ (*i.e.*, beyond data trees).

Recall that *root*($\mathbb{A}$) stands for the root of a τ-tree, and $\mathbb{A}{\restriction}a$ is the $\{\mathit{ch}\} \sqcup \tau$-substructure induced by the subtree rooted at a. For τ-trees $\mathbb{A}_1, \ldots, \mathbb{A}_m \in \mathbf{C}$ we define $\Sigma_{\mathbf{C}}(\mathbb{A}_1, \ldots, \mathbb{A}_m)$ as the set

$$\begin{aligned}\{\mathbb{A} \in \mathbf{C} \mid \ & \mathit{root}(\mathbb{A}) \text{ has exactly } m \text{ children } a_1, \ldots, a_m, \\ & \text{and for all } i \leq m \text{ we have } \mathbb{A}{\restriction}a_i = \mathbb{A}_i\}.\end{aligned}$$

We say that $\mathbf{C}$ is an *effective inductive class* of τ-trees if three conditions hold:

1. Testing whether a structure $\mathbb{A}$ is in $\mathbf{C}$ is decidable;
2. $\mathbf{C}$ is closed under subtrees, *i.e.*, for every $\mathbb{A} \in \mathbf{C}$ and $a \in A$, the structure $\mathbb{A}{\restriction}a$ is in $\mathbf{C}$;
3. for each $r \geq 0$ and structures $\mathbb{A}_1, \ldots, \mathbb{A}_m \in \mathbf{C}$, the set $\{\mathrm{N}_r^{\mathbb{A}|\sigma}(\mathit{root}(\mathbb{A})) \mid \mathbb{A} \in \Sigma_{\mathbf{C}}(\mathbb{A}_1, \ldots, \mathbb{A}_m)\}$ is computable from $\{\mathrm{N}_r^{\mathbb{A}_i|\sigma}(\mathit{root}(\mathbb{A}_i)) \mid 1 \leq i \leq m\}$.

Theorem 6.1. *Let $\mathbf{C}$ be an effective inductive class of k-ranked τ-trees. Then $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}(\{\mathit{ch}\} \sqcup \tau)$ is decidable on $\mathbf{C}$.*

Before sketching the proof, we give a few corollaries. An example of an effective inductive class is the class of trees with

- the sibling order relation *sib*,
- n equivalence relations $\sim_1, \ldots, \sim_n$,

- m linear orders $\leq_1, \dots, \leq_m$, and
- l partial orders $\preceq_1, \dots, \preceq_l$.

This is an extension of the class of data trees, since one of the equivalence relations can be viewed as the data equality predicate.

Corollary 6.2. *The pattern logic* $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}$ *is decidable on ranked data trees, even expanded with a sibling order and an arbitrary number of equivalence relations and partial and linear orders.*

The satisfiability problem for $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}$ is not affected by adding a prefix of existential monadic second order quantifiers: the proof goes through as before, with extra unary relations. Thus, if we define $\exists\mathrm{MSO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ as formulae of the form $\exists X_1 \dots X_n\, \varphi$, where φ is an $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma \sqcup \tau \sqcup \{X_1, \dots, X_n\})$ formula and $X_1, \dots, X_n$ are unary predicates, then we obtain:

Corollary 6.3. *The logic* $\exists\mathrm{MSO}^{\mathrm{PAT}}_{\mathit{ch}}$ *is decidable on ranked data trees even with a sibling order and an arbitrary number of equivalence relations and partial and linear orders.*

The reason Corollary 6.3 is useful is that many static analysis tasks for XML are considered with respect to a regular tree language [41], as regular languages provide a standard abstraction of XML schema languages [44]. Thus, satisfiability of a logic $\mathcal{L}$ *relative to a tree automaton* is the following problem: given a tree automaton $\mathcal{TA}$, and a formula φ of $\mathcal{L}$, is there a tree (with extra relations such as the data equality relation) that satisfies φ and is accepted by $\mathcal{TA}$?

Since encoding a tree automaton on a ranked tree is expressible in $\exists$MSO over vocabulary $\{\mathit{ch}, \mathit{sib}\}$, with references to *sib* local (*i.e.*, restricted to *ch*-neighborhoods of radius 2), we obtain:

Corollary 6.4. *If* $\mathbf{C}$ *is an effective inductive class of ranked* τ*-trees, then the satisfiability problem for* $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}(\{\mathit{ch}\} \sqcup \tau)$ *relative to a tree automaton is decidable on* $\mathbf{C}$.

In particular, it is decidable whether a data tree satisfying a $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}$ sentence exists in a given regular tree language.

Sketch of the proof of Theorem 6.1. The decidability result is another consequence of Gaifman locality theorem for the pattern logic. Let $\sigma = \{\mathit{ch}\}$, and let φ be a $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}(\sigma\tau)$ sentence, where τ is as in the statement of the theorem. In the proof we assume that τ contains a sibling order relation *sib*. This is without any loss of generality, since it can be shown that any effective inductive class of ranked trees continues to be effective inductive if we add the sibling order relation. By Theorem 4.4, φ is equivalent to a Boolean combination of basic $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}$ local sentences. Let r be the bound on the locality radius and let n be the bound on the width of these basic $\mathrm{FO}^{\mathrm{PAT}}_{\sigma}$ local sentences.

The *sibling number* of an element a of a tree $\mathbb{A}$ is one plus the number of previous siblings it has in the order *sib* (note that the sibling number of the root is 1). Given two elements a, a' from a tree $\mathbb{A}$ so that a is an ancestor of a', we write $\mathit{path}(a, a') \in \mathbb{N}^*$ to denote the string representing the path between a and a':

- if $a = a'$ then $\mathit{path}(a, a') = \varepsilon$, the empty string,
- otherwise, $\mathit{path}(a, a') = \mathit{path}(a, a'') \cdot i$ where i is the sibling number of a', and a'' is the parent of a'.

For any k-ranked ordered τ-tree $\mathbb{A}$ with root a, consider the function $\xi_{\mathbb{A}}$ with domain $[1, r] \times [1, n]$, so that for every $\hat{r} \in [1, r]$ and $\hat{n} \in [1, n]$, its value $\xi_{\mathbb{A}}(\hat{r}, \hat{n})$ is the set of all $\hat{n}$-tuples $\langle(\mathrm{N}^{\mathbb{A}|\sigma}_{\hat{r}}(a_1), p_1), \dots, (\mathrm{N}^{\mathbb{A}|\sigma}_{\hat{r}}(a_{\hat{n}}), p_{\hat{n}})\rangle$ so that $a_1, \dots, a_{\hat{n}} \in A$, the distance $\delta_\sigma(a_i, a_j)$ is greater than $2 \cdot \hat{r}$ whenever $i \neq j$, and p_i is defined as $\mathit{path}(a, a_i)$ if $\delta_\sigma(a, a_i) \leq 2 \cdot \hat{r}$ and as ∞ otherwise, for all $i \leq n$. Note that there can be at most one $i \in [1, \hat{n}]$ so that $|p_i| \leq \hat{r}$ in each tuple (if there were two, they must be at distance $\leq 2 \cdot \hat{r}$ from each other). Note also that the sizes of $\mathrm{N}^{\mathbb{A}|\sigma}_{\hat{r}}(a_i)$ are bounded by a function on k and r since $\mathbf{C}$ contains only k-ranked trees. Thus, it follows that once r and n are fixed,

$$\Xi \overset{\text{def}}{=} \{(\xi_{\mathbb{A}}, \mathrm{N}^{\mathbb{A}|\sigma}_{2r}(\mathit{root}(\mathbb{A}))) \mid \mathbb{A} \in \mathbf{C}\}$$

is finite (assuming, of course, that we do not distinguish isomorphic structures). We will call $(\xi_{\mathbb{A}}, \mathrm{N}^{\mathbb{A}|\sigma}_{2r}(\mathit{root}(\mathbb{A})))$ the *neighborhood description* of $\mathbb{A}$. Note that by Gaifman locality, whether $\mathbb{A} \models \varphi$ holds is determined by $\xi_{\mathbb{A}}$ (and there is an effective procedure to verify this). We now show how to compute Ξ.

The key property to do this is the following lemma.

Lemma 6.5. *Let* a *be the root of* $\mathbb{A} \in \mathbf{C}$ *and let* $a_1, \dots, a_m$*, for* $m \leq k$*, be the children of* a*. Then* $\xi_{\mathbb{A}}$ *is computable from* $\xi_{\mathbb{A}\restriction a_1}, \dots, \xi_{\mathbb{A}\restriction a_m}$ *and* $\mathrm{N}^{\mathbb{A}|\sigma}_{2r}(a)$.

Using the lemma, we can show that the following is the correct procedure to compute Ξ:

1. Let Ξ_0 contain all $(\xi_{\mathbb{A}}, \mathbb{A})$ such that $\mathbb{A} \in \mathbf{C}$ and A is a singleton.
2. Set $\Xi := \Xi_0$.
3. Let $(\xi_1, \mathbb{A}_1), \dots, (\xi_m, \mathbb{A}_m) \in \Xi$ for some $m \in [1, k]$. Thus, there are structures $\mathbb{A}'_i \in \mathbf{C}$ so that $\mathbb{A}_i = \mathrm{N}^{\mathbb{A}'_i|\sigma}_{2r}(\mathit{root}(\mathbb{A}'_i))$ and $\xi_i = \xi_{\mathbb{A}'_i}$ for all $i \in [1, m]$.
4. Let $\mathbb{A}' \in \{\mathrm{N}^{\mathbb{A}|\sigma}_{2r}(\mathit{root}(\mathbb{A})) \mid \mathbb{A} \in \Sigma_{\mathbf{C}}(\mathbb{A}_1, \dots, \mathbb{A}_m)\}$. All such $\mathbb{A}'$ can be computed, since $\mathbf{C}$ is effective inductive and $\mathbb{A}_i = \mathrm{N}^{\mathbb{A}'_i|\sigma}_{2r}(\mathit{root}(\mathbb{A}'_i))$ for each $i \leq m$.
5. Compute ξ' from $\{\xi_i\}_{i \in [1,m]}$ and $\mathbb{A}'$ as in Lemma 6.5.
6. Add $(\xi', \mathbb{A}')$ to Ξ.
7. Repeat steps 3–6 until there is no change to Ξ.

Now the decision procedure is simple: to verify whether φ is satisfiable, it suffices to test whether there is a neighborhood description in Ξ of a structure in which φ holds. Due to Gaifman locality theorem, whether $\mathbb{A} \models \varphi$ depends only on the neighborhood description of $\mathbb{A}$ and φ, and thus the theorem follows. □

6.2 Undecidability results

A natural question is whether the decidability result continues to hold if trees need not be ranked. The answer to this is negative. We start with a simple observation that gives us undecidability for pattern logics with the next-sibling order relation *sib*.

Corollary 6.6. *The pattern logic* $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch,sib}}$ *is undecidable on unranked data trees, even of height* 1.

This is a consequence of the undecidability of FO on data words [15], since the entire tree of height one becomes a radius-one neighborhood of the root, and in the presence of the sibling ordering, it contains full FO on the data word of the root's children.

But even without the sibling order, the pattern logic is undecidable once the restriction to ranked trees is removed.

Theorem 6.7. *The logic* $\mathrm{FO}^{\mathrm{PAT}}_{\mathit{ch}}$ *is undecidable over arbitrary data trees.*

The proof is by encoding of 2-counter Minsky machines and reduction from the undecidability of their emptiness problem.

7. Conclusions

We have shown that pattern logics over structures with auxiliary relations behave quite well and preserve several properties of the underlying logic. We showed strong versions of Hanf and Gaifman locality theorems that do not use auxiliary relations in defining vertices of neighborhoods. We established a general homomorphism preservation theorem that covers classes of structures not covered by previously known homomorphism preservation results. Finally, we proved decidability of the pattern logic for a large class of data trees. Note that the very notion of pattern logic is directly inspired by querying modern data models such as XML and graph databases. In fact our results can be applied to querying XML and graph data with incomplete information and to reasoning about XML with data values.

A few questions remain. The best homomorphism preservation result that we have is for unranked trees of moderate degree. We do not yet know if this restriction can be lifted. For instance, we do not know if the homomorphism preservation theorem holds if the σ-reduct has bounded treewidth, nor do we know what happens on unranked trees with the sibling order/successor. We also would like to explore connections between our results and those on invariance, in particular, we would like to see whether our techniques can be used to provide new results on the behavior of invariant queries. The fact that some common auxiliary relations (*e.g.*, successor) can be defined locally, suggests that this might be a possible route for establishing new results in a notoriously hard area.

Acknowledgements We are very grateful to anonymous reviewers, who pointed out the current simple proofs for the locality theorems in Section 4. Work supported by EPSRC grant J015377.

References

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

[2] S. Abiteboul, L. Segoufin, and V. Vianu. Representing and querying XML with incomplete information. *ACM TODS*, 31(1):208–254, 2006.

[3] S. Amer-Yahia, S. Cho, L. V. S. Lakshmanan, and D. Srivastava. Tree pattern query minimization. *VLDB J.*, 11(4):315–331, 2002.

[4] M. Anderson, D. van Melkebeek, N. Schweikardt, and L. Segoufin. Locality of queries definable in invariant first-order logic with arbitrary built-in predicates. In *ICALP*, pages 368–379, 2011.

[5] A. Atserias. On digraph coloring problems and treewidth duality. In *LICS*, pages 106–115, 2005.

[6] A. Atserias, A. Dawar, and P. Kolaitis. On preservation under homomorphisms and unions of conjunctive queries. *J. ACM*, 53(2):208–237, 2006.

[7] P. Barceló. Querying graph databases. In *PODS*, pages 175–188, 2013.

[8] P. Barceló, L. Libkin, A. Poggi, and C. Sirangelo. XML with incomplete information. *J. ACM*, 58(1):4, 2010.

[9] P. Barceló, L. Libkin, and J. Reutter. Querying regular graph patterns. *J. ACM*, 61(1), 2014.

[10] M. Benedikt and L. Segoufin. Towards a characterization of order-invariant queries over tame graphs. *J. Symb. Log.*, 74(1):168–186, 2009.

[11] M. Bienvenu, B. ten Cate, C. Lutz, and F. Wolter. Ontology-based data access: a study through disjunctive datalog, CSP, and MMSNP. In *PODS*, pages 213–224, 2013.

[12] H. Björklund, W. Martens, and T. Schwentick. Optimizing conjunctive queries over trees using schema information. In *MFCS*, pages 132–143, 2008.

[13] H. Björklund, W. Martens, and T. Schwentick. Conjunctive query containment over trees. *J. Comput. Syst. Sci.*, 77(3):450–472, 2011.

[14] M. Bojańczyk. Automata for data words and data trees. In *RTA*, pages 1–4, 2010.

[15] M. Bojańczyk, C. David, A. Muscholl, T. Schwentick, and L. Segoufin. Two-variable logic on data words. *ACM TOCL*, 12(4):27, 2011.

[16] M. Bojańczyk, A. Muscholl, T. Schwentick, and L. Segoufin. Two-variable logic on data trees and XML reasoning. *J. ACM*, 56(3), 2009.

[17] C. David, A. Gheerbrant, L. Libkin, and W. Martens. Containment of pattern-based queries over data trees. In *ICDT*, pages 201–212, 2013.

[18] A. Dawar. Homomorphism preservation on quasi-wide classes. *J. Comput. Syst. Sci.*, 76(5):324–332, 2010.

[19] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Perspectives in Mathematical Logic. Springer, 1995.

[20] K. Etessami and N. Immerman. Reachability and the power of local ordering. *Theor. Comput. Sci.*, 148(2):261–279, 1995.

[21] R. Fagin and P. Kolaitis. Local transformations and conjunctive-query equivalence. In *PODS*, pages 179–190, 2012.

[22] R. Fagin, L. J. Stockmeyer, and M. Y. Vardi. On monadic NP vs. monadic co-NP. *Inf. & Comput.*, 120(1):78–92, 1995.

[23] W. Fan. Graph pattern matching revised for social network analysis. In *ICDT*, pages 8–21, 2012.

[24] H. Gaifman. On local and non-local properties. In *Proceedings Herbrand Symposium Logic Colloquium, North Holland, 1981*, pages 105–135, 1982.

[25] A. Gheerbrant, L. Libkin, and C. Sirangelo. When is naïve evaluation possible? In *PODS*, pages 75–86, 2013.

[26] E. Grädel, P. Kolaitis, L. Libkin, M. Marx, J. Spencer, M. Vardi, and S. Weinstein. *Finite Model Theory and its Applications*. Springer, 2008.

[27] M. Grohe and T. Schwentick. Locality of order-invariant first-order formulas. *ACM Trans. Comput. Log.*, 1(1):112–130, 2000.

[28] W. P. Hanf. Model-theoretic methods in the study of elementary logic. In *The Theory of Models*, pages 132–145. North Holland, 1965.

[29] T. Imielinski and W. Lipski. Incomplete information in relational databases. *J. ACM*, 31(4):761–791, 1984.

[30] N. Immerman. *Descriptive Complexity*. Springer, 1999.

[31] J. J. Karaganis. On the cube of a graph. *Canadian Math. Bull.*, 11:295–296, 1968.

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

[33] L. Libkin. Incomplete information and certain answers in general data models. In *PODS*, pages 59–70, 2011.

[34] L. Libkin and D. Vrgoč. Regular path queries on graphs with data. In *ICDT*, pages 74–85, 2012.

[35] M. Marx. Conditional XPath. *ACM TODS*, 30(4):929–959, 2005.

[36] G. Miklau and D. Suciu. Containment and equivalence for a fragment of XPath. *J. ACM*, 51(1):2–45, 2004.

[37] H. Niemistö. On locality and uniform reduction. In *LICS*, pages 41–50, 2005.

[38] B. Rossman. Successor-invariant first-order logic on finite structures. *J. Symb. Log.*, 72(2):601–618, 2007.

[39] B. Rossman. Homomorphism preservation theorems. *J. ACM*, 55(3), 2008.

[40] T. Schwentick. On winning Ehrenfeucht games and monadic NP. *Ann. Pure Appl. Logic*, 79(1):61–92, 1996.

[41] L. Segoufin. Static analysis of XML processing with data values. *SIGMOD Record*, 36(1):31–38, 2007.

[42] T. Tan. An automata model for trees with ordered data values. In *LICS*, pages 586–595, 2012.

[43] B. ten Cate and P. Kolaitis. Structural characterizations of schema-mapping languages. In *ICDT*, pages 63–72, 2009.

[44] V. Vianu. A Web odyssey: From Codd to XML. In *PODS*, pages 1–15, 2001.