Logics Capturing Local Properties

LEONID LIBKIN
University of Toronto and Bell Laboratories

Well-known theorems of Hanf and Gaifman establishing locality of first-order definable properties have been used in many applications. These theorems were recently generalized to other logics, which led to new applications in descriptive complexity and database theory. However, a logical characterization of local properties that correspond to Hanf's and Gaifman's theorems is still lacking. Such a characterization only exists for structures of bounded valence. In this paper, we give logical characterizations of local properties behind Hanf's and Gaifman's theorems. We first deal with an infinitary logic with counting terms and quantifiers that is known to capture Hanf-locality on structures of bounded valence. We show that testing isomorphism of neighborhoods can be added to it without violating Hanf-locality, while increasing its expressive power. We then show that adding local second-order quantification to it captures precisely all Hanf-local properties. To capture Gaifman-locality, one must also add a (potentially infinite) `case` statement. We further show that the hierarchy based on the number of variants in the `case` statement is strict.

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

General Terms: Languages, Theory

Additional Key Words and Phrases: Locality, logic, counting

1. INTRODUCTION

It is well known that first-order logic (FO) only expresses local properties. Two best known formal results stating locality of FO are Hanf's and Gaifman's theorems [Hanf 1965; Gaifman 1982]. They both found numerous applications in computer science, due to the fact that they are among relatively few results in first-order model theory that apply to both finite and infinite structures. Gaifman's theorem itself works for both finite and infinite structures, while for Hanf's theorem an extension to finite structures was formulated by Fagin, Stockmeyer, and Vardi [1995].

More recently, the statements underlying Hanf's and Gaifman's theorems have been abstracted from the statements of the theorems, and used in their own right. In essence, Hanf's theorem states that two structures cannot be distinguished by sentences of quantifier rank k whenever they realize the same multiset of d-

Part of this work done while visiting INRIA. Preliminary version of this paper appeared in the 17th Symposium on Theoretical Aspects of Computer Science (STACS'2000). Address: Department of Computer Science, 6 King's College Road, Room PT370, University of Toronto, Ontario M5S 3H5, Canada, email: libkin@cs.toronto.edu.

neighborhoods of points; here d depends only on k. Gaifman's theorem states that in a given structure, two tuples cannot be distinguished by formulae of quantifier rank k whenever d-neighborhoods of these tuples are isomorphic; again d is determined by k.

It was shown that Hanf's theorem is strictly stronger than Gaifman's, and that both apply to a variety of logics that extend FO with counting mechanisms and limited infinitary connectives [Grohe and Schwentick 2000; Hella et al. 1999a; Hella et al. 1999b; Libkin 2000; Nurmonen 1996]. These results found applications in descriptive complexity and database theory. Since the complexity class TC^0 (with the appropriate notion of uniformity) can be captured by FO with counting quantifiers [Barrington et al. 1990], locality can be used to prove lower bounds for logics coming very close to capturing TC^0 [Etessami 1995; Libkin and Wong 1998]. In database theory, logics with counting mechanisms model aggregate functions commonly found in commercial query languages. Thus, locality was used to prove expressivity bounds for query languages with aggregation [Dong et al. 2000; Hella et al. 1999b]. For applications to automata, see [Schwentick and Barthelmann 1998].

The above-mentioned papers considered a sequence of more and more powerful logics, each of which was proved to be local, starting with FO with counting quantifiers, and ending with a logic that permits arbitrary predicates on natural numbers, a limited form of infinitary connectives [Libkin 2000] and even aggregate functions [Hella et al. 1999b]. However, it was not clear how much one can add to these logics and still preserve its locality. Our goal, therefore, is to give a precise characterization of local logics.

Note that the abstract notions of locality were previously characterized on finite structures of *bounded valence* (e.g., for graphs of fixed maximum degree). The characterization for Hanf-locality uses a logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ introduced in [Libkin 2000]. This logic subsumes a number of counting extensions of FO (such as FO with counting quantifiers [Immerman and Lander 1990], FO with unary generalized quantifiers [Hella 1996; Kolaitis and Väänänen 1995], FO with unary counters [Benedikt and Keisler. 1997]) and is quite easy to deal with. A result in [Hella et al. 1999a] states that Hanf-local properties on structures of bounded valence are precisely those definable in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$.

The question naturally arises whether this continues to hold for arbitrary finite structures. We show in this paper that this is not the case. We do so by first finding a simple direct proof of Hanf-locality of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$, and then using it to show that adding new atomic formulae testing isomorphism of neighborhoods of a fixed radius does not violate Hanf-locality, while strictly increasing the expressive power. We next define a logic that captures precisely the Hanf-local properties. It is obtained by adding *local second-order* quantification to $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$. That is, second-order quantifiers bind predicates that are only allowed to range over fixed radius neighborhoods of free first-order variables. We will also show that this amounts to adding arbitrarily powerful computations to $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ as long as they are bound to some neighborhoods.

For Gaifman-locality, a characterization theorem in [Hella et al. 1999a] stated that it is equivalent, over structures of bounded valence, to first-order definition by cases. That is, there are $m > 0$ classes of structures and m FO formulae φ_i such that, over the ith class, the given property is described by φ_i. Again, this falls short of a general characterization. We show that over the class of all finite structures (no restriction on valence), Gaifman-locality is equivalent to definition by cases, where the number of classes can be infinite. Furthermore, the hierarchy given by the number of those classes (that is, the number of cases) is strict.

Organization. Section 2 introduces notations and notions of locality. Section 3 gives a new simple proof of Hanf-locality of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ which is then used to show that adding tests for neighborhood isomorphism preserves locality. Section 4 characterizes Hanf-local properties as those definable in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ with local second-order quantification. Section 5 characterizes Gaifman-local properties as those definable by (finite or infinite) case statements, and shows the strictness of the hierarchy.

2. NOTATION

Finite Structures and Neighborhoods. All structures are assumed to be *finite*. A relational signature σ is a set of relation symbols $\{R_1, ..., R_l\}$, with associated arities $p_i > 0$. A σ-structure is $\mathcal{A} = \langle A, R_1^{\mathcal{A}}, \ldots, R_l^{\mathcal{A}}\rangle$, where A is a finite set, and $R_i^{\mathcal{A}} \subseteq A^{p_i}$ interprets R_i. The class of finite σ-structures is denoted by $\mathrm{STRUCT}[\sigma]$. When there is no confusion, we write R_i in place of $R_i^{\mathcal{A}}$. Isomorphism is denoted by $\cong$. The carrier of a structure $\mathcal{A}$ is always denoted by A, and the carrier of $\mathcal{B}$ is denoted by B.

Given a structure $\mathcal{A}$, its *Gaifman graph* $\mathcal{G}(\mathcal{A})$ is defined as $\langle A, E\rangle$ where (a, b) is in E iff there is a tuple $\vec{c} \in R_i^{\mathcal{A}}$ for some i such that both a and b are in $\vec{c}$. The distance $d(a, b)$ is defined as the length of the shortest path from a to b in $\mathcal{G}(\mathcal{A})$; we assume $d(a, a) = 0$. If $\vec{a} = (a_1, \ldots, a_n)$ and $\vec{b} = (b_1, \ldots, b_m)$, then $d(\vec{a}, \vec{b}) = \min_{ij} d(a_i, b_j)$. Given $\vec{a}$ over A, its *r-sphere* $S_r^{\mathcal{A}}(\vec{a})$ is $\{b \in A \mid d(\vec{a}, b) \leq r\}$. Its *$r$-neighborhood* $N_r^{\mathcal{A}}(\vec{a})$ is defined as a structure in the signature that extends σ with n new constant symbols:

$$\langle S_r^{\mathcal{A}}(\vec{a}), R_1^{\mathcal{A}} \cap S_r^{\mathcal{A}}(\vec{a})^{p_1}, \ldots, R_l^{\mathcal{A}} \cap S_r^{\mathcal{A}}(\vec{a})^{p_l}, a_1, \ldots, a_n\rangle$$

That is, the carrier of $N_r^{\mathcal{A}}(\vec{a})$ is $S_r^{\mathcal{A}}(\vec{a})$, the interpretation of the σ-relations is inherited from $\mathcal{A}$, and the n extra constants are the elements of $\vec{a}$. If $\mathcal{A}$ is understood, we write $S_r(\vec{a})$ and $N_r(\vec{a})$.

If $\mathcal{A}, \mathcal{B} \in \mathrm{STRUCT}[\sigma]$, and there is an isomorphism $N_r^{\mathcal{A}}(\vec{a}) \to N_r^{\mathcal{B}}(\vec{b})$ (that sends $\vec{a}$ to $\vec{b}$), we write $\vec{a} \approx_r^{\mathcal{A},\mathcal{B}} \vec{b}$. If $\mathcal{A} = \mathcal{B}$, we write $\vec{a} \approx_r^{\mathcal{A}} \vec{b}$.

Given a tuple $\vec{a} = (a_1, \ldots, a_n)$, we write $\vec{a}c$ for the tuple $(a_1, \ldots, a_n, c)$.

The quantifier rank of a formula is denoted by $\mathsf{qr}(\cdot)$.

Hanf's and Gaifman's Theorems. An *m-ary query* on σ-structures, Q, is a mapping that associates to each $\mathcal{A} \in \mathrm{STRUCT}[\sigma]$ a structure $\langle A, S\rangle$, where $S \subseteq A^m$. We always assume that queries are invariant under isomorphisms. We write $\vec{a} \in Q(\mathcal{A})$ if $\vec{a} \in S$, where $\langle A, S\rangle = Q(\mathcal{A})$. A query Q is definable in a logic $\mathcal{L}$ if there exists an $\mathcal{L}$ formula $\varphi(x_1, \ldots, x_m)$ such that $Q(\mathcal{A}) = \langle A, \{\vec{a} \mid \mathcal{A} \models \varphi(\vec{a})\}\rangle$. If $m = 0$, then Q is naturally associated with a subclass of $\mathrm{STRUCT}[\sigma]$, and definability means definability by a sentence of $\mathcal{L}$.

Definition 2.1 *(Gaifman-Locality).* (See [Dong et al. 2000; Hella et al. 1999a]). An m-ary query Q, $m \geq 1$, is called *Gaifman-local* if there exists a number $r \geq 0$

such that, for any structure $\mathcal{A}$ and any $\vec{a}, \vec{b} \in A^m$

$$\vec{a} \approx_r^{\mathcal{A}} \vec{b} \quad \textit{implies} \quad \vec{a} \in Q(\mathcal{A}) \text{ iff } \vec{b} \in Q(\mathcal{A}).$$

The minimum such r is called the *locality rank* of Q, and is denoted by $\mathsf{lr}(Q)$.

THEOREM 2.2 (GAIFMAN [1982]). *Every FO formula $\varphi(x_1, \ldots, x_m)$ defines a Gaifman-local query Q with $\mathsf{lr}(Q) \leq (7^{\mathsf{qr}(\varphi)} - 1)/2$.*

The statement of Gaifman's theorem actually provides more information about FO-definable properties; it also states that every first-order definable property can be expressed in terms of types of neighborhoods realized in a given structure. An abstract formulation of this property was introduced in [Hella et al. 1999a] under the name of strong Gaifman-locality, and was shown to be equivalent to first-order definability over finite structures of bounded degree. However, it is the above statement that is used in most applications for proving expressivity bounds, and it also extends beyond FO. Note also that better bounds of the order $O(2^{\mathsf{qr}(\varphi)})$ are known for $\mathsf{lr}(Q)$; see [Libkin 2000].

For $\mathcal{A}, \mathcal{B} \in \mathrm{STRUCT}[\sigma]$, we write $\mathcal{A} \leftrightarrows_d \mathcal{B}$ if the multisets of isomorphism types of d-neighborhoods of points are the same in $\mathcal{A}$ and $\mathcal{B}$. That is, $\mathcal{A} \leftrightarrows_d \mathcal{B}$ if there exists a bijection $f : A \to B$ such that $N_d^{\mathcal{A}}(a) \cong N_d^{\mathcal{B}}(f(a))$ for every $a \in A$. We also write $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$ if there is a bijection $f : A \to B$ such that $N_d^{\mathcal{A}}(\vec{a}c) \cong N_d^{\mathcal{B}}(\vec{b}f(c))$ for every $c \in A$.

Definition 2.3 *(Hanf-Locality).* (See [Hanf 1965; Fagin et al. 1995; Hella et al. 1999a]). An m-ary query Q, $m \geq 0$, is called *Hanf-local* if there exist a number $d \geq 0$ such that for any two structures $\mathcal{A}, \mathcal{B}$ and any $\vec{a} \in A^m, \vec{b} \in B^m$,

$$(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b}) \quad \textit{implies} \quad \vec{a} \in Q(\mathcal{A}) \text{ iff } \vec{b} \in Q(\mathcal{B}).$$

The minimum d for which this holds is called *Hanf locality rank* of Q, and is denoted by $\mathsf{hlr}(Q)$.

For a Boolean query Q ($m = 0$) this means that Q cannot distinguish two structures $\mathcal{A}$ and $\mathcal{B}$ whenever $\mathcal{A} \leftrightarrows_d \mathcal{B}$.

THEOREM 2.4 (HANF [1965], FAGIN-STOCKMEYER-VARDI [1995]). *Every FO sentence φ defines a Hanf-local Boolean query Q with $\mathsf{hlr}(Q) \leq 3^{\mathsf{qr}(\Phi)}$.* $\square$

An extension to open formulae, although easily derivable from the proof of [Fagin et al. 1995], was probably first explicitly stated in [Hella et al. 1999a]: every FO formula $\varphi(\vec{x})$ defines a Hanf-local query. Better bounds on $\mathsf{hlr}(Q)$ of the order $O(2^{\mathsf{qr}(\varphi)})$ are also known for Hanf-locality [Immerman 1999; Libkin 2000].

We shall use the following result that connects the binary relations $\leftrightarrows$ and $\approx$.

LEMMA 2.5 (SEE [HELLA ET AL. 1999A]). *(a) Let $\mathcal{A} \leftrightarrows_d \mathcal{B}$ and $\vec{a} \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}$. Then $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$.*

(b) Let $(\mathcal{A}, \vec{a}) \leftrightarrows_{3d+1} (\mathcal{B}, \vec{b})$. Then there exists a bijection $f : A \to B$ such that $(\mathcal{A}, \vec{a}c) \leftrightarrows_d (\mathcal{B}, \vec{b}f(c))$ for every $c \in A$.

Note that Lemma 2.5, part (b) is in fact an easy corollary of Lemma 2.5, (a): If $(\mathcal{A}, \vec{a}) \leftrightarrows_{3d+1} (\mathcal{B}, \vec{b})$, then there is a bijection $f : A \to B$ such that $\vec{a}c \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}f(c)$; since $\mathcal{A} \leftrightarrows_{3d+1} \mathcal{B}$ and thus $\mathcal{A} \leftrightarrows_d \mathcal{B}$, this implies $(\mathcal{A}, \vec{a}c) \leftrightarrows_d (\mathcal{B}, \vec{b}f(c))$.

Another easy corollary of Lemma 2.5, (a), is that every Hanf-local m-ary query Q, $m \geq 1$, is Gaifman-local [Hella et al. 1999a]. Indeed, let $d = 3\mathsf{hlr}(Q) + 1$, and let $\vec{a} \approx_d^{\mathcal{A}} \vec{b}$. Since $\mathcal{A} \leftrightarrows_{\mathsf{hlr}(Q)} \mathcal{A}$, we obtain $(\mathcal{A}, \vec{a}) \leftrightarrows_{\mathsf{hlr}(Q)} (\mathcal{A}, \vec{b})$ and thus $\vec{a} \in Q(\mathcal{A})$ iff $\vec{b} \in Q(\mathcal{A})$, by Hanf-locality.

Logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$. The logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ subsumes a number of counting extensions of FO, such as FO with counting quantifiers [Etessami 1995; Immerman and Lander 1990], unary quantifiers [Hella 1996], and unary counters [Benedikt and Keisler. 1997]. (When we speak of counting extensions of FO, we mean extensions that only add a counting mechanism, as opposed to those – extensively studied in the literature, see [Cai et al. 1992; Otto 1997] – that add both counting and fixpoint.) It is a two-sorted logic, with one sort being the universe of a finite structure, and the other sort being $\mathbb{N}$, and it uses counting terms that produce constants of the second sort, similarly to the logics studied in [Grädel and Gurevich 1998]. The formal definition is as follows.

We denote the infinitary logic by $\mathcal{L}_{\infty\omega}$; it extends FO by allowing infinite conjunctions $\bigwedge$ and disjunctions $\bigvee$. Then $\mathcal{L}_{\infty\omega}(\mathbf{C})$ is a two-sorted logic that extends $\mathcal{L}_{\infty\omega}$. Its structures are of the form $(\mathcal{A}, \mathbb{N})$, where $\mathcal{A}$ is a finite relational structure, and $\mathbb{N}$ is a copy of natural numbers. We shall use $\vec{x}, \vec{y}$, etc., for variables ranging over the first (nonnumerical) sort, and $\vec{\imath}, \vec{\jmath}$, etc., for variables ranging over the second (numerical) sort. Assume that every constant $n \in \mathbb{N}$ is a second-sort term. To $\mathcal{L}_{\infty\omega}$, add *counting quantifiers* $\exists ix$ for every $i \in \mathbb{N}$, and *counting terms:*

—If φ is a formula and $\vec{x}$ is a tuple of free first-sort variables in φ, then $\#\vec{x}.\varphi$ is a term of the second sort, and its free variables are those in φ except $\vec{x}$. Its interpretation is the number of $\vec{a}$ over the finite first-sort universe that satisfy φ. That is, given a structure $\mathcal{A}$, a formula $\varphi(\vec{x}, \vec{y}; \vec{\jmath})$, $\vec{b} \subseteq A$, and $\vec{\jmath}_0 \subset \mathbb{N}$, the value of the term $\#\vec{x}.\varphi(\vec{x}, \vec{b}; \vec{\jmath}_0)$ is the cardinality of the (finite) set $\{\vec{a} \subseteq A \mid \mathcal{A} \models \varphi(\vec{a}, \vec{b}; \vec{\jmath}_0)\}$. For example, the interpretation of $\#x.E(x, y)$ is the in-degree of node y in a graph with the edge-relation E.

—The interpretation of a counting quantifier $\exists ix\varphi$ is $\#x.\varphi \geq i$. Note that this quantifier binds x, but i remains free.

As this logic is too powerful (it expresses every property of finite structures), we restrict it by means of the *rank* of formulae and terms, denoted by rk. It is defined as quantifier rank, but without taking into account quantification over $\mathbb{N}$. That is:

—The rank of a variable or a constant is 0.

—The rank of an atomic formula is the maximum rank of a term in it.

—$\mathsf{rk}(\bigvee_i \varphi_i) = \mathsf{rk}(\bigwedge_i \varphi_i) = \sup_i \mathsf{rk}(\varphi_i)$.

—$\mathsf{rk}(\neg\varphi) = \mathsf{rk}(\varphi)$.

—$\mathsf{rk}(\exists x\varphi) = \mathsf{rk}(\exists ix\varphi) = \mathsf{rk}(\varphi) + 1$.

—$\mathsf{rk}(\exists n\varphi) = \mathsf{rk}(\varphi)$, where n ranges over $\mathbb{N}$.

—$\mathsf{rk}(\#\vec{x}.\psi) = \mathsf{rk}(\psi) + |\vec{x}|$.

Definition 2.6. (See [Libkin 2000].) The logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ is defined to be the restriction of $\mathcal{L}_{\infty\omega}(\mathbf{C})$ to terms and formulae of finite rank.

It is known [Libkin 2000] that $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ is closed under finitary Boolean connectives and all quantification, and that every predicate on $\mathbb{N} \times \ldots \times \mathbb{N}$ is definable by a $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ formula of rank 0. Thus, we assume that $+, *, -, \leq$, and in fact *every* predicate on $\mathbb{N}$ is available. Furthermore, counting terms can be eliminated in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ without increasing the rank; that is, counting quantifiers suffice. In fact, there exists an alternative presentation of this logic, which is one-sorted, and uses arbitrary unary generalized quantifiers [Hella 1996; Hella et al. 1999a]; however, expressing counting properties with unary quantifiers is often quite awkward, and thus we chose to use a two-sorted version with counting terms here.

FACT 2.7. (See [Hella et al. 1999b; Libkin 2000].) *Queries expressed by $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ formulae without free variables of the second-sort are Hanf-local and Gaifman-local.*

Gaifman-locality of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ was proved by a simple direct argument in [Libkin 2000]; Hanf-locality was shown in [Hella et al. 1999b] using *bijective Ehrenfeuct-Fraïssé games* of [Hella 1996]. The game is played by two players, called the spoiler and the duplicator, on two structures $\mathcal{A}, \mathcal{B} \in \text{STRUCT}[\sigma]$. For the n-round game, in each round $i = 1, \ldots, n$, the duplicator selects a bijection $f_i : A \to B$, and the spoiler selects a point $a_i \in A$ (if $card(A) \neq card(B)$, then the spoiler wins). The duplicator wins after n rounds if the relation $\{(a_i, f_i(a_i)) \mid 1 \leq i \leq n\}$ is a partial isomorphism $\mathcal{A} \to \mathcal{B}$; otherwise the spoiler wins. If the duplicator has a winning strategy in the n-move bijective game on $\mathcal{A}$ and $\mathcal{B}$, we write $\mathcal{A} \equiv^{bij}_n \mathcal{B}$. It was shown in [Hella et al. 1999b] (building upon [Hella 1996]) that bijective games characterize elementary equivalence in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$: $\mathcal{A}$ and $\mathcal{B}$ agree on $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ sentences of rank up to n iff $\mathcal{A} \equiv^{bij}_n \mathcal{B}$.

Structures of Bounded Valence (Degree). We use the notation $\text{STRUCT}_k[\sigma]$ for the set of structures $\mathcal{A} \in \text{STRUCT}[\sigma]$ such that in the Gaifman graph $\mathcal{G}(\mathcal{A})$, every node has degree at most k.

FACT 2.8. (See [Hella et al. 1999a].) *For any fixed k, a query Q on $\text{STRUCT}_k[\sigma]$ is Hanf-local iff it is expressed by a formula of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ (without free second-sort variables).*

An m-ary query Q on a class $\mathcal{C} \subseteq \text{STRUCT}[\sigma]$ is given by a *first-order definition by cases* if there exists a number p, a partition $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \ldots \cup \mathcal{C}_p$ and first-order formulae $\alpha_1(x_1, \ldots, x_m), \ldots, \alpha_p(x_1, \ldots, x_m)$ in the language σ such that on all structures $\mathcal{A} \in \mathcal{C}_i$, Q is definable by α_i. That is, for all $1 \leq i \leq p$ and $\mathcal{A} \in \mathcal{C}_i$, $\vec{a} \in Q(\mathcal{A})$ iff $\mathcal{A} \models \alpha_i(\vec{a})$. Note that $\mathcal{C}_i$'s are not required to be first-order-definable.

FACT 2.9. (See [Hella et al. 1999a].) *For any fixed k, a query Q on $\text{STRUCT}_k[\sigma]$ is Gaifman-local iff it is given by a first-order definition by cases.*

3. ISOMORPHISM OF NEIGHBORHOODS AND $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$

We start with a slightly modified definition of locality that makes it convenient to work with two-sorted logics, like $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$. We say that such a logic expresses Hanf-local (or Gaifman-local) queries if for every formula $\varphi(\vec{x}, \vec{\imath})$ there exists a number d such that for every $\vec{\imath}_0 \subset \mathbb{N}$, the formula $\varphi_{\vec{\imath}_0}(\vec{x}) = \varphi(\vec{x}, \vec{\imath}_0)$ (without free second-sort variables) expresses a query Q with $\mathsf{hlr}(Q) \leq d$ ($\mathsf{lr}(Q) \leq d$, respectively).

Consider a set θ of relation symbols, disjoint from σ, and define $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \theta$ by allowing for each k-ary $U \in \theta$ and a k-tuple $\vec{x}$ of variables of the first sort, $U(\vec{x})$ to be a new atomic formula. The rank of this formula is 0. An interpretation of predicates in θ is said to be Hanf-local if there exists a number d such that each predicate in θ defines a Hanf-local query Q with $\mathsf{hlr}(Q) \le d$.

THEOREM 3.1. *If the interpretation of predicates in θ is Hanf-local, then every query definable in $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \theta$ is Hanf-local.*

PROOF. Let d witness Hanf-locality of θ. We shall show that every $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \theta$ formula of rank m defines a Hanf-local query Q with $\mathsf{hlr}(Q) \le 3^m \cdot d + \frac{3^m - 1}{2}$ (for all instantiations of free variables of the second sort). That is, for a sequence defined by $d_0 = d, d_1 = 3d_0 + 1, \ldots, d_{i+1} = 3d_i + 1, \ldots$, we have $\mathsf{hlr}(Q) \le d_m$.

The proof of this is by induction on a formula. The atomic case follows from the assumption that θ is Hanf-local (note that atomic σ-formulae define queries of Hanf locality rank 0). The cases of Boolean or infinitary connectives are simple: for example, if formulae $\varphi_j(\vec{x}, \vec{\imath})$ define queries of Hanf locality rank at most r for all instantiations $\vec{\imath}_0$ for $\vec{\imath}$, then the same is true for $\varphi \equiv \bigvee_j \varphi_j$. Indeed, if $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$, then $\mathcal{A} \models \varphi_j(\vec{a}, \vec{\imath}_0)$ iff $\mathcal{B} \models \varphi_j(\vec{b}, \vec{\imath}_0)$, and thus the same is true for φ. The cases of negation and quantification over the numerical sort clearly do not change the value of hlr, since hlr is independent of $\vec{\imath}_0$.

It thus remains to consider the case of $\psi(\vec{x}, \vec{\imath}) \equiv \exists i y (\varphi(y, \vec{x}, \vec{\imath}))$ (as counting terms can be eliminated without increasing the rank [Libkin 2000]) and to show that if φ defines a query of Hanf locality rank r for every $\vec{\imath}_0$, then ψ defines a query Q with $\mathsf{hlr}(Q) \le 3r + 1$. We then fix $\vec{\imath}_0$ and assume $(\mathcal{A}, \vec{a}) \leftrightarrows_{3r+1} (\mathcal{B}, \vec{b})$. By Lemma 2.5, b), there exists a bijection $f : A \to B$ such that $(\mathcal{A}, \vec{a}c) \leftrightarrows_r (\mathcal{B}, \vec{b}f(c))$ for all $c \in A$. Thus, $\mathcal{A} \models \varphi(c, \vec{a}, \vec{\imath})$ iff $\mathcal{B} \models \varphi(f(c), \vec{b}, \vec{\imath})$, due to Hanf-locality of φ, and hence $\mathcal{A} \models \psi(\vec{a}, \vec{\imath})$ iff $\mathcal{B} \models \psi(\vec{b}, \vec{\imath})$, as the number of elements satisfying $\varphi(\cdot, \vec{a}, \vec{\imath})$ and $\varphi(\cdot, \vec{b}, \vec{\imath})$ is the same. This completes the proof. $\square$

We now consider the following example. For each d, k, define a $2k$-ary predicate $I^k_d(x_1, \ldots, x_k, y_1, \ldots, y_k)$ to be interpreted as follows: $\mathcal{A} \models I^k_d(\vec{a}, \vec{b})$ iff $N^{\mathcal{A}}_d(\vec{a}) \cong N^{\mathcal{A}}_d(\vec{b})$. Clearly, $(\mathcal{A}, \vec{a}_1\vec{a}_2) \leftrightarrows_d (\mathcal{B}, \vec{b}_1\vec{b}_2)$ implies $N^{\mathcal{A}}_d(\vec{a}_1\vec{a}_2) \cong N^{\mathcal{B}}_d(\vec{b}_1\vec{b}_2)$, and thus $\vec{a}_1 \approx^{\mathcal{A}}_d \vec{a}_2$ iff $\vec{b}_1 \approx^{\mathcal{B}}_d \vec{b}_2$. This shows Hanf-locality of I^k_d and gives us

COROLLARY 3.2. *For any fixed d, $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I^k_d \mid k > 0\}$ only expresses Hanf-local properties.*

We next show that this gives us an increase in expressive power. The result below is proved using bijective games.

PROPOSITION 3.3. *For any $d, k > 0$, $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + I^k_d$ is strictly more expressive than $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$.*

PROOF. It suffices to show this proposition for the case of $d = k = 1$. Consider the signature of one binary relation E and a formula $\varphi(x) \equiv E(x,x) \wedge \exists y I^1_1(x, y)$. Assume to the contrary that this is definable by a $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ formula ψ of rank m. Let $r = 3^{m+1}$. We now construct a graph G with the set of nodes $V = \{a, b, c\} \cup \{a_i, b_i, c_i \mid 1 \le i \le 2r\}$. First, we have loops $(a, a), (b, b), (c, c)$ and edges $(a, a_i), (b, b_i), (c, c_i)$ for each $i \le 2r$. Furthermore, on the a_i's we have two cycles of length r: $(a_1,a_2),\ldots,(a_{r-1},a_r),(a_r,a_1)$ and $(a_{r+1},a_{r+2}),\ldots,(a_{2r-1},a_{2r}),(a_{2r},a_{r+1})$, and likewise for the b_i. On the nodes c_i's, we have one cycle of length $2r$: $(c_1,c_2),\ldots,(c_{2r-1},c_{2r}),(c_{2r},c_1)$. There are no other edges.

Note that the output of φ on G is $\{a,b\}$. We next show that $(G,a)\equiv_m^{bij}(G,c)$ which would imply that $G\models\psi(a)$ iff $G\models\psi(b)$, contradicting definability of φ in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$. Let G_a be the subgraph of G whose nodes are the a_j's and let G_c be the subgraph whose nodes are the c_js. Since $G_a \leftrightarrows_{3^m} G_c$, it follows from [Hella et al. 1999a; Nurmonen 1996] that $G_a\equiv_m^{bij}G_c$. Then the duplicator wins in the m-round bijective game on (G,a) and (G,c) as follows. For the ith round, the duplicator looks at the points played so far on G_a and G_c and, assuming he is playing round $i+1$ in the bijective game on G_a and G_c, constructs a bijection $f_0: G_a\to G_c$. Then this bijection is extended to the bijection f from (G,a) to (G,c) as follows. First, $f(a)=c, f(c)=a, f(b)=b$. Secondly, $f(b_j)=b_j$ for all j. Finally, $f(a_j)=f_0(a_j)$ and $f(c_j)=f_0^{-1}(a_j)$. It follows immediately from the construction and from $G_a\equiv_m^{bij}G_c$ that with this strategy, the duplicator maintains partial isomorphism. □

COROLLARY 3.4. *The logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ fails to capture Hanf-local properties over arbitrary finite structures.*

Note that we only used I_d^ks as atomic formulae. A natural extension would be to use them as generalized quantifiers. In this case we extend the definition of the logic by a rule that if $\varphi_1(\vec{v}_1,\vec{z}),\ldots,\varphi_l(\vec{v}_l,\vec{z})$ are formulae with $\vec{v}_i$ being an m_i-tuple of first-sort variables, then $\psi(\vec{x},\vec{y},\vec{z})\equiv\mathbf{I}_d^k[m_1,\ldots,m_l](\vec{v}_1,\ldots,\vec{v}_l)(\varphi_1(\vec{v}_1,\vec{z}),\ldots,\varphi_l(\vec{v}_l,\vec{z}))$ is a formula with $\vec{x}$ and $\vec{y}$ being k-tuples of fresh free variables of the first sort. The semantics is that for each $\mathcal{A}$ and $\vec{c}$, one defines a new structure on A in which the ith predicate of arity m_i is interpreted as $\{\vec{u}\in A^{m_i}\mid\mathcal{A}\models\varphi_i(\vec{u},\vec{c})\}$. Then $\mathcal{A}\models\psi(\vec{a},\vec{b},\vec{c})$ if in this structure the d-neighborhoods of $\vec{a}$ and $\vec{b}$ are isomorphic. However, this generalization does not preserve locality.

PROPOSITION 3.5. *Adding $\mathbf{I}_d^k[m_1,\ldots,m_l]$ to $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ violates Hanf-locality. In fact, with addition of $\mathbf{I}_1^1[2]$ to FO one can define properties that are neither Hanf-local nor Gaifman-local.*

PROOF. Consider a signature (E,C_1,C_2) where E is binary and C_1, C_2 unary (that is, we deal with 2-colored graphs). Let $\varphi(u,v)$ be $(E(u,v)\wedge C_1(u))\vee(C_2(u)\wedge C_1(v))$. We now form $\psi(x,y)\equiv\mathbf{I}_1^1[2](u,v)(\varphi(u,v))$ testing if 1-neighborhoods of u and v are isomorphic in the output of φ. Assume that this defines a Gaifman-local query Q with $\mathsf{lr}(Q)\le r, r>0$. Take $m=4r$ and construct a 2-colored graph G as follows. The set of nodes is $\{a_i,b_i,c_i,e_i\mid 1\le i\le m\}$. The edges are $(a_i,a_{i+1}),(b_i,b_{i+1}),(c_i,c_{i+1}),(e_i,e_{i+1})$ for $1\le i<m$ as well as $(a_i,b_i),(e_i,c_i)$ for all i. The interpretation of C_1 is $\{a_i,e_i\mid 1\le i\le m\}$, and the interpretation of C_2 is $\{b_i,c_i\mid 1\le i\le m\}$.

For each b_i, its 1-neighborhood in the output of φ consists of $\{b_i\}\cup\{a_j,e_j\mid 1\le j\le m\}$, with all the E-edges between the a_j's and e_j, as well as (a_i,b_i) and (e_i,b_i). Likewise, the 1-neighborhood of c_k in the output of φ consists of $\{c_k\}\cup\{a_j,e_j\mid 1\le j\le m\}$, with all the E-edges between the a_j's and e_j, and the edges $(a_k, c_k), (e_k, c_k)$. Thus, those neighborhoods are isomorphic iff $i = k$. However, our choice of m guarantees that there is $i < m$ such that $(b_i, c_i) \approx_r^G (b_{i+1}, c_i)$ which would imply $\psi(b_i, c_i)$ iff $\psi(b_{i+1}, c_i)$, by the locality of ψ. However, we have $\psi(b_i, c_i)$ and $\neg\psi(b_{i+1}, c_i)$. This contradiction shows that ψ is not Gaifman-local; consequently, it is not Hanf-local either. □

4. CHARACTERIZING HANF-LOCAL PROPERTIES

We have seen that the logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ fails to capture Hanf-local properties over arbitrary finite structures. To fill the gap between $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ and Hanf-locality, we introduce the notion of *local second-order quantification*. The idea is similar to local first-order quantification which restricts quantified variables to fixed radius neighborhoods of free variables. This kind of quantification was used in Gaifman's locality theorem [Gaifman 1982] as well as in translations of various modal logics into fragments of FO [van Benthem 1985; Grädel 1999].

Definition 4.1. Fix $r \geq 0$ and a relational signature σ. Suppose that we have, for every arity $k > 0$, a countably infinite set of k-ary relational symbols T_k^i, $i \in \mathbb{N}$, disjoint from σ. Define a set of formulae $\mathcal{F}$ by starting with $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ atomic formulae involving symbols from σ as well as T_k^i's, and closing under the formation rules of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ and the following rule: if $\varphi(\vec{x}, \vec{\imath})$ is a formula, $\vec{y}$ is a subtuple of $\vec{x}$ and $d \leq r$, then

$$\psi_1(\vec{x}, \vec{\imath}) \;\equiv\; \exists T_k^i \sqsubseteq S_d(\vec{y})\; \varphi(\vec{x}, \vec{\imath}) \quad \text{and} \quad \psi_2(\vec{x}, \vec{\imath}) \;\equiv\; \forall T_k^i \sqsubseteq S_d(\vec{y})\; \varphi(\vec{x}, \vec{\imath})$$

are formulae of rank $\mathsf{rk}(\varphi) + 1$. We say that the symbol T_k^i is bound in these formulae.

We then define $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ over STRUCT$[\sigma]$ as the set of all formulae in $\mathcal{F}$ of finite rank in which all occurrences of the symbols T_k^i's are bound. The logic $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ (*local second-order with counting*) is defined as $\bigcup_{r \geq 0} \mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$.

The semantics of the new construct is as follows. Given a σ-structure $\mathcal{A}$ and an interpretation $\mathcal{T}$ for all the symbols T_k^i's occurring freely in ψ_1, we have $(\mathcal{A}, \mathcal{T}) \models \psi_1(\vec{a}, \vec{\imath})$ iff there exists a set $T \subseteq S_d(\vec{b})^k$, where $\vec{b}$ is the subtuple of $\vec{a}$ corresponding to $\vec{y}$, such that $(\mathcal{A}, \mathcal{T}, T) \models \varphi(\vec{a}, \vec{\imath})$. For ψ_2, one replaces "exists" by "for all." □

For example, the formula

$$\exists x \exists T \sqsubseteq S_r(x) \exists T' \sqsubseteq S_r(x) \left(\begin{array}{l} \quad\;\; \forall y \in S_r(x)\ (T(y) \wedge \neg T'(y)) \vee (\neg T(y) \wedge T'(y)) \\ \wedge\ \forall z, v\ (T(z) \wedge E(z,v) \to \\ \qquad\qquad\qquad T'(v)) \wedge (T'(z) \wedge E(z,v) \to T(v)) \end{array} \right)$$

tests if there is a 2-colorable r-neighborhood of a node in a graph. Note that local first-order quantification $\forall y \in S_r(x)$ is definable in FO for every fixed r.

Our main result can now be stated as follows.

THEOREM 4.2. *An m-ary query Q, $m \geq 0$, is Hanf-local iff it is definable by a formula of $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ (without free second-sort variables).*

PROOF. We first show that queries definable in $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ are Hanf-local. As the first observation, we note that counting terms can be eliminated from $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ without increasing the rank of a formula; in fact, the proof of this result for $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$

from [Libkin 2000] applies verbatim. Thus, we shall always assume in this direction of the proof that we deal with formulae without counting terms.

Suppose we are given a signature σ' disjoint from σ. If $\mathcal{A} \in \mathrm{STRUCT}[\sigma]$, $\vec{a}$ is a k-tuple of elements of A, and $\vec{C}$ is an interpretation of σ' predicates as relations of appropriate arity over A, we write $(\mathcal{A}, \vec{C}, \vec{a})$ for the corresponding structure in the language of $\sigma \cup \sigma'$ union constants for elements of $\vec{a}$. By $adom(\vec{C})$ we mean the active domain of $\vec{C}$, that is, the set of all elements of A that occur in relations from $\vec{C}$. We then write, for $d \geq r$,

$$(\mathcal{A}, \vec{C}, \vec{a}) \quad \sim_d^r \quad (\mathcal{B}, \vec{D}, \vec{b})$$

if $\vec{D}$ interprets σ' over B, if $\vec{a}$, $\vec{b}$ are of the same length, and if the following three conditions hold:

(1) $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$,

(2) $adom(\vec{C}) \subseteq S_r^{\mathcal{A}}(\vec{a})$ and $adom(\vec{D}) \subseteq S_r^{\mathcal{B}}(\vec{b})$, and

(3) there exists an isomorphism $h : N_d^{\mathcal{A}}(\vec{a}) \to N_d^{\mathcal{B}}(\vec{b})$ such that $h(\vec{C}) = \vec{D}$.

We next prove the following lemma, which implies the *if* direction of the theorem by simply taking σ' to be empty. From now on, we shall often be listing free second-order variables explicitly, for bookkeeping convenience.

LEMMA 4.3. *Let $\varphi(\vec{x}, \vec{\imath}, \vec{X})$ be a $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ formula. Then there exists a number $d \geq r$ such that, for every interpretation $\vec{\imath}_0$ of $\vec{\imath}$, it is the case that $(\mathcal{A}, \vec{a}, \vec{C}) \sim_d^r (\mathcal{B}, \vec{b}, \vec{D})$ implies*

$$\mathcal{A} \models \varphi(\vec{a}, \vec{\imath}_0, \vec{C}) \quad \textit{iff} \quad \mathcal{B} \models \varphi(\vec{b}, \vec{\imath}_0, \vec{D}).$$

PROOF. By induction on formulae. Let $\mathsf{rk}_0(\varphi)$ be defined as $\mathsf{rk}(\varphi)$ but without taking into account second-order quantification (in particular, $\mathsf{rk}_0(\varphi) \leq \mathsf{rk}(\varphi)$). We show that d can be taken to be $9^m r + \frac{9^m - 1}{2}$ where $m = \mathsf{rk}_0(\varphi)$. That is, for the sequence defined by $d_0 = r$, …, $d_{i+1} = 9d_i + 4$, …, it is the case that d in the lemma can be taken to be d_m.

The case of atomic formulae not involving symbols from σ' is straightforward, as $\vec{a}$ and $\vec{b}$ satisfy all the same atomic σ-formulae if $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$ for any $d \geq 0$. For the case of atomic σ'-formulae, one can take $d = r$. Indeed, assume $(\mathcal{A}, \vec{a}, \vec{C}) \sim_r^r (\mathcal{B}, \vec{b}, \vec{D})$, and $T(\vec{a}_0)$ holds, where $\vec{a}_0$ is a subtuple of $\vec{a}$ of the same length as the arity of a σ'-symbol T. Let $\vec{b}_0$ be the corresponding subtuple of $\vec{b}$. We must show that $T(\vec{b}_0)$ holds. Assume T is interpreted by $C_0 \in \vec{C}$ over $\mathcal{A}$ and $D_0 \in \vec{D}$ over $\mathcal{B}$. We have an isomorphism $h : N_r^{\mathcal{A}}(\vec{a}) \to N_r^{\mathcal{B}}(\vec{b})$ with $h(\vec{C}) = \vec{D}$, and in particular $h(C_0) = D_0$. Since $h(\vec{a}_0) = \vec{b}_0$, we obtain from $\vec{a}_0 \in C_0$ that $\vec{b}_0 \in D_0$, thus showing that $T(\vec{b}_0)$ holds over $\mathcal{B}$.

The cases of negation, infinitary connectives, and quantification over the numerical sort are proved just as in the proof of Theorem 3.1.

Next, consider the case of local second-order quantification. Given a formula

$$\psi(\vec{x}_1\vec{x}_2, \vec{\imath}, \vec{X}) \equiv \exists Y \sqsubseteq S_{r'}(\vec{x}_1)\ \varphi(\vec{x}_1\vec{x}_2, \vec{\imath}, Y, \vec{X})$$

for some l-ary symbol Y and $r' \leq r$, let d be given by applying the hypothesis to φ. We must show that this d works for ψ. Fix $\vec{\imath}_0$. Assume $(\mathcal{A}, \vec{a}, \vec{C}) \sim_d^r (\mathcal{B}, \vec{b}, \vec{D})$ where

$\vec{C}$ and $\vec{D}$ are interpretations of $\vec{X}$. We then have an isomorphism $h : N_d^{\mathcal{A}}(\vec{a}) \to N_d^{\mathcal{B}}(\vec{b})$ such that $h(\vec{C}) = \vec{D}$. Assume that $\mathcal{A} \models \psi(\vec{a}, \vec{\imath}_0, \vec{C})$. Then we can find a set $V \subseteq (S_{r'}^{\mathcal{A}}(\vec{a}_1))^l$ such that $\mathcal{A} \models \varphi(\vec{a}, \vec{\imath}_0, V, \vec{C})$. Here $\vec{a}_1$ is the subtuple of $\vec{a}$ corresponding to $\vec{x}_1$. Note that $adom(V) \subseteq S_d^{\mathcal{A}}(\vec{a})$; thus we can define $U = h(V)$. Since h is an isomorphism, $U \subseteq (S_{r'}^{\mathcal{B}}(\vec{b}_1))^l$, and hence all the conditions for $(\mathcal{A}, \vec{a}, V, \vec{C}) \sim_d^r (\mathcal{B}, \vec{b}, U, \vec{D})$ hold. Since $\mathcal{A} \models \varphi(\vec{a}, \vec{\imath}_0, V, \vec{C})$, by the hypothesis this implies $\mathcal{B} \models \varphi(\vec{b}, \vec{\imath}_0, U, \vec{D})$ and thus $\mathcal{B} \models \psi(\vec{b}, \vec{\imath}_0, \vec{D})$. The converse (that is, $\mathcal{B} \models \psi(\cdots)$ implies $\mathcal{A} \models \psi(\cdots)$) is identical, which proves the case of second-order quantification.

In preparation for the case of counting quantifiers, we need the following.

Claim 4.4. Assume $(\mathcal{A}, \vec{a}) \leftrightarrows_{9d+4} (\mathcal{B}, \vec{b})$. Let h be an arbitrary isomorphism $N_{9d+4}^{\mathcal{A}}(\vec{a}) \to N_{9d+4}^{\mathcal{B}}(\vec{b})$. Then there exists a bijection $f : A \to B$ such that on $S_{6d+3}(\vec{a})$ it coincides with h, and $(\mathcal{A}, \vec{a}c) \leftrightarrows_d (\mathcal{B}, \vec{b}f(c))$ for every $c \in A$.

Proof. By Lemma 2.5, part (a), $\mathcal{A} \leftrightarrows_d \mathcal{B}$ and $\vec{a} \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}$ imply $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$. We use this as follows. The assumptions show that $\mathcal{A} \leftrightarrows_{9d+4} \mathcal{B}$ and $\vec{a} \approx_{9d+4}^{\mathcal{A},\mathcal{B}} \vec{b}$. Fix an isomorphism $h : N_{9d+4}^{\mathcal{A}}(\vec{a}) \to N_{9d+4}^{\mathcal{B}}(\vec{b})$; clearly it maps $S_{6d+3}^{\mathcal{A}}(\vec{a})$ onto $S_{6d+3}^{\mathcal{B}}(\vec{b})$ as it preserves distances. Consider any isomorphism type τ of a $3d+1$-neighborhood of a single point. Suppose $c \in S_{6d+3}^{\mathcal{A}}(\vec{a})$ realizes τ; since $S_{3d+1}^{\mathcal{A}}(c) \subseteq S_{9d+4}^{\mathcal{A}}(\vec{a})$, it follows that $f(c) \in S_{6d+3}^{\mathcal{B}}(\vec{b})$ realizes τ in $\mathcal{B}$. Thus, there are equally many realizers of τ in $S_{6d+3}^{\mathcal{A}}(\vec{a})$ and $S_{6d+3}^{\mathcal{B}}(\vec{b})$. Since $\mathcal{A} \leftrightarrows_{9d+4} \mathcal{B}$ implies $\mathcal{A} \leftrightarrows_{3d+1} \mathcal{B}$ (cf. [Fagin et al. 1995]), there are equally many realizers of τ in $\mathcal{A}$ and $\mathcal{B}$, and thus there exists a bijection $g : A - S_{6d+3}^{\mathcal{A}}(\vec{a}) \to B - S_{6d+3}^{\mathcal{B}}(\vec{b})$ that preserves isomorphism types of $3d+1$-neighborhoods.

We now define $f : A \to B$ as follows: $f(c) = h(c)$ if $c \in S_{6d+3}^{\mathcal{A}}(\vec{a})$, and $f(c) = g(c)$ otherwise. Clearly, this is a bijection, that coincides with h on $S_{6d+3}^{\mathcal{A}}(\vec{a})$. Now consider an arbitrary $c \in A$. If $c \in S_{6d+3}^{\mathcal{A}}(\vec{a})$, then $S_{3d+1}^{\mathcal{A}}(c) \subseteq S_{9d+4}^{\mathcal{A}}(\vec{a})$ and hence $\vec{a}c \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}f(c)$, since $f(c) = h(c)$ and since h is an isomorphism. If $c \notin S_{6d+3}^{\mathcal{A}}(\vec{a})$, then $f(c) = g(c) \notin S_{6d+3}^{\mathcal{B}}(\vec{b})$ has the same type of its $3d+1$-neighborhood as c, and again $\vec{a}c \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}f(c)$ since there cannot be elements from $S_{3d+1}^{\mathcal{A}}(\vec{a})$ and $S_{3d+1}^{\mathcal{A}}(\vec{c})$ that occur together in a tuple of a σ-relation in $\mathcal{A}$ (because the distance between $\vec{a}$ and c is at least $6d+4$) and likewise for $\vec{b}$ and $f(c)$. Thus, we have $\vec{a}c \approx_{3d+1}^{\mathcal{A},\mathcal{B}} \vec{b}f(c)$ for every c, which together with $\mathcal{A} \leftrightarrows_d \mathcal{B}$ implies $(\mathcal{A}, \vec{a}c) \leftrightarrows_d (\mathcal{B}, \vec{b}f(c))$. This proves the claim. □

We now consider the case of a formula

$$\psi(\vec{x}, \vec{\imath}, \vec{X}) \equiv \exists iz \; \varphi(\vec{x}, z, \vec{\imath}, \vec{X}).$$

Applying the hypothesis to φ, we obtain a number $d \geq r$ such that for every $\vec{\imath}_0$, $(\mathcal{A}, \vec{a}, c, \vec{C}) \sim_d^r (\mathcal{B}, \vec{b}, e, \vec{D})$ implies that $\mathcal{A} \models \varphi(\vec{a}, c, \vec{\imath}_0, \vec{C})$ iff $\mathcal{B} \models \varphi(\vec{b}, e, \vec{\imath}_0, \vec{D})$. To conclude, we must prove that $(\mathcal{A}, \vec{a}, \vec{C}) \sim_{9d+4}^r (\mathcal{B}, \vec{b}, \vec{D})$ implies that $\mathcal{A} \models \psi(\vec{a}, \vec{\imath}_0, \vec{C})$ iff $\mathcal{B} \models \psi(\vec{b}, \vec{\imath}_0, \vec{D})$. For this, it will suffice to establish a bijection $f : A \to B$ such that for every c, $(\mathcal{A}, \vec{a}, c, \vec{C}) \sim_d^r (\mathcal{B}, \vec{b}, f(c), \vec{D})$. Then clearly the number of elements satisfying φ will be preserved.

Since $(\mathcal{A}, \vec{a}, \vec{C}) \sim^r_{9d+4} (\mathcal{B}, \vec{b}, \vec{D})$ and $d \geq r$, we have $(\mathcal{A}, \vec{a}) \leftrightarrows_{9d+4} (\mathcal{B}, \vec{b})$, and $h(\vec{C}) = \vec{D}$ for some isomorphism $h : N^{\mathcal{A}}_{9d+4}(\vec{a}) \to N^{\mathcal{B}}_{9d+4}(\vec{b})$; moreover, $adom(\vec{C})$ is contained in $S^{\mathcal{A}}_r(\vec{a}) \subseteq S^{\mathcal{A}}_d(\vec{a})$, and likewise for $\vec{D}$ in $\mathcal{B}$. Applying Claim 4.4, we obtain a bijection $f : A \to B$ that coincides with h on $S^{\mathcal{A}}_{6d+3}(\vec{a})$ and such that $(\mathcal{A}, \vec{a}c) \leftrightarrows_d (\mathcal{B}, \vec{b}f(c))$ for every c.

Thus, to conclude that $(\mathcal{A}, \vec{a}, c, \vec{C}) \sim^r_d (\mathcal{B}, \vec{b}, f(c), \vec{D})$ we must only show that for every c, there is an isomorphism $h_c : N^{\mathcal{A}}_d(\vec{a}c) \to N^{\mathcal{B}}_d(\vec{b}f(c))$ with $h_c(\vec{C}) = \vec{D}$, as other conditions are clearly satisfied. First, assume $c \notin S^{\mathcal{A}}_{2d+1}(\vec{a})$. Then $f(c) \notin S^{\mathcal{B}}_{2d+1}(\vec{b})$, since f coincides with h on $S^{\mathcal{A}}_{6d+3}(\vec{a})$. Hence, $S^{\mathcal{A}}_d(\vec{a}c)$ is a disjoint union of $S^{\mathcal{A}}_d(\vec{a})$ and $S^{\mathcal{A}}_d(c)$ (and likewise for $S^{\mathcal{B}}_d(\vec{b}f(c))$), and thus there exists an isomorphism $h_c : N^{\mathcal{A}}_d(\vec{a}c) \to N^{\mathcal{B}}_d(\vec{b}f(c))$ which coincides with h on $S^{\mathcal{A}}_d(\vec{a})$; as $adom(\vec{C}) \subseteq S^{\mathcal{A}}_d(\vec{a})$, this implies $h(\vec{C}) = \vec{D}$. Assuming $c \in S^{\mathcal{A}}_{2d+1}(\vec{a})$, we have $f(c) = h(c)$ and $S^{\mathcal{A}}_d(c) \subseteq S^{\mathcal{A}}_{3d+1}(\vec{a})$, and $S^{\mathcal{B}}_d(f(c)) \subseteq S^{\mathcal{B}}_{3d+1}(\vec{b})$. Thus, in this case h maps $N^{\mathcal{A}}_d(\vec{a}c)$ isomorphically onto $N^{\mathcal{B}}_d(\vec{b}f(c))$, and hence $h(\vec{C}) = \vec{D}$ for h_c being a proper restriction of h. This concludes the proof for the case of counting quantifiers, and thus the proof of the lemma and the *if* part of the theorem.

PROOF. (Only if) Let Q be an m-ary query with $\mathsf{hlr}(Q) \leq r$, $r > 0$. We show that Q is definable by a formula of $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$. Consider some enumeration $\tau_i, i \in \mathbb{N}_+$ of all isomorphism types of r-neighborhoods of $m+1$-tuples in structures from STRUCT$[\sigma]$. Note that there countably many of those. Suppose $K = \{(i_1, j_1), \ldots, (i_l, j_l)\}$ is a finite subset of $\mathbb{N}_+ \times \mathbb{N}_+$ with all i_ps being distinct. We write $ntp_r(\mathcal{A}, \vec{a}) \triangleright K$ if there are exactly j_p elements c such that the type of $N^{\mathcal{A}}_r(\vec{a}c)$ is τ_{i_p}, and the cardinality of A is $j_1 + \ldots + j_p$ (that is, $\tau_{i_1}, \ldots, \tau_{i_l}$ are the only isomorphism types of $N^{\mathcal{A}}_r(\vec{a}c)$ as c ranges over A). Then Q is uniquely determined by a collection $\mathbf{B}_Q$ of finite subsets K of $\mathbb{N}_+ \times \mathbb{N}_+$ which are graphs of partial functions. That is, there exists a collection $\mathbf{B}_Q$ of such sets K such that $\vec{a} \in Q(\mathcal{A})$ iff $ntp_r(\mathcal{A}, \vec{a}) \triangleright K$ for some $K \in \mathbf{B}_Q$. Conversely, for any collection $\mathbf{B}$ of finite partial functions $K \subset \mathbb{N}_+ \times \mathbb{N}_+$, the query defined by $\vec{a} \in Q(\mathcal{A})$ iff $ntp_r(\mathcal{A}, \vec{a}) \triangleright K$ for some $K \in \mathbf{B}$ is Hanf-local with $\mathsf{hlr}(Q) \leq r$. This follows directly from the definition of Hanf-locality. Thus, the $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ formula defining Q is

$$\bigvee_{K \in \mathbf{B}_Q} \psi_K(\vec{x}),$$

where $\mathcal{A} \models \psi_K(\vec{a})$ iff $ntp_r(\mathcal{A}, \vec{a}) \triangleright K$. Furthermore, the formulae ψ_K are defined in such a way that there is an upper bound on $\mathsf{rk}(\psi_K)$ that depends only on m, r and σ; this ensures that the infinite disjunction above is a $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ formula.

It thus remains to show how to define ψ_K by a formula whose rank is determined by m, r, and σ only. For $K = \{(i_1, j_1), \ldots, (i_l, j_l)\}$, it is defined as

$$\psi_K(\vec{x}) \quad \equiv \quad \bigwedge_{p=1}^{l} \exists^{=j_p} y \ \nu_r^{\tau_{i_p}}(\vec{x}, y)$$

where $\exists^{=j} y\varphi$ is an abbreviation for $\exists j y\varphi \wedge \neg\exists(j+1)y\varphi$ (or $\#y.\varphi = j$) and $\mathcal{A} \models \nu_r^{\tau_{i_p}}(\vec{a}, c)$ iff the isomorphism type of $N^{\mathcal{A}}_r(\vec{a}c)$ is τ_{i_p}.

To conclude the proof, we show, for arbitrary r, n, and an isomorphism type τ of an n-tuple, $n > 0$, how to define $\nu_r^\tau(\vec{x})$ such that $\mathcal{A} \models \nu_r^\tau(\vec{a})$ iff $N_r^{\mathcal{A}}(\vec{a})$ is of type τ. Let neighborhoods of type τ contain N elements. (Note that for this construction, we only need to consider the case when $\vec{x}$ is nonempty, and hence $N > 0$.) Fix a neighborhood $\mathcal{N}$ realizing τ, with $a_1, \ldots, a_n$ interpreting $\vec{x}$, and let $e_1, \ldots, e_{N-n}$ be any enumeration of the remaining elements. For each k-ary relation R from σ, a k-tuple $\vec{t}$ over $\vec{a}, \vec{e}$, and a binary relational symbol L not in σ, define a $\sigma \cup \{L\}$-formula $\alpha_R^{\vec{t}}(\vec{x})$ of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ as follows. Suppose $e_{j_1}, \ldots, e_{j_s}$ is the subtuple of $\vec{t}$ containing the elements of $\vec{e}$. Then $\alpha_R^{\vec{t}}(\vec{x})$ is

$$\exists y_1, \ldots, y_s \; R(\vec{x}, \vec{y}) \;\wedge\; \bigwedge_i \left(\begin{array}{l} (y_i \in S_r(\vec{x})) \wedge \\ (\#z.(z \in S_r(\vec{x}) \wedge L(z, y_i) \wedge \bigwedge_l \neg(z = x_l)) = j_i) \end{array} \right)$$

where by $R(\vec{x}, \vec{y})$ we mean that the position corresponding to a_i in $\vec{t}$ is occupied by x_i, and the position corresponding to e_{j_i} is occupied by by y_i. This formula says that for L defining the linear ordering corresponding to $e_1, \ldots, e_{N-n}$ on $S_r(\vec{a}) - \vec{a}$, the tuple extending $\vec{a}$ with elements occurring in the positions of $e_{j_1}, \ldots, e_{j_s}$ in the ordering, belongs to R. Note that the membership in $S_r(\vec{x})$ can be tested by an FO formula whose rank is at most $r + p_\sigma - 1$, where p_σ is the maximum arity of a relation in σ (with σ being nonempty, $p_\sigma > 0$). Thus, $\alpha_R^{\vec{t}}$ is an $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ formula, whose rank is at most $s + r + p_\sigma \leq r + 2p_\sigma$.

We now define a formula $\beta_r(\vec{x}, L)$ as

$$\bigwedge_{R(\vec{t}) \in \mathrm{Diag}(\mathcal{N})} \alpha_R^{\vec{t}}(\vec{x}) \;\;\wedge \bigwedge_{R(\vec{t}) \notin \mathrm{Diag}(\mathcal{N})} \neg\alpha_R^{\vec{t}}(\vec{x}),$$

where $\mathrm{Diag}(\mathcal{N})$ is the diagram of the neighborhood $\mathcal{N}$. This formula says that exactly $\mathrm{Diag}(\mathcal{N})$ atomic formulae hold in $N_r(\vec{x})$, assuming L defines an ordering on $S_r(\vec{x}) - \vec{x}$ consistent with that chosen on $\mathcal{N}$. Let $\gamma(\vec{x}, L)$ be an FO formula saying that L defines a linear order on $S_r(\vec{x}) - \vec{x}$. We then conclude that

$$\exists L \sqsubseteq S_r(\vec{x}) \;\; ((\#z.z \in S_r(\vec{x}) = N) \wedge \beta(\vec{x}, L) \wedge \gamma(\vec{x}, L))$$

defines $\nu_r^\tau(\vec{x})$. Indeed, if $\mathcal{A} \models \nu_r^\tau(\vec{a})$, then the diagram of $N_r^{\mathcal{A}}(\vec{a})$ is the same as that of $\mathcal{N}$ for some ordering on $S_r^{\mathcal{A}}(\vec{a}) - \vec{a}$, and thus the type of $N_r^{\mathcal{A}}(\vec{a})$ is τ. If the type of $N_r^{\mathcal{A}}(\vec{a})$ is τ, the choose the ordering as in $\mathcal{N}$ to see that $\nu_r^\tau(\vec{a})$ holds. We finally note that ν increases the rank of the α's by at most $r + p_\sigma + 1$. Then $\mathsf{rk}(\nu_r^{\tau_{i_p}}) \leq r + p_\sigma + 1 + r + 2p_\sigma = 2r + 3p_\sigma + 1$ and hence $\mathsf{rk}(\psi_K) \leq 2r + 3p_\sigma + 2$. This concludes the proof of definability of Q in $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$, and thus proves the theorem. $\square$

There are several corollaries to the proof. First notice that if we defined $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ without increasing the rank of a formula for every second-order local quantifier, the proof would go through verbatim. We can also define a logic $\mathbb{L}^r_{\infty\omega}(\mathbf{C})$ just as $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ except that first-order local quantification $\exists z \in S_r(\vec{x})$ and $\forall z \in S_r(\vec{x})$ is used in place of second-order local quantifiers, and those local quantifiers do not increase the rank (in particular, the depth of their nesting can be infinite, which allows one to define arbitrary computations on those neighborhoods). Let then $\mathbb{L}^*_{\infty\omega}(\mathbf{C})$ be $\bigcup_r \mathbb{L}^r_{\infty\omega}(\mathbf{C})$. The proof of Hanf-locality of $\mathbb{L}^*_{\infty\omega}(\mathbf{C})$

goes through as before, and proving that every Hanf-local query is definable in $\mathbb{L}^*_{\infty\omega}(\mathbf{C})$ is very similar to that of $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ as with infinitely many local first-order quantifiers we can write out diagrams of neighborhoods. We thus obtain:

COROLLARY 4.5. *The following have the same expressive power as* $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ *(and thus capture Hanf-local properties):*

—*the logic obtained from* $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ *by allowing the depth of nesting of local quantifiers to be infinite and*
—*the logic* $\mathbb{L}^*_{\infty\omega}(\mathbf{C})$.

Analyzing the proof of Theorem 4.2, we also obtain the following normal form for $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ formulae, which shows that the depth of nesting of local second-order quantifiers need not exceed 1.

COROLLARY 4.6. *Every* $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ *formula* $\varphi(\vec{x})$ *is equivalent to a formula in the form*

$$\bigvee_i \bigwedge_j (n_{ij} = \#y.(\exists S \sqsubseteq S_d(\vec{x})\ \psi_{ij}(\vec{x}, y, S)))$$

where the conjunctions are finite, S *is binary, and each* ψ_{ij} *is a* $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ *formula.*

As a final remark, we note that $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ is strictly more expressive than $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ extended with tests for neighborhood isomorphisms.

PROPOSITION 4.7. $\bigcup_{d>0}(\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I^k_d \mid k > 0\}) \subsetneq \mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$.

PROOF. Consider a signature σ that consists of three binary relations E_1, E_2, and T. We shall use the notation $adom(E_i)$ for the set of elements of σ-structures that occur in E_i-tuples, $i = 1, 2$. We now define the following Boolean query Q on STRUCT$[\sigma]$: $Q(\mathcal{A})$ is true iff T is the total relation on $A \neq \emptyset$ ($T = A \times A$), and E_1- and E_2-reducts of $\mathcal{A}$ are isomorphic as graphs. This is definable in $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$. First note that if T is the total relation, then for every $a \in A$, $S_1^{\mathcal{A}}(a) = A$. Thus, we define Q by the conjunction of $\forall x \forall y\ T(x, y)$ and the sentence

$$\exists x \exists F \sqsubseteq S_1(x) \left(\begin{array}{l} \quad \mathit{function}(F) \\ \wedge\ \mathit{dom}(D) = \mathit{adom}(E_1) \\ \wedge\ \mathit{codom}(D) = \mathit{adom}(E_2) \\ \wedge\ \forall x, y, u, v\ F(x, u) \wedge F(y, v) \rightarrow (E_1(x, y) \leftrightarrow E_2(u, v)) \end{array} \right)$$

which asserts that T is total and that an isomorphism F exists (since $S_1^{\mathcal{A}}(a) = A$, the second-order quantification is over the entire universe). Here $\mathit{function}(F)$ is a first-order sentence stating that F is a 1-1 function, $\mathit{dom}(D) = \mathit{adom}(E_1)$ is an FO sentence saying that F's domain is $adom(E_1)$, and $\mathit{codom}(D) = \mathit{adom}(E_2)$ is an FO sentence saying that F's codomain is $adom(E_2)$.

To prove that Q is not definable in $\bigcup_{d>0}(\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I^k_d \mid k > 0\})$, define a class $\mathcal{C}$ of nonempty σ-structures as follows. In a structure $\mathcal{A}$ in $\mathcal{C}$, T is interpreted as a total relation (that is, A^2), A is the disjoint union of $adom(E_1)$ and $adom(E_2)$, and E_1 and E_2 are successor relations, possibly with loops on some nodes.

We now assume that Q is definable by a sentence Φ of $\bigcup_{d>0}(\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I^k_d \mid k > 0\})$ of rank m. Consider any occurrence of $I^k_d(\vec{x}, \vec{y})$ in Φ. Suppose we have a structure $\mathcal{A}$ from $\mathcal{C}$. Since $S_1^{\mathcal{A}}(a) = A$ for all $a \in A$, $\mathcal{A} \models I^k_d(\vec{a}, \vec{b})$, $\vec{a}, \vec{b} \in A^k$ means that there exists an automorphism $h : \mathcal{A} \to \mathcal{A}$ such that $h(\vec{a}) = \vec{b}$. However, since E_1 and E_2 are disjoint successor relations (perhaps with loops on some nodes), the structure $\mathcal{A}$ is rigid, and thus h must be the identity. Hence, $\mathcal{A} \models I_d^k(\vec{a}, \vec{b})$ iff $\vec{a} = \vec{b}$. Using this, construct a sentence Φ' of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ of rank m by replacing each $I_d^k(\vec{x}, \vec{y})$ in Φ with $\bigwedge_i (x_i = y_i)$. We thus showed that for any $\mathcal{A}$ in $\mathcal{C}$, $\mathcal{A} \models \Phi$ iff $\mathcal{A} \models \Phi'$.

It remains to show that Q cannot be expressed by an $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ sentence of rank m on $\mathcal{C}$. Construct two structures $\mathcal{A}, \mathcal{B}$ in the class $\mathcal{C}$. In both of them, the E_1- and E_2-successor relations have length $2 \cdot 3^m + 3$. In $\mathcal{A}$, there are loops on the nodes in E_1 and E_2 at the same distance $3^m + 1$ from the start node. In $\mathcal{B}$, there is one loop on E_1 at the distance $3^m + 1$ from the start, and one loop on E_2 at the distance $3^m + 2$ from the start. Hence, Q is true on $\mathcal{A}$ and false on $\mathcal{B}$.

Let $\mathcal{A}'$ and $\mathcal{B}'$ be the (E_1, E_2) reducts of $\mathcal{A}$ and $\mathcal{B}$ respectively. Then $\mathcal{A}' \leftrightarrows_{3^m} \mathcal{B}'$, since the nodes with loops are at the distance at least $3^m + 1$ from the start and end nodes of the successor relations. Hence, by [Hella et al. 1999a; Nurmonen 1996], the duplicator wins the m-round bijective Ehrenfeuct-Fraïssé game on $\mathcal{A}'$ and $\mathcal{B}'$. This shows in turn that $\mathcal{A} \equiv_m^{bij} \mathcal{B}$. Indeed, for each round of the game, the duplicator just forgets the T-relation, and uses the strategy for $\mathcal{A}'$ and $\mathcal{B}'$ to pick his bijection. We know that after each round i, the points $(a_1, \ldots, a_i)$ and $(b_1, \ldots, b_i)$ played in A and B respectively define a partial isomorphism with respect to E_1 and E_2. Since $(a_l, a_k) \in T$ iff $(b_l, b_k) \in T$ for all l, k, it follows that they define a partial isomorphism $\mathcal{A} \to \mathcal{B}$. We thus found two structures $\mathcal{A} \equiv_m^{bij} \mathcal{B}$ in $\mathcal{C}$ that disagree on Q, showing that on $\mathcal{C}$, Q cannot be defined by an $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ sentence of rank m. Hence, by the above, Q cannot be defined by a sentence of Φ of $\bigcup_{d>0}(\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I_d^k \mid k > 0\})$. This completes the proof. □

5. CHARACTERIZING GAIFMAN-LOCAL PROPERTIES

We now turn to Gaifman's notion of locality, which states that a query Q is local with $\mathsf{lr}(Q) \le r$ if $N_r^{\mathcal{A}}(\vec{a}_1) \cong N_r^{\mathcal{A}}(\vec{a}_2)$ implies that $\vec{a}_1 \in Q(\mathcal{A})$ iff $\vec{a}_2 \in Q(\mathcal{A})$. For structures of bounded valence, this notion was characterized by first-order definition by cases. An extended version of this notion captures Gaifman-locality in the general case.

Definition 5.1. An m-ary query, $m > 0$, on STRUCT$[\sigma]$ is given by a *Hanf-local definition by cases* if there exists a finite or countable partition of STRUCT$[\sigma]$ into classes $\mathcal{C}_i$, $i \in \mathbb{N}$, a number $d \ge 0$, and Hanf-local queries Q_i, $i \in \mathbb{N}$, with $\mathsf{hlr}(Q_i) \le d$, such that for every i and every $\mathcal{A} \in \mathcal{C}_i$, it is the case that $Q(\mathcal{A}) = Q_i(\mathcal{A})$.

THEOREM 5.2. *A query is Gaifman-local iff it is given by a Hanf-local definition by cases.*

PROOF. Assume that Q is given by a Hanf-local definition by cases. Let d be an upper bound on $\mathsf{hlr}(Q_i)$. We claim that Q is Gaifman-local and $\mathsf{lr}(Q) \le 3d + 1$. Fix $\mathcal{A}$, and assume $\mathcal{A} \in \mathcal{C}_i$. Let $\vec{a}_1 \approx_{3d+1}^{\mathcal{A}} \vec{a}_2$. Then by Lemma 2.5 we obtain $(\mathcal{A}, \vec{a}_1) \leftrightarrows_d (\mathcal{A}, \vec{a}_2)$, and Hanf-locality of Q_i implies $\vec{a}_1 \in Q_i(\mathcal{A}) = Q(\mathcal{A})$ iff $\vec{a}_2 \in Q_i(\mathcal{A}) = Q(\mathcal{A})$.

Conversely, let a Gaifman-local Q be given, with $\mathsf{lr}(Q) = d$. Let $\tau_1, \tau_2 \ldots$ be an enumeration of isomorphism types of finite σ-structures. Let $\mathcal{C}_i$ be the class of structures of type τ_i. We define Q_i as follows: $\vec{b} \in Q_i(\mathcal{B})$ iff there exists $\mathcal{A}$ of type τ_i and $\vec{a} \in A^m$ such that $(\mathcal{B}, \vec{b}) \leftrightarrows_d (\mathcal{A}, \vec{a})$ and $\vec{a} \in Q(\mathcal{A})$.

First show that each Q_i is Hanf-local, with $\mathsf{hlr}(Q_i) \leq d$. Let $(\mathcal{B}_1, \vec{b}_1) \leftrightarrows_d (\mathcal{B}_2, \vec{b}_2)$. Assume $\vec{b}_1 \in Q_i(\mathcal{B}_1)$. Then for some $\mathcal{A}$ of type τ_i and $\vec{a} \in A^m$ such that $(\mathcal{B}_1, \vec{b}_1) \leftrightarrows_d (\mathcal{A}, \vec{a})$ we have $\vec{a} \in Q(\mathcal{A})$. We thus have $(\mathcal{B}_2, \vec{b}_2) \leftrightarrows_d (\mathcal{A}, \vec{a})$, and hence $\vec{b}_2 \in Q_i(\mathcal{B}_2)$. The converse (that $\vec{b}_2 \in Q_i(\mathcal{B}_2)$ implies $\vec{b}_1 \in Q_i(\mathcal{B}_1)$) is identical.

To conclude, we must show that for every $\mathcal{A}$ of type τ_i, $Q(\mathcal{A}) = Q_i(\mathcal{A})$. Assume first that $\vec{a} \in Q_i(\mathcal{A})$. Then for some $\mathcal{A}' \cong \mathcal{A}$ and $\vec{a}'$ such that $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{A}', \vec{a}')$ we have $\vec{a}' \in Q(\mathcal{A}')$. Let h be an isomorphism $\mathcal{A} \to \mathcal{A}'$. Since the isomorphism type of the d-neighborhood of $h(\vec{a})$ in $\mathcal{A}'$ is the same as that of the d-neighborhood of $\vec{a}'$, it follows from Gaifman-locality that $h(\vec{a}) \in Q(\mathcal{A}')$. Since queries are closed under isomorphisms, applying h^{-1} we get $\vec{a} \in Q(\mathcal{A})$. Conversely, assume $\vec{a} \in Q(\mathcal{A})$. Since $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{A}, \vec{a})$ we obtain $\vec{a} \in Q_i(\mathcal{A})$. This concludes the proof. □

Unlike in Fact 2.9, the number of cases in a Hanf-local definition by cases can be infinite. A natural question to ask is whether a finite number of cases is sufficient (in particular, whether the statement of Fact 2.9 holds for arbitrary finite structures). We now show that the infinite number of cases is unavoidable. In fact, we show a stronger result.

Definition 5.3. For $k > 0$, let LOCAL_k be the class of queries given by a Hanf-local definition by cases, where the number of cases is at most k. Let LOCAL^* be $\bigcup_{k>0} \text{LOCAL}_k$, and G_LOCAL be the class of all Gaifman-local queries.

Note that LOCAL_1 is precisely the class of Hanf-local queries.

THEOREM 5.4. *The hierarchy*

$$\text{LOCAL}_1 \subset \text{LOCAL}_2 \subset \ldots \subset \text{LOCAL}^* \subset \text{G\_LOCAL}$$

is strict.

PROOF. We first exhibit a query $Q \in \text{LOCAL}_{l+1} - \text{LOCAL}_l$. Intuitively, a query from LOCAL_l cannot make $l+1$ choices, and thus is different from every query in LOCAL_{l+1} on some class of the partition. More precisely, we define a class $\mathcal{C}_i^{l+1}$, $1 \leq i \leq l+1$, of graphs that consists of graphs with the number of connected components being $i-1$ modulo $l+1$. Define Q_i^{l+1} as a query returning the set of nodes that can be reached by a path of length $i-1$ from a node of indegree 0. That is, if the input is a successor relation, this query returns the ith node. Clearly, Q_i^{l+1} is FO-definable and thus Hanf-local. We now form a query Q that coincides with Q_i^{l+1} on $\mathcal{C}_i^{l+1}$. (Note that Q is not FO, as the classes $\mathcal{C}_i^{l+1}$ are not FO-definable.) From Theorem 5.2, this is a Gaifman-local query, and it belongs to LOCAL_{l+1}.

Suppose Q is in LOCAL_l; that is, there is a partition of the class of all finite graphs into l classes $\mathcal{C}_1^l, \ldots, \mathcal{C}_l^l$ and Hanf-local queries Q_i^l such that on $\mathcal{C}_i^l$, Q coincides with Q_i^l, $i = 1, \ldots, l$. Let $d = 1 + \max \mathsf{hlr}(Q_i^l)$. Let G_0 be a successor relation on $l+1$ nodes. Define a graph H_i^{l+1} as the union of i cycles with $\frac{(l+1)!(2d+1)}{i}$ nodes each, $i = 1, \ldots, l+1$. As the total number of nodes in each H_i^{l+1} is $(l+1)!(2d+1)$ and all d-neighborhoods are isomorphic, we have $H_i^{l+1} \leftrightarrows_d H_j^{l+1}$ for all $i, j \leq l+1$. Let now G_i^{l+1} be the disjoint union of G_0 and H_i^{l+1}, $i = 1, \ldots, l+1$. If x and y are the nodes in the G_0 part of G_i^{l+1} and G_j^{l+1} respectively at the same distance from the start node, then $(G_i^{l+1}, x) \leftrightarrows_d (G_j^{l+1}, y)$.

By the pigeonhole principle, there exists a class $\mathcal{C}'_k$ and $i \neq j, i, j \leq l+1$ such that $G_i^{l+1}, G_j^{l+1} \in \mathcal{C}'_k$. Let x, y be the nodes at distance $i-1$ from the start node of the G_0 part of G_i^{l+1} and G_j^{l+1}, resp. Let z be the node at distance $j-1$ from the start node of the G_0 part of G_j^{l+1}; note that $z \neq y$. By definition of Q, it returns x on G_i^{l+1} and z on G_j^{l+1}. However, $(G_i^{l+1}, x) \leftrightarrows_d (G_j^{l+1}, y)$, and since Q is given on $\mathcal{C}'_k$ by Q'_k of $\mathsf{hlr}(Q'_k) \leq d$, it must return y on G_j^{l+1} if it returns x on G_i^{l+1}. This contradiction shows that $Q \notin \text{LOCAL}_l$.

To separate G_LOCAL from LOCAL*, we exhibit a query Q of $\mathsf{lr}(Q) = 1$ such that $Q \notin \text{LOCAL}^*$. Consider a signature consisting of two binary relations E_1 and E_2. Let Q be as follows: if no element of the universe occurs in an E_1-tuple and an E_2-tuple, if E_1 is a linear ordering, and if its length is at least the number k of connected components of E_2, then return the kth element in the linear order E_1; otherwise return nothing. Clearly this Q is of locality rank 1. In inputs on which the output of Q is not empty, two points with isomorphic 1-neighborhoods may only occur in E_2, and thus no such point belongs to the output of Q. We next show that $Q \notin \text{LOCAL}_l$ for each l. We consider the example we used to separate LOCAL_{l+1} from LOCAL_l, and modify it in such a way that in a structure G_i^{i+1}, G_0, which will interpret E_1 is a linear order of length $l+1$, and H_i^{l+1}, which interprets E_2, is the same as before. It again follows that $(G_i^{l+1}, x) \leftrightarrows_d (G_j^{l+1}, y)$, where x and y are in the same position in the linear order part G_0 of G_i^{l+1} and G_j^{l+1}. We then use the same pigeonhole argument as before to prove that $Q \notin \text{LOCAL}_l$. This concludes the proof. □

Thus, similarly to the case of Hanf-local queries, the characterization for structures of bounded valence fails to extend to the class of all finite structures.

COROLLARY 5.5. *There exist Gaifman-local queries that cannot be given by first-order definition by cases.*

6. CONCLUSION

Notions of locality have been used in logic numerous times. The local nature of first-order logic is particularly transparent when one deals with fragments corresponding to various modal logics; in general, Gaifman's and Hanf's theorems state that FO can only express local properties. These theorems were generalized, and, being applicable to finite structures, they found applications in areas such as complexity and databases.

However, while more and more powerful logics were proved to be local, there was no clear understanding of what kind of mechanisms can be added to logics while preserving locality. Here we answered this question by providing logical characterizations of local properties on finite structures. For Hanf-locality, arbitrary counting power and testing arbitrary properties of small neighborhoods can be added to first-order logic while retaining locality; moreover, with a limited form of infinitary connectives, such a logic captures all Hanf-local properties. For Gaifman-locality, one can in addition permit definition by cases, and the number of cases be either finite or infinite.

ACKNOWLEDGMENTS

I thank the anonymous referees for their suggestions.

REFERENCES

BARRINGTON, D.A.M., IMMERMAN, N., AND STRAUBING, H. 1990. On uniformity within NC^1. *JCSS*, 41:274–306.

BENEDIKT, M. AND KEISLER, H.J 1997. Expressive power of unary counters. In *Proceedings of International Conference on Database Theory (ICDT'97)*, Springer Lecture Notes in Computer Science, vol. 1186. Springer-Verlag, 291–305.

CAI, J., FÜRER, M., AND IMMERMAN, N. 1992. On optimal lower bound on the number of variables for graph identification. *Combinatorica 12*, 4, 389–410.

DONG, G., LIBKIN, L., AND WONG, L 2000. Local properties of query languages. *Theoretical Computer Science 239*, 1, 277–308.

EBBINGHAUS, H.-D., AND FLUM, J. 1995. *Finite Model Theory.* Springer Verlag, Berlin.

ETESSAMI, K. 1997. Counting quantifiers, successor relations, and logarithmic space, *J. Comput. Syst. Sci. 54*, 3, 400–411.

FAGIN, R., STOCKMEYER, L., AND VARDI, M.Y. 1995. On monadic NP vs monadic co-NP, *Information and Computation 120*, 1, 78–92.

GAIFMAN, H. 1982. On local and non-local properties, *Proceedings of the Herbrand Symposium, Logic Colloquium '81*, North Holland Publishing Co., Amsterdam.

GRÄDEL, E. 1999. On the restraining power of guards. *J. Symb. Logic 64*, 4, 1719–1742.

GRÄDEL, E. AND GUREVICH, Y. 1998. Metafinite model theory. *Information and Computation 140*, 1, 26–81.

GROHE, M. AND SCHWENTICK, T. 2000. Locality of order-invariant first-order formulas. *ACM TOCL*, 1, 1, 112–130.

HANF, W. 1965. Model-theoretic methods in the study of elementary logic. In J.W. Addison et al, eds, *The Theory of Models*, North Holland Publishing Co., Amsterdam, 132–145.

HELLA, L. 1996. Logical hierarchies in PTIME. *Information and Computation 129*, 1, 1–19.

HELLA, L., LIBKIN, L. AND NURMONEN, J. 1999a. Notions of locality and their logical characterizations over finite models. *J. Symb. Logic 64*, 4, 1751–1773.

HELLA, L., LIBKIN, L., NURMONEN, J. AND WONG, L. 1999b. Logics with aggregate operators. In *Proceedings of the 14th IEEE Symposium on Logic in Computer Science* (LICS'99, Trento, Italy, July), IEEE Press, Piscataway, NJ, 35–44. Full version to appear in *J. ACM.*

IMMERMAN, N. 1999. *Descriptive Complexity.* Springer Verlag, New York.

IMMERMAN, N. AND LANDER, E. 1990. Describing graphs: A first order approach to graph canonization. In *"Complexity Theory Retrospective"*, Springer Verlag, Berlin.

KOLAITIS, PH. AND VÄÄNÄNEN, J. 1995. Generalized quantifiers and pebble games on finite structures. *Annals of Pure and Applied Logic 74*, 23–75.

LIBKIN, L. 2000. On counting logics and local properties. *ACM TOCL*, 1, 1, 33–59.

LIBKIN, L. AND WONG, L. 1998. Unary quantifiers, transitive closure, and relations of large degree. In *Proceedings of Symp. on Theoretical Aspects of Comp. Sci.*, Springer Lecture Notes in Computer Science, vol. 1377, pages 183–193.

NURMONEN, J. 1996. On winning strategies with unary quantifiers. *J. Logic and Computation 6*, 3, 779–798.

OTTO, M. 1997. *Bounded Variable Logics and Counting: A Study in Finite Models.* Springer Verlag, New York.

SCHWENTICK, TH. AND BARTHELMANN, K. 1998. Local normal forms for first-order logic with applications to games and automata. In *Proceedings of Symp. on Theoretical Aspects of Comp. Sci.*, Springer Lecture Notes in Computer Science, vol. 1377, pages 444–454.

VAN BENTHEM, J.F.A.K. 1985. *Modal Logic and Classical Logic.* Bibliopolis, Naples.

Received March 2000; revised August 2000; accepted August 2000