Logics Capturing Local Properties

Leonid Libkin[1*]

Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974, USA.
Email: `libkin@research.bell-labs.com`

Abstract. Well-known theorems of Hanf's and Gaifman's 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.

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 [12, 8]. 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 extend to *finite* 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 [7].

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-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.

[*] Part of this work done while visiting INRIA.

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 [11, 14, 15, 19, 22]. Since the complexity class TC^0 (with the appropriate notion of uniformity) can be captured by FO with counting quantifiers [1], these results found applications in descriptive complexity, where they were used to prove lower bounds for logics coming very close to capturing TC^0 [6, 21]. They were also applied in database theory, where they were used to prove expressivity bounds for relational query languages with aggregation [4, 15] that correspond to practical query languages such as SQL. For applications to automata, see [24].

The abstract notions of locality were themselves characterized only 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 [19] as a counterpart of a finite variable logic $\mathcal{L}^\omega_{\infty\omega}$. While $\mathcal{L}^\omega_{\infty\omega}$ subsumes a number of fixpoint logics and is easier to study, $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ subsumes a number of counting extensions of FO (such as FO with counting quantifiers [17], FO with unary generalized quantifiers [13, 18], FO with unary counters [2]) and is quite easy to deal with. A result in [14] 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 [14] 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 show the strictness of the hierarchy. All proofs can be found in the full version [20].

2 Notation

Finite Structures and Logics 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 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* [5,8,7] $\mathcal{G}(\mathcal{A})$ 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:

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

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 $\mathcal{A}, \mathcal{B} \in$ 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 tuples $\vec{a} = (a_1, \ldots, a_n)$ and $\vec{b} = (b_1, \ldots, b_m)$, and an element c, we write $\vec{a}\vec{b}$ for the tuple $(a_1, \ldots, a_n, b_1, \ldots, b_m)$, and $\vec{a}c$ for $(a_1, \ldots, a_n, c)$.

Hanf's and Gaifman's theorems An *m-ary query* on σ-structures, Q, is a mapping that associates to each $\mathcal{A} \in$ 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 STRUCT$[\sigma]$ and definability means definability by a sentence of $\mathcal{L}$.

Definition 1. (cf. [4,14]) *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 $\vec{a}, \vec{b} \in A^m$*

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

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

Theorem 1 (Gaifman). *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 states that every formula is a Boolean combination of sentences of a special form and open formulae in which quantifiers are restricted to certain neighborhoods. 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 [19].

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 (Hanf-locality). (see [12, 7, 14]) *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 $\vec{a} \in A^m, \vec{b} \in B^m$,*

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

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 (Hanf, Fagin-Stockmeyer-Vardi). *Every FO sentence Φ defines a Hanf-local Boolean query Q with $\mathsf{lr}(Q) \leq 3^{\mathsf{qr}(\Phi)}$.* □

An extension to open formulae, although easily derivable from the proof of [7], was probably first explicitly stated in [14]: every FO formula $\varphi(\vec{x})$ defines a Hanf-local query. Better bounds of the order $O(2^{\mathsf{qr}(\varphi)})$ are also known for Hanf-locality [16, 19].

It was shown in [14] that every Hanf-local m-ary query, $m \geq 1$, is Gaifman-local.

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 [6, 17], unary quantifiers [13], and unary counters [2]. (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 [3, 23] – 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 [10]. 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 infinitary logic $\mathcal{L}_{\infty\omega}$. 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 (non-numerical) 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 $\exists ix\varphi$ is $\#x.\varphi \geq i$.

As this logic is too powerful (it expresses every property of finite structures), we restrict it by means of the *rank* of a formulae and terms, denoted by rk. It is defined as quantifier rank (that is, it is 0 for atomic formulae, $\mathsf{rk}(\bigvee_i \varphi_i) = \max_i \mathsf{rk}(\varphi_i), \mathsf{rk}(\neg\varphi) = \mathsf{rk}(\varphi), \mathsf{rk}(\exists x\varphi) = \mathsf{rk}(\exists ix\varphi) = \mathsf{rk}(\varphi) + 1$) but does not take into account quantification over $\mathbb{N}$: $\mathsf{rk}(\exists i\varphi) = \mathsf{rk}(\varphi)$. Furthermore, $\mathsf{rk}(\#\vec{x}.\psi) = \mathsf{rk}(\psi) + |\vec{x}|$.

Definition 3. (see [19]) *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 [19] that $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ formulae are closed under 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 natural numbers is available. Furthermore, counting terms can be eliminated in $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ without increasing the rank (that is, counting quantifiers suffice, although expressing properties with just counting quantifiers is often quite awkward).

Fact 3 *(see [15, 19]) 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 [19]; Hanf-locality was then shown in [15] using *bijective Ehrenfeuct-Fraïssé games* of [13].

Structures of bounded valence (degree) If $\mathcal{A} \in \mathrm{STRUCT}[\sigma]$, and R_i is of arity p_i, then $\mathit{degree}_j(R_i^{\mathcal{A}}, a)$ for $1 \leq j \leq p_i$ is the number of tuples $\vec{a}$ in $R_i^{\mathcal{A}}$ having a in the jth position. In the case of directed graphs, this gives us the usual notions of in- and out-degree. By $\mathit{deg\_set}(\mathcal{A})$ we mean the set of all degrees realized in $\mathcal{A}$. We use the notation $\mathrm{STRUCT}_k[\sigma]$ for $\{\mathcal{A} \in \mathrm{STRUCT}[\sigma] \mid \mathit{deg\_set}(\mathcal{A}) \subseteq \{0, 1, \ldots, k\}\}$.

Fact 4 *(see [14]) For any fixed k, a query Q on $\mathrm{STRUCT}_k[\sigma]$ 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 $\mathcal{C} \subseteq \mathrm{STRUCT}[\sigma]$ 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 $\mathcal{A} \models \alpha_i(\vec{a})$.

Fact 5 *(see [14]) For any fixed k, a query Q on $\mathrm{STRUCT}_k[\sigma]$ 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. Assume that we fix a semantics of predicates from θ. We then say that θ is Hanf-local if there exists a number d such that each predicate in θ defines a Hanf-local query Q with $\mathsf{hlr}(Q) \leq d$.

Theorem 6. *Let θ be Hanf-local. Then $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \theta$ expresses only Hanf-local queries.*

Proof sketch. Let d witness Hanf-locality of θ. We show that every $\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \theta$ formula of rank m defines a Hanf-local query Q with $\mathsf{hlr}(Q) \leq 3^m \cdot d + \frac{3^m - 1}{2}$ (for all instantiations of free variables of the second sort).

The proof 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 and infinitary connectives, as well as negation and quantification over the numerical sort are simple.

It 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 [19]) 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 $\mathsf{hlr}(Q) \leq 3r + 1$. For this, we need the following result from [14]: if $(\mathcal{A}, \vec{a}) \leftrightarrows_{3r+1} (\mathcal{B}, \vec{b})$, then 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 $c \in A$. We then fix $\vec{\imath}_0$ and assume $(\mathcal{A}, \vec{a}) \leftrightarrows_{3r+1} (\mathcal{B}, \vec{b})$. Then, for f as above, it is the case that $\mathcal{A} \models \varphi(c, \vec{a}, \vec{\imath})$ iff $\mathcal{B} \models \varphi(f(c), \vec{b}, \vec{\imath})$, due to Hanf-locality of φ, and thus $\mathcal{A} \models \psi(\vec{a}, \vec{\imath})$ 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. □

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

Corollary 2. *The logic $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ fails to capture Hanf-local properties over arbitrary finite structures.* □

Note that we only used I^k_ds 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 $\psi(\vec{x}, \vec{y}, \vec{z}) \equiv \mathbf{I}^k_d[m_1, \ldots, m_l](\vec{v}_1, \ldots, \vec{v}_l)(\varphi_1(\vec{v}_1, \vec{z}), \ldots, \varphi_l(\vec{v}_l, \vec{z}))$ 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 2. *Adding $\mathbf{I}^k_d[m_1, \ldots, m_l]$ 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.* □

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 [8] as well as in translations of various modal logics into fragments of FO [9, 25].

Definition 4. *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^i_k, $i \in \mathbb{N}$, 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^i_ks, 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*

$$\psi_1(\vec{x}, \vec{\imath}) \;\equiv\; \exists T^i_k \sqsubseteq S_d(\vec{y})\; \varphi(\vec{x}, \vec{\imath}) \qquad \text{and} \qquad \psi_2(\vec{x}, \vec{\imath}) \;\equiv\; \forall T^i_k \sqsubseteq S_d(\vec{y})\; \varphi(\vec{x}, \vec{\imath})$$

are formulae of rank $\mathsf{rk}(\varphi) + 1$. We say that the symbol T^i_k 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^i_ks 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 as follows. Given a σ-structure $\mathcal{A}$ and an interpretation $\mathcal{T}$ for all the symbols T^i_ks occurring freely in ψ, 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

$$\exists x \exists T \sqsubseteq S_r(x) \exists T' \sqsubseteq S_r(x) \left(\begin{array}{r} \forall y \in S_r(x)\ (T(y) \wedge \neg T'(y)) \vee (\neg T(y) \wedge T'(y)) \\ \wedge\, \forall z, v\ (T(z) \wedge E(z,v) \rightarrow \\ T'(v)) \wedge (T'(z) \wedge E(z,v) \rightarrow T(v)) \end{array} \right)$$

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 7. *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 sketch. We first show that queries definable in $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$ are Hanf-local. The same argument as in [19] shows that counting terms can be eliminated from $\mathcal{LSO}^r_{\infty\omega}(\mathbf{C})$ without increasing the rank of a formula. Suppose we are given a signature σ' disjoint from σ. If $\mathcal{A} \in \text{STRUCT}[\sigma]$, $\vec{a}$ 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 $d \geq r$,

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

if $\vec{D}$ interprets σ' over B, $\vec{a}$, $\vec{b}$ are of the same length, and the following three conditions hold: (1) $(\mathcal{A}, \vec{a}) \leftrightarrows_d (\mathcal{B}, \vec{b})$; (2) $adom(\vec{C}) \subseteq S^{\mathcal{A}}_r(\vec{a})$ and $adom(\vec{D}) \subseteq S^{\mathcal{B}}_r(\vec{b})$; and (3) there exists an isomorphism $h : N^{\mathcal{A}}_d(\vec{a}) \rightarrow N^{\mathcal{B}}_d(\vec{b})$ such that $h(\vec{C}) = \vec{D}$. The *if* direction is now implied by the lemma below, simply by taking σ' to be empty.

Lemma 1. *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 $(\mathcal{A}, \vec{a}, \vec{C}) \sim^r_d (\mathcal{B}, \vec{b}, \vec{D})$ 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 of the lemma is 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 $m = \mathsf{rk}_0(\varphi)$. The case requiring most work is that of counting quantifiers; that is, 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 $\vec{\imath}_0$, $(\mathcal{A}, \vec{a}, c, \vec{C}) \sim^r_d (\mathcal{B}, \vec{b}, e, \vec{D})$

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 $(\mathcal{A}, \vec{a}, \vec{C}) \sim^r_{9d+4} (\mathcal{B}, \vec{b}, \vec{D})$ 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 suffices to establish a bijection $f : A \to B$ such that for every c, $(\mathcal{A}, \vec{a}, c, \vec{C}) \sim^r_d (\mathcal{B}, \vec{b}, f(c), \vec{D})$ – then clearly the number of elements satisfying φ will be preserved. This proof in turn is based on the following combinatorial lemma: Assume that $(\mathcal{A}, \vec{a}) \leftrightarrows_{9d+4} (\mathcal{B}, \vec{b})$, and h is an arbitrary isomorphism $N^{\mathcal{A}}_{9d+4}(\vec{a}) \to N^{\mathcal{B}}_{9d+4}(\vec{b})$. 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 $c \in A$.

To prove the *only if* part, we show that with local second-order quantification, one can define local orderings on neighborhoods, and then the counting power of $\mathcal{L}^*_{\infty\omega}(\mathbf{C})$ allows one to code neighborhoods with numbers. The construction can be carried out in such a way that the entire multiset of isomorphism types of neighborhoods in a structure is coded by a formula whose rank is only determined by the radius of neighborhoods and the signature σ. Using this, one can express any Hanf-local query in $\mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$. □

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 3. *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 7, 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. *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 3. $\bigcup_{d>0}(\mathcal{L}^*_{\infty\omega}(\mathbf{C}) + \{I^k_d \mid k > 0\}) \subsetneq \mathcal{LSO}^*_{\infty\omega}(\mathbf{C})$. □

5 Characterizing Gaifman-local properties

We now turn to Gaifman's notion of locality, which states that a query Q is local with $\mathsf{lr}(Q) \leq r$ if $N^{\mathcal{A}}_r(\vec{a}_1) \cong N^{\mathcal{A}}_r(\vec{a}_2)$ 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. *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 $d \geq 0$, and Hanf-local queries Q_i, $i \in \mathbb{N}$, with $\mathsf{hlr}(Q_i) \leq d$, 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 8. *A query is Gaifman-local iff it is given by a Hanf-local definition by cases.*

Proof sketch. Assume that Q is given by a Hanf-local definition by cases. Let d be an upper bound on $\mathsf{hlr}(Q_i)$. Then Q is Gaifman-local and $\mathsf{lr}(Q) \leq 3d + 1$. Fix $\mathcal{A}$, and assume $\mathcal{A} \in \mathcal{C}_i$. Let $\vec{a}_1 \approx^{\mathcal{A}}_{3d+1} \vec{a}_2$. Then by [14], $(\mathcal{A}, \vec{a}_1) \leftrightarrows_d (\mathcal{A}, \vec{a}_2)$, 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 $\vec{a} \in Q(\mathcal{A})$. One then shows that each Q_i is Hanf-local, with $\mathsf{hlr}(Q_i) \leq d$, and for every $\mathcal{A}$ of type τ_i, $Q(\mathcal{A}) = Q_i(\mathcal{A})$. □

Unlike in Fact 5, 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 5 holds for arbitrary finite structures). We now show that the infinite number of cases is unavoidable. In fact, we show a stronger result.

Definition 6. *For $k > 0$, let* $\textsc{Local}_k$ *be the class of queries given by a Hanf-local definition by cases, where the number of cases is at most k. Let* $\textsc{Local}^*$ *be* $\bigcup_{k>0} \textsc{Local}_k$, *and* G_$\textsc{Local}$ *be the class of all Gaifman-local queries.*

Note that $\textsc{Local}_1$ is precisely the class of Hanf-local queries.

Theorem 9. *The hierarchy*

$$\textsc{Local}_1 \subset \textsc{Local}_2 \subset \ldots \subset \textsc{Local}^* \subset \text{G\_}\textsc{Local}$$

is strict.

Proof sketch. 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 of the classes of structures. More precisely, we define a class $\mathcal{C}_i^{l+1}$, $1 \leq i \leq l+1$ to be the class of graphs with the number of connected components being $i-1$ modulo $l+1$. Let Q_i^{l+1} be a FO-definable query returning the set of nodes reachable by a path of length $i-1$ from a node of indegree 0. Form the query Q that coincides with Q_i^{l+1} on $\mathcal{C}_i^{l+1}$. (Note that it is not FO, as the classes $\mathcal{C}_i^{l+1}$ are not FO-definable.) From Theorem 8, 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 $\mathcal{C}'_1, \ldots, \mathcal{C}'_l$ and Hanf-local queries Q'_i such that on $\mathcal{C}'_i$, Q coincides with Q'_i, $i = 1, \ldots, l$. Let $d = 1 + \max \mathsf{hlr}(Q'_i)$. 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}, $i = 1, \ldots, l+1$. By pigeonhole, there exists a class $\mathcal{C}'_k$ and $i \neq j, i, j \leq l+1$ such that $G_i^{l+1}, G_j^{l+1} \in \mathcal{C}'_k$. We then show that Q cannot give correct results on both G_i^{l+1} and G_j^{l+1}. The separation G_Local from Local^* is proved by a minor modification of the construction above. □

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. *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 arbitrary computations over small neighborhoods and 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.

References

1. D.A. Barrington, N. Immerman, H. Straubing. On uniformity within NC^1. *JCSS*, 41:274–306,1990.
2. M. Benedikt, H.J. Keisler. Expressive power of unary counters. *Proc. Int. Conf. on Database Theory (ICDT'97)*, Springer LNCS 1186, January 1997. *ICDT'97*, pages 291–305.
3. J. Cai, M. Fürer and N. Immerman. On optimal lower bound on the number of variables for graph identification. *Combinatorica*, 12 (1992), 389–410.
4. G. Dong, L. Libkin and L. Wong. Local properties of query languages. *Theoretical Computer Science*, to appear. Extended abstract in *ICDT'97*, pages 140–154.
5. H.-D. Ebbinghaus and J. Flum. *Finite Model Theory.* Springer Verlag, 1995.
6. K. Etessami. Counting quantifiers, successor relations, and logarithmic space, *JCSS*, 54 (1997), 400–411.
7. R. Fagin, L. Stockmeyer and M. Vardi, On monadic NP vs monadic co-NP, *Information and Computation*, 120 (1994), 78–92.
8. H. Gaifman. On local and non-local properties, *Proceedings of the Herbrand Symposium, Logic Colloquium '81*, North Holland, 1982.
9. E. Grädel. On the restraining power of guards. *J. Symb. Logic*, to appear.
10. E. Grädel and Y. Gurevich. Metafinite model theory. *Information and Computation* 140 (1998), 26–81.
11. M. Grohe and T. Schwentick. Locality of order-invariant first-order formulas. In *MFCS'98*, pages 437–445.
12. W. Hanf. Model-theoretic methods in the study of elementary logic. In J.W. Addison et al, eds, *The Theory of Models*, North Holland, 1965, pages 132–145.
13. L. Hella. Logical hierarchies in PTIME. *Information and Computation*, 129 (1996), 1–19.
14. L. Hella, L. Libkin and J. Nurmonen. Notions of locality and their logical characterizations over finite models. *J. Symb. Logic*, to appear. Extended abstract in *LICS'97*, pages 204–215 (paper by the 2nd author).
15. L. Hella, L. Libkin, J. Nurmonen and L. Wong. Logics with aggregate operators. In *LICS'99*, pages 35–44.
16. N. Immerman. *Descriptive Complexity.* Springer Verlag, 1999.
17. N. Immerman and E. Lander. Describing graphs: A first order approach to graph canonization. In *"Complexity Theory Retrospective"*, Springer Verlag, Berlin, 1990.
18. Ph. Kolaitis and J. Väänänen. Generalized quantifiers and pebble games on finite structures. *Annals of Pure and Applied Logic*, 74 (1995), 23–75.
19. L. Libkin. On counting logics and local properties. In *LICS'98*, pages 501–512.
20. L. Libkin. Logics capturing local properties. Bell Labs Technical Memo, 1999.
21. L. Libkin and L. Wong. Unary quantifiers, transitive closure, and relations of large degree. In *STACS'98*, Springer LNCS 1377, pages 183–193.
22. J. Nurmonen. On winning strategies with unary quantifiers. *J. Logic and Computation*, 6 (1996), 779–798.
23. M. Otto. *Bounded Variable Logics and Counting: A Study in Finite Models.* Springer Verlag, 1997.
24. T. Schwentick and K. Barthelmann. Local normal forms for first-order logic with applications to games and automata. In *STACS'98*, Springer LNCS 1377, 1998, pages 444–454.
25. M. Vardi. Why is monadic logic so robustly decidable? In *Proc. DIMACS Workshop on Descriptive Complexity and Finite Models*, AMS 1997.