Constant Delay Enumeration for FO Queries over Databases with Local Bounded Expansion

Luc Segoufin[1] **and Alexandre Vigny**[2]

1 **Inria, France; and**
ENS Cachan, Cachan, France
2 **Université Paris Diderot Paris 7, Paris, France; and**
ENS Cachan, Cachan, France

Abstract

We consider the evaluation of first-order queries over classes of databases with *local bounded expansion*. This class was introduced by Nešetřil and Ossona de Mendez and generalizes many well known classes of databases, such as bounded degree, bounded tree width or bounded expansion. It is known that over classes of databases with local bounded expansion, first-order sentences can be evaluated in pseudo-linear time (pseudo-linear time means that for all ϵ there exists an algorithm working in time $O(n^{1+\epsilon})$). Here, we investigate other scenarios, where queries are not sentences. We show that first-order queries can be enumerated with constant delay after a pseudo-linear preprocessing over any class of databases having locally bounded expansion. We also show that, in this context, counting the number of solutions can be done in pseudo-linear time.

1998 ACM Subject Classification H.2.4 [Database Management] Systems, Query Processing, H.2.3 [Database Management] Languages, Query Languages

Keywords and phrases Enumeration, First-Order queries, Local bounded expansion

Digital Object Identifier 10.4230/LIPIcs.ICDT.2017.20

1 Introduction

Query evaluation is a fundamental task in databases and a vast literature is devoted to the complexity of this problem. Given a database $\mathbf{D}$ and a query q the goal is to compute the set $q(\mathbf{D})$ of all solutions for q over $\mathbf{D}$. Unfortunately, the set $q(\mathbf{D})$ might be way bigger than the database itself as the number of solutions could be exponential in the arity of the query. It can therefore be unrealistic to compute all solutions, even for small queries. One could imagine many scenarios to overcome this situation. We could for instance only want to compute the number of solutions or just compute the k most relevant solutions relative to some ranking function.

We consider here the complexity of the enumeration of the set $q(\mathbf{D})$, i.e. generating one by one all the solutions for q over $\mathbf{D}$. In this context two parameters play an important role. The first one is the *preprocessing time*, i.e. the time it takes to produce the first solution. The second one is the *delay*, i.e. the maximum time between the output of any two consecutive solutions. An enumeration algorithm is then said to be *efficient* if these two parameters are small. For the delay, the best we can hope for is constant time: depending only on the query and independent from the size of the database. For the preprocessing time an ideal goal would be linear time: linear in the size of the database with a constant factor depending on the query. When both are achieved we say that the query can be enumerated with constant delay after linear preprocessing.

Constant delay enumeration after linear preprocessing cannot be achieved for all queries. However, for restricted classes of queries and databases several efficient enumeration algorithms have been obtained. This is the case for instance for first-order (FO) queries over databases with bounded degree [3, 11], monadic second-order (MSO) queries over databases with bounded tree-width [2, 13] and FO queries over databases with bounded expansion [12]. Bounded expansion is a large class of databases as it contains in particular all structures excluding at least one minor (planarity, bounded tree-width etc.) and all structures of bounded degree [16].

In some scenarios only pseudo-linear preprocessing time has been achieved. A query can be enumerated with constant delay after a pseudo-linear preprocessing time if for all ϵ there exists an enumeration procedure with constant delay and preprocessing time in $O(\|\mathbf{D}\|^{1+\epsilon})$. This is the case for FO queries over databases with low degree [4].

A special case of enumeration is when the query is boolean. In this case the preprocessing computes the answer to the query. In order to be able to enumerate queries of a given language efficiently, it is therefore necessary to be able to solve the boolean case efficiently.

It has been shown recently that boolean FO queries could be computed in pseudo-linear time over nowhere dense databases [9]. Nowhere dense is an important class of databases generalizing bounded expansion [16]. Amongst classes of databases closed under sub-databases, Nowhere dense is the largest possible class enjoying efficient evaluation for FO queries [14].

It's a major open problem to show that over nowhere dense databases the boolean case can be extended to a constant delay enumeration for FO queries of higher arities.

In this paper we make one step towards solving this problem, extending the bounded expansion result to databases having local bounded expansion. Local bounded expansion lies strictly between bounded expansion and nowhere dense. It requires that for all r the class of neighbors of radius r has bounded expansion. It contains for instance all databases having local bounded tree-width, or excluding locally a minor. It strictly extends bounded expansion as there exist classes of local bounded tree-width that do not have bounded expansion [7].

For FO queries over a class of databases with local bounded expansion we provide:

- an enumeration procedure with constant delay after pseudo-linear preprocessing,
- a pseudo-linear time algorithm counting the number of solutions.

Our proof for enumeration follows a classical scheme. Our first ingredient is Gaifman's theorem, decomposing a formula into local ones with distance constraints. In order to evaluate the local formulas we would need to compute local neighborhoods. However this would not be linear as each neighborhood may be of linear size and we have linearly many of them. Our second key ingredient is the result that one can compute in pseudo-linear time a representative "cover" of the database by means of neighborhoods [9]. Because these neighborhoods have bounded expansion we can use the bounded expansion case in order to evaluate the local formulas. It remains to take care of the distance constraints and this is the main technical contribution of this paper.

The paper is organized as follows. We start by giving a new proof of the boolean case in Section 5. We then extend it to constant delay enumeration in Section 6 and to counting in Section 7.

Related work. Our presentation for the model checking, Section 5, uses the same tricks that were used in [7] to lift the model checking from the bounded tree-width case to the local bounded tree-width case. The model checking results presented in Section 5 were already obtained in [5] with a very similar argument. We give the proofs again here for completeness and in order to fix the notations.

An algorithm for counting in linear time the number of solutions for FO queries over classes of databases with "nice" local bounded tree-width was presented in [6]. The restriction "nice" requires that the neighborhood cover can be computed in linear time and that one part of the cover intersects only a constant number of other parts. It is more restrictive than the one we use, given by [9], and is designed to make the counting easy with a simple exclusion/inclusion argument. This argument does not seem to extend to the cover we have and our algorithm for counting, presented in Section 7, is done by induction on the number of free variables.

In [10] a labeling scheme was presented for first-order queries over graphs with "nice" local bounded tree-width. If constant delay enumeration may be derived from the labeling scheme, this one is computed in polynomial time while we aim for pseudo-linear time. It is unclear whether this result can be generalized to classes of graphs with local bounded expansion using the tools we develop in this paper.

2 Preliminaries

For a positive integer k, $[k]$ denotes the set $\{1, \cdots, k\}$. Thereafter, ϵ will always denote an element of $\mathbb{R}^+$, p, r, s, i, j and k positive integers and f a function of $\mathbb{N} \to \mathbb{N}$.

Databases and First-Order queries. A relational signature σ is a tuple $(R_1, \ldots, R_s)$ where each R_i is a relational symbol of arity r_i. By database, we mean a finite structure over a relational signature σ, that is a tuple $\mathbf{D} = (D, R_1^{\mathbf{D}}, \ldots, R_s^{\mathbf{D}})$, where D, the domain of $\mathbf{D}$, is a finite set and for each i, $R_i^{\mathbf{D}}$ is a subset of D^{r_i}. If $\mathbf{D}$ is a database and $A \subseteq D$ a subset of its universe, we denote by $\mathbf{D}[A]$ the database given by the substructure of $\mathbf{D}$ induced by A. We fix a classical encoding of structures as input, see for example [1]. We denote by $\|\mathbf{D}\|$ the size of (the encoding of) $\mathbf{D}$. Without loss of generality we assume that the domain D comes with a linear order. If not, we arbitrarily choose one, for instance the one induced by the encoding of $\mathbf{D}$. This order induces a lexicographical order among the tuples over D.

A query is a first-order (FO) formula built from atomic formulas, "$x = y$" and $R_i(x_1, \ldots, x_{r_i})$, and closed under boolean combinations, $\wedge, \vee, \neg$, existential and universal quantifications, $\exists, \forall$. We write $q(\overline{x})$ if $\overline{x}$ are the free variables of q. The length of $\overline{x}$ is called the arity of the query. Queries of arity 0 are called sentences. The size of q is written $|q|$.

We write $\mathbf{D} \models q(\bar{a})$ to denote the fact that $\bar{a}$ is a solution for q over $\mathbf{D}$. We write $q(\mathbf{D})$ to denote the set of tuples $\overline{a}$ such that $\mathbf{D} \models q(\overline{a})$.

Given a database $\mathbf{D}$ and a sentence q, the problem of testing whether $\mathbf{D} \models q$ or not is called *the model checking problem*. It may be restricted to a class $\mathcal{C}$ of databases.

Model of computation and complexity. As usual when dealing with linear time, we use Random Access Machines (RAM) with addition and uniform cost measure as a model of computation.

All problems encountered in this paper have two inputs, a database $\mathbf{D}$ and a query q. However they play different roles as $\|\mathbf{D}\|$ is large while $|q|$ is small. We therefore consider the data complexity point of view. We say that a problem is *linear time* if it can be solved in time $O(\|\mathbf{D}\|)$. Here, and in the rest of the paper, the constants hidden behind the "big O" depend on q. We say that a problem is *pseudo-linear time* if, for all ϵ, it can be solved in time $O(\|\mathbf{D}\|^{1+\epsilon})$. In this case the constant factor also depends on ϵ. If a subroutine of a procedure depending on ϵ produces an output of size $O(\|\mathbf{D}\|^{\epsilon})$ we will then say that the output is *pseudo-constant*.

Neighborhoods and bounded expansion. Fix a database $\mathbf{D}$ of domain D. The *Gaifman graph* of $\mathbf{D}$ is the non-directed graph which set of vertices is D and which edges are the pairs $\{a, b\}$ such that a and b occur in a tuple of some relation of $\mathbf{D}$. Given two elements a and b of D, the *distance* between a and b is the length of a shortest path between a and b in the Gaifman graph of $\mathbf{D}$. The notion of distance extends to tuples in the usual way.

Given a positive integer r, the r-neighborhood of a in $\mathbf{D}$ is the substructure of $\mathbf{D}$ induced by the elements of D at distance at most r from a. It is denoted by $N_r^{\mathbf{D}}(a)$. Similarly we define $N_r^{\mathbf{D}}(\bar{a})$ as the union of the r-neighborhoods of the elements of $\bar{a}$.

Given a graph G with a linear order on its vertices, and two of its vertices a, b, we say that b is weakly r-accessible from a if there exists a path of length at most r between a and b such that b is smaller than all vertices of the path.

A class of graphs $\mathcal{C}$ has bounded expansion if for all r, there is a constant N_r, such that for all graphs G of $\mathcal{C}$, there is a linear order on the vertices of G, such that for all vertex a of G, the number of vertices weakly r-accessible from a is bounded by N_r [16]. This is a robust class of graphs with many equivalent definitions [16]. The precise definition will not be important for this paper as we will use this notion via its known algorithmic properties, in particular the fact that constant-delay enumeration algorithms exists for any class of databases with bounded expansion, see Section 3.

It is easy to see that if $\mathcal{C}$ has bounded expansion then the class of all subgraphs of all graphs of $\mathcal{C}$ also has bounded expansion.

A class $\mathcal{C}$ of graphs has *local bounded expansion* if, for any radius r, the class $\mathcal{C}_r$ of all subgraphs of all r-neighborhoods of all graphs in $\mathcal{C}$, has bounded expansion [16].

A class $\mathcal{C}$ of databases has (local) bounded expansion if the class of their Gaifman graphs has the same property.

Normal form for FO queries. We will make use of Gaifman Normal Form and Gaifman Locality Theorem for FO queries. This is rather classical in this context.

For all r there exists FO queries $\mathrm{dist}_r(y, \bar{x})$ expressing the fact that y is at distance at most r from $\bar{x}$. A query $q(\bar{x})$ is said to be r-local if all its quantifications are relative to elements at distance at most r from one of its free variables $\bar{x}$. This can be achieved using quantifications of the form $\exists y\ \mathrm{dist}_r(y, \bar{x}) \wedge \cdots$ and $\forall y\ \mathrm{dist}_r(y, \bar{x}) \rightarrow \cdots$.

It is known as Gaifman Normal Form that for any FO query there is an r such that the query is equivalent to a boolean combination of r-local queries and sentences of the form

$$\exists x_1 \dots x_k \Big(\bigwedge_{1 \leq i \leq j \leq k} \mathrm{dist}(x_i, x_j) > 2r \wedge \bigwedge_{1 \leq i \leq k} \psi(x_i) \Big),$$

where ψ is r-local. A proof can be found for example in [15].

For r-local queries $q(\bar{x})$ it is convenient to refine this normal form in order to know which of the free variables are close together. Any r-local query $q(\overline{x})$ is equivalent to a disjunction of the form:

$$\bigvee_{(\overline{x}_1;\dots;\overline{x}_p) \in P(\overline{x})} \alpha_1(\overline{x}_1) \wedge \dots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_1;\dots;\overline{x}_p), \tag{1}$$

where:

- $P(\overline{x})$ is the set of partitions of $\overline{x}$.
- $\alpha_i(\overline{x}_i)$ is r-local.
- $\tau_r(\overline{x}_1;\dots;\overline{x}_p)$ express the fact that the distance between $\overline{x}_i$ and $\overline{x}_j$ is bigger than $2r$ and that no refinement of P has this property. We will sometime refer to τ_r as a *distance type*.

Note that this implies that each α_i is $(r|\overline{x}_i|)$-local around any of its free variables, hence in particular the first one. Notice also that in (1) the disjunction is strict: no two outputs can satisfy two disjuncts. We will use this fact later to reduce our attention to a single disjunct.

Counting and enumeration. The *counting problem* is, given a database $\mathbf{D}$ and a query q, to compute the number of solutions to q over $\mathbf{D}$, i.e. the size of $q(\mathbf{D})$, noted $\#q(\mathbf{D})$.

We will now focus on the *enumeration problem*. An enumeration algorithm for a database $\mathbf{D}$ and a query q is divided into two consecutive phases:

- a preprocessing phase,
- an enumeration phase, outputting one by one and without repetitions the set $q(\mathbf{D})$.

The *preprocessing time* of the enumeration algorithm is the time taken by the preprocessing phase. Its *delay* is the maximum time between any two consecutive outputs.

One can view an enumeration algorithm as a compression algorithm computing a representation of $q(\mathbf{D})$ together with a streaming decompression algorithm. We aim for constant delay and pseudo-linear preprocessing time enumeration algorithms. By this we mean that for all ϵ, there is a preprocessing phase working in time $O(\|D\|^{1+\epsilon})$ and an enumeration phase with constant delay. Note that the multiplicative constants, for both the preprocessing phase and the delay, may depend on q and on ϵ.

All our enumeration procedures will output their tuples in lexicographical order. We will see that this is useful for queries in disjunctive normal form.

For the sake of readability, in the reminder of the paper, we only consider classes of graphs. All results and proofs can be easily adapted to the database case using standard techniques.

3 Main results

We will build on several known results over classes of databases with bounded expansion. The first is a linear time model checking algorithm for sentences:

▶ **Theorem 1** (Dvorak-Kral-Thomas [5]). *Let $\mathcal{C}$ be a class of graphs with bounded expansion. Then the model checking problem for FO queries over $\mathcal{C}$ can be solved is linear time.*

The second one solves the unary query case:

▶ **Theorem 2** (Dvorak-Kral-Thomas [5]). *Let $\mathcal{C}$ be a class of graphs with bounded expansion and let $q(x)$ be a query with one free variable. We can compute the set $q(G)$ in linear time.*

For queries with bigger arities, we cannot hope to evaluate their output in linear time anymore. A constant delay enumeration algorithm after linear preprocessing time has been obtained by Kazana and Segoufin in [12]. We present their result using a stronger statement than enumeration that will be useful for us later. Here $\geq$ is the lexicographical order on tuples over the domain. Recall that the constant factor depends on the query.

▶ **Theorem 3** (Kazana-Segoufin [12]). *Let $\mathcal{C}$ be a class of graphs with bounded expansion. Then there is an algorithm such that for all graph G in $\mathcal{C}$, and for any FO query q, after a preprocessing in linear time, on input any tuple $\overline{a}$, the algorithm computes in constant time the minimal tuple $\overline{b}$ such that:*

- $\overline{b} \geq \overline{a}$
- $G \models q(\overline{b})$

If there is no such tuple (i.e. $\overline{a}$ is bigger than all solutions), it outputs Null.

Our first result extends Theorem 3 to classes with local bounded expansion, replacing linear preprocessing time with pseudo-linear preprocessing time:

▶ **Theorem 4.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion. Then there is an algorithm such that for all graph G in $\mathcal{C}$, and for any FO query q, after a preprocessing in pseudo-linear time, on input any tuple $\overline{a}$, the algorithm computes in constant time the minimal tuple $\overline{b}$ such that:*

- $\overline{b} \geq \overline{a}$
- $G \models q(\overline{b})$

If there is no such tuple (i.e. $\overline{a}$ is bigger than all solutions), it outputs Null.

It immediately yields the constant delay enumeration after pseudo-linear preprocessing time.

▶ **Corollary 5.** *The enumeration of first-order query over class of graphs with local bounded expansion can be done with constant delay, after pseudo-linear preprocessing. Moreover the output tuples are given in lexicographiacal order.*

Our second result shows that counting the number of solutions can be done in pseudo-linear time.

▶ **Theorem 6.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion and $q(\overline{x})$ be a first-order query. Then for all graph G in $\mathcal{C}$, we can compute $\#q(G)$ in pseudo-linear time.*

Our proof works by induction on the arity of the query. It uses a partition of the database into representative neighborhoods that we describe next. It then combines this partition with the bounded expansion case.

4 Neighborhood covers and partitions

Because of the definition of local bounded expansion it is natural to examine the neighborhoods of our graphs. However, the sum of the sizes of all neighborhoods could be quadratic in the size of the input, which is too big as we aim for pseudo-linear time algorithms. To overcome this we selects some representative neighborhoods that cover the entire graph. The result presented here actually works for the more general notion of nowhere dense[1] graphs and is based on [9].

A (r, s)-*neighborhood cover* of a graph G is a set T of bags $U_1, \ldots, U_\omega$ such that:

- $\forall a \in G,\ \exists \lambda \leq \omega \quad \mathcal{N}_r^G(a) \subseteq U_\lambda$
- $\forall \lambda \leq \omega,\ \exists a \in G \quad U_\lambda \subseteq \mathcal{N}_s^G(a)$

The *size* of the cover T is the sum of the bag sizes: $\|T\| = \sum_{\lambda \leq \omega} \|U_\lambda\|$. Its *degree* is the number $\delta(T) := \max_{a \in G} |\{\lambda \leq \omega \mid a \in U_\lambda\}|$.

▶ **Theorem 7** (Grohe et al. [9])**.** *Let $\mathcal{C}$ be a nowhere-dense class of graphs. Then for all integer s and for all graph G in $\mathcal{C}$, we can compute in pseudo-linear time a $(s, 2s)$-neighborhood cover of G with a pseudo-constant degree. In particular the size of the neighborhood cover is pseudo-linear.*

[1] Nowhere dense requires that for all r, the number of weakly r-accessible nodes is pseudo-constant, instead of constant for bounded expansion.

Let A be a set of vertices of G. The *s-kernel* of A is the set $K_s(A) := \{a \in A \mid N_s^G(a) \subseteq A\}$.

We deduce from a $(s, 2s)$-neighborhood cover of G a partition of the vertices of G as follows:

$$P_\lambda := K_s^G(U_\lambda) \setminus \bigcup_{\mu<\lambda} K_s^G(U_\mu).$$

It follows from the definitions that the P_λ form a partition of the vertices of G. Moreover, modulo an extra linear preprocessing time, given an element a we have access in constant time to the unique λ such that $a \in P_\lambda$. This is a consequence of the following simple lemma.

▶ **Lemma 8.** *For all graph G, for all set A of vertices of G, and for all integer s, $K_s^G(A)$ is computable in time $O(s \cdot \|A\|)$.*

Proof. We prove the lemma by induction on s.

- If $s = 1$, let L be a list initialized empty. Then for each element a of A, we go through every neighbor of a. If we find one that is not in A, we add a to L and we go to the following element of A. At the end, we have $K_1^G(A) = A \setminus L$.
- If $s = i + 1$, from $K_{i+1}^G(A) = K_1^G(K_i^G(A))$ we get $K_s^G(A)$ is in time $O(s \cdot \|A\|)$. ◀

In the following sections, we will often say: compute $T = \{(U_1, P_1), \ldots, (U_\omega, P_\omega)\}$ that is the $(s, 2s)$-neighborhood cover paired with the s-kernel partition.

The previous observations can be synthesized in the following corollary.

▶ **Corollary 9.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion. Then for all graph G in $\mathcal{C}$ and for all integer s, we can compute in pseudo-linear time a $(s, 2s)$-neighborhood cover with pseudo-constant degree and the associated s-kernel partition.*

If a $(s, 2s)$-neighborhood cover can be computed efficiently on any nowhere dense class of graphs, a key property of the covers that works only for the local bounded expansion case is that all U_λ are in a class of graphs with bounded expansion. This is because each U_λ is included in the $2s$-neighborhood of some point and the later has bounded expansion by definition. We can therefore enumerate any FO query on each U_λ using Theorem 3 in time $O(\|U_\lambda\|)$, for a total time $O(\sum_{\lambda \leq \omega} \|U_\lambda\|)$, that is pseudo-linear. We will use this property implicitly in the rest of the paper. Note that this does not solve the general case as some solutions may have parts in different U_λ.

5 Model-Checking

Since every class of graphs with local bounded expansion is nowhere dense, we already know that the model checking problem of first-order queries over graphs with local bounded expansion can be done in pseudo-linear time [9]. Before that, another proof specific to local bounded expansion was given in [5]. In order to illustrate the tools presented in the previous sections, we give a new proof of this result. As in [5], it is based on the ideas of Grohe and Frick for graphs with local bounded tree-width [7].

▶ **Theorem 10.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion. Given a graph G in $\mathcal{C}$ and a FO sentence q, we can decide in pseudo-linear time whether $G \models q$.*

The rest of this section is devoted to the proof of Theorem 10. Fix G in $\mathcal{C}$ and a FO sentence q. In view of Gaifman Normal Form, we can assume wlog, that q is of the form:

$$q := \exists x_1 \ldots x_k \Big(\bigwedge_{1 \leq i \leq j \leq k} \mathrm{dist}(x_i, x_j) > 2r \wedge \bigwedge_{1 \leq i \leq k} \psi(x_i) \Big),$$

where ψ is r-local for some r.

Our strategy is as follows: we will first compute the set of nodes satisfying ψ and then test whether k of them are far apart from each other. The next lemma takes care of the first step.

▶ **Lemma 11.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion. For all graph G in $\mathcal{C}$, for all integer r, and for all unary and r-local FO query ψ, we can compute $\psi(G)$ in pseudo-linear time.*

Proof. Recall that $\mathcal{C}_s$ denotes the class of all subgraphs of s-neighborhoods of graphs from $\mathcal{C}$. Fix r, ψ a unary and r-local query and $G \in \mathcal{C}$.

We first compute $T = \{(U_1, P_1), \dots, (U_\omega, P_\omega)\}$, a $(2r, 4r)$-neighborhood cover paired with the $2r$-kernel partition. This can be done in pseudo linear time by Corollary 9. We can then view the P_λ as new unary predicates.

$\forall \lambda \leq \omega$, we set $\psi_\lambda(x) := \psi(x) \wedge P_\lambda(x)$. Because ψ is r-local, $\psi(G)$ is the disjoint union of all $\psi_\lambda(U_\lambda)$. By definition, $U_\lambda \in \mathcal{C}_{4r}$ which has bounded expansion. Consequently, it is possible to compute $\psi_\lambda(U_\lambda)$ in time $O(\|U_\lambda\|)$ by Theorem 2. Therefore, we are able to compute the set $\psi(G)$ in time $O\left(\sum_{\lambda=1}^{\omega} \|U_\lambda\|\right) = O(\|T\|)$ that is pseudo-linear in the size of G. ◀

Now we want to find k elements far apart in $\psi(G)$. We use a trick found in [7].

▶ **Lemma 12.** *Let $\mathcal{C}$ be a class of graphs with local bounded expansion. For all graph G in $\mathcal{C}$, for all integer r and k, and for all set A of vertices of G, we can decide in pseudo-linear time whether A contains a subset of k elements that are pairwise at distance more than $2r$.*

Proof. We proceed as follows:

We first compute $T = \{(U_1, P_1), \dots, (U_\omega, P_\omega)\}$, a $(2r, 4r)$-neighborhood cover paired with the $2r$-kernel partition as in Corollary 9.

Let L be a list, initialized as empty.

While A is not empty and $|L| < k$, we select (and remove) an element a in A.

If for all b in L we have: $(b \in P_\lambda \Rightarrow a \notin U_\lambda)$ then we add a in L. Notice that every b belongs to some P_λ, and hence $N_{2r}(b) \subseteq U_\lambda$. If furthermore $a \notin U_\lambda$ then a and b must be at distance more than $2r$.

At the end, we have tree different cases:

- 1^{st} case, $|L| = 0$. Then $A = \emptyset$.
- 2^{nd} case, $|L| = k$. Then we are done because all elements of L are far apart from each other by construction.
- 3^{rd} case, $|L| = m$, with $0 < m < k$. Let $L = \{b_1, \cdots, b_m\}$. Notice that $A \subseteq N^G_{4r}(b_1, \dots, b_m)$. We can see that $H := N^G_{4r}(b_1, \dots, b_m)$ is in $\mathcal{C}_{4rm}$. Therefore, from Theorem 1 it is possible to check in linear time if:

$$H \models \exists\, x_1, \dots, x_k \bigwedge_{1 \leq i \leq k} A(x_i) \;\wedge \bigwedge_{1 \leq i < j \leq k} \mathrm{dist}(x_i, x_j) > 2r.$$ ◀

Theorem 10 now easily follows from Lemma 11 and Lemma 12.

6 Enumeration

In this section we provide a constant delay enumeration procedure for FO queries over graphs with local bounded expansion. We actually prove a stronger result as stated in Theorem 4. Let G be a graph, q a FO query and $\overline{a}$ any tuple from G, not necessarily in $q(G)$. We denote

by $\text{FOLW}_q(\overline{a})$ the smallest tuple $\overline{b} \in q(G)$ such that $\overline{b}$ is bigger than $\overline{a}$ in the lexicographical order i.e. $\overline{b} \geq \overline{a}$. If there is no such $\overline{b}$, we say that $\text{FOLW}_q(\overline{a}) = \mathsf{Null}$.

In the rest of this section we show that for any class $\mathcal{C}$ with local bounded expansion, given $G \in \mathcal{C}$ and q, after a pseudo-linear time preprocessing, we can compute $\text{FOLW}_q(\overline{a})$ in constant time. Recall that this means that given ϵ there is a preprocessing algorithm working in time $O(\|G\|^{1+\epsilon})$ computing a structure that can then be used to compute $\text{FOLW}_q(\overline{a})$ from $\overline{a}$ in constant time. All constants depend on q and ϵ. This proves Theorem 4. We proceed by induction on the arity of the query.

▶ Remark 1. Assume q is $q_1 \vee q_2$. Then $\text{FOLW}_q(\overline{a})$ is the smallest tuple among $\text{FOLW}_{q_1}(\overline{a})$ and $\text{FOLW}_{q_2}(\overline{a})$. Hence if q is a disjunction of queries, it is enough to prove Theorem 4 on each of the disjunct to get the result for q.

Thanks to Gaifman Normal Form, we can assume that the query is a boolean combination of r-local formulas and sentences. By Theorem 10 the sentences can be precomputed during the preprocessing phase. We are then left with a r-local query (any boolean combination of r-local queries is a r-local query). Moreover, in view of Remark 1 and Gaifman Theorem for local queries, we can assume without loss of generality that our query q has the form:

$$q(\overline{x}) = \alpha_1(\overline{x}_1) \wedge \ldots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_1; \ldots; \overline{x}_p)$$

where the α_i and τ_r satisfy the conditions described in Section 2.

We start with some examples in order to illustrate the difficulty of the task. Assume that the query returns the pairs of blue-red nodes that are sufficiently far apart:

$$q(x, y) := \text{dist}(x, y) > 2r \wedge B(x) \wedge R(y).$$

Given a blue node a and a node b we can compute $\text{FOLW}_q(a, b)$ as follows.

During the preprocessing phase, thanks to Corollary 9, we compute in pseudo-linear time a $(2r, 4r)$-cover with its associated partition $\{(U_1, P_1), \ldots, (U_\omega, P_\omega)\}$. Given a the λ such that $a \in P_\lambda$ can then be obtained in constant time. Assume we could compute in pseudo-linear time a structure that, given λ and b, returns in constant time the smallest red node $c \geq b$ outside of U_λ. With this, from $a \in P_\lambda$ we have that $q(a, c)$ holds and therefore $\text{FOLW}_q(a, b) \leq c$. It remains to test whether there is also a node $c' \geq b$ within U_λ that is far from a. As we are within U_λ, we can invoke Theorem 3 and compute the smallest such c' in constant time after a preprocessing linear in $\|U_\lambda\|$. As we don't know a, and therefore λ, in advance, we perform that preprocessing for all λ, for a total time linear in $O(\Sigma_\lambda \|U_\lambda\|)$, which is pseudo linear. The minimum element among c and c' is then the desired $\text{FOLW}_q(a, b)$. Unfortunately we cannot afford to construct the structure returning c from λ and b because this is a function with two parameters that can potentially have a quadratic size. We will see in the proof (this is essentially Claim 13 and Claim 14 bellow) that we can compute in pseudo-linear time a subset of this function that is sufficient for our needs.

The situation is even worse for higher arities. To see this, assume the query is now

$$q(x_1, x_2, x_3) := B(x_1) \wedge Y(x_2) \wedge R(x_3) \wedge \bigwedge_{1 \leq i < j \leq 3} \text{dist}(x_i, x_j) > 2r.$$

Given a blue node a, a yellow node a' that are both far apart, and a node b, we are looking for $\text{FOLW}_q(aa', b)$.

Let's see what happens when extending the previous reasoning. Given a and a' we derive in constant time λ and λ' such that $a \in P_\lambda$ and $a' \in P_{\lambda'}$. As above we could get in constant time the smallest red node $c \geq b$ outside of U_λ. But if this node is certainly far from a it

might be close to a'. We can then imagine precomputing the smallest red node $c \geq b$ outside of $U_\lambda \cup U_{\lambda'}$, a ternary function. Again, we will see that we can compute in pseudo-linear time a subset of this function good enough for us. It remains to test whether there is also a node $c' \geq b$ within $U_\lambda \cup U_{\lambda'}$ that is far from a and a'. We could use again Theorem 3, but we would need a preprocessing linear in $\Sigma_{\lambda,\lambda'}(\|U_\lambda \cup U_{\lambda'}\|)$, which is unfortunately quadratic. To overcome this problem we introduce an intermediate bag V_λ between P_λ and U_λ together with a more complex algorithm based on the positions of a and a' in all the bags we have, that we will describe below.

We now turn to the formal details.

Base case. Assume q is unary. Because q is also r-local, by Lemma 11 we can compute $q(G)$ during the preprocessing phase. In order to compute $\mathrm{FOLW}_q(a)$ for all $a \in G$, we go through all vertices of G starting from the maximal one. For each vertex a, if $a \in q(G)$ then we set $\mathrm{FOLW}_q(a) = a$. If $a \notin q(G)$ and a is the maximal element, we set $\mathrm{FOLW}_q(a)$ to Null. In all remaining cases, we set $\mathrm{FOLW}_a(a)$ to $\mathrm{FOLW}_q(b)$, where b is the successor of a in the linear order on the vertices.

Inductive case. Assume now that $q(\bar{x}, y)$ is an r-local query of arity $k+1$. Let $q'(\bar{x})$ be the query $\exists y\ q(\bar{x}, y)$.

We claim that, modulo a pseudo-linear preprocessing, given a tuple $\bar{a}$ such that $G \models q'(\bar{a})$ and a vertex b, we can compute in constant time the smallest $b' \geq b$ such $G \models q(\overline{a}, b)$, outputting Null if no such b' exists.

Before proving the claim we show that it implies the constant time computation of FOLW_q. Let a_{min} be the minimal element of G. Let $\overline{a}b$ be a tuple. Let $\overline{a}' = \mathrm{FOLW}_{q'}(\overline{a})$. By induction, $\overline{a}'$ can be computed in constant-time. If $\overline{a}'$ is Null we output Null and we are done. If $\overline{a}' > \overline{a}$ then we apply the claim with $\overline{a}'$ and a_{min} and we are done. If $\overline{a}' = \overline{a}$ then we apply the claim with $\overline{a}$ and b. If b' is not Null, then $\overline{a}b'$ is the desired tuple. If $b' =$ Null then let $\overline{a}'$ be the successor of $\overline{a}$ in the lexicographical order and $\overline{a}'' = \mathrm{FOLW}_{q'}(\overline{a}')$. We apply again the claim for $\overline{a}''$ and a_{min} and we are done. All this clearly takes constant time.

In the rest of this section we prove the claim. Recall that $q(\bar{x}, y)$ is of the form:

$$q(\overline{x}, y) = \alpha_1(\overline{x}_1, y) \wedge \ldots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_1, y; \ldots; \overline{x}_p).$$

Let $\bar{w} = \overline{x}_2 \cup \ldots \cup \overline{x}_p$. We have:

$$q(\overline{x}, y) = q_1(\overline{x}_1, y) \wedge q_2(\bar{w}) \wedge \tau_r(\overline{x}_1, y; \ldots; \overline{x}_p).$$

We will distinguish two cases, depending whether $\overline{x}_1$ is empty or not. Let k be the arity of q.

Elements far away

We assume here that $\overline{x}_1$ is empty. By Lemma 11 we can precompute in pseudo-linear time the set L of nodes satisfying q_1. It remains to compute from $\overline{a}$, and b the smallest element of L that is at distance $2r$ from $\overline{a}$ and is greater than b.

We compute a $(4r, 8r)$-neighborhood cover and the associated $4r$-kernel partition according to Corollary 9. We then compute the $2r$-neighborhood V_λ of each P_λ. We now have $T := \{(P_1, V_1, U_1), \ldots, (P_\omega, V_\omega, U_\omega)\}$ such that $N^G_{2r}(P_\lambda) = V_\lambda$ and $N^G_{2r}(V_\lambda) \subseteq U_\lambda$.

We define for all vertex b and all set $I \subseteq \{1, \ldots, \omega\}$ such that $|I| \leq k$ the function.

$$\mathrm{NEXT}(b, I) = \min \left\{ b' \mid b' \geq b \wedge b' \notin \bigcup_{\lambda \in I} V_\lambda \wedge b' \in L \right\}$$

The domain of this function is too big (recall that ω is linear in $\|G\|$) so we cannot compute it. Fortunately, computing only a small part of it will be good enough for our needs. For each vertex b we define by induction the following set $SC(b)$ of elements $I \subseteq \{1, \dots, \omega\}$ with $|I| \leq k$:

- For all b in G and for all λ with $b \in V_\lambda$, we add $\{\lambda\}$ to $SC(b)$.
- For all b in G, for all I, and for all λ, if $|I| < k$ and $I \in SC(b)$ and $\text{NEXT}(b, I) \in V_\lambda$, then we add $\{I \cup \{\lambda\}\}$ to $SC(b)$.

Our aim is to compute all $\text{NEXT}(b, I)$ for all b and $I \in SC(b)$. We first show that it will be enough to compute in constant time $\text{NEXT}(b, I)$ for all b and I.

▶ **Claim 13.** *Given a vertex b, a set I, and $NEXT(c, J)$ for all vertices $c > b$ and sets $J \in SC(c)$, then we can compute $NEXT(b, I)$ in constant time.*

Proof.

- Case 1, $b \in L$ and $b \notin \bigcup_{\lambda \in I} V_\lambda$, then b is $\text{NEXT}(b, I)$.
- Case 2, $b \notin L$ or $b \in \bigcup_{\lambda \in I} V_\lambda$, then let c be the smallest element of L strictly bigger than b. If there is no such c then $\text{NEXT}(b, I) = \mathsf{Null}$, otherwise:
  - Case 2.1, $c \notin \bigcup_{\lambda \in I} V_\lambda$, then c is $\text{NEXT}(b, I)$.
  - Case 2.2, $c \in V_\lambda$ with $\lambda \in I$. Therefore $\{\lambda\} \in SC(c)$. Let J be a maximal (for inclusion) subset of I in $SC(c)$. Since $\{\lambda\} \in SC(c)$, we know that J is non empty. We claim that $\text{NEXT}(c, J) = \text{NEXT}(b, I)$. To see this, assume that $\text{NEXT}(c, J) \in V_\mu$ with $\mu \in I$ hence $|J| < k$, then by definition of $SC(c)$, $J \cup \{\mu\} \in SC(c)$ and J was not maximal. Moreover, by definition of $\text{NEXT}(c, J)$, every point between c and $\text{NEXT}(c, J)$ is either not in L or in one of the V we want to avoid. As all nodes between b and c are not in L, the claim follows. ◀

We now show that $SC(b)$ is small for all b and that we can compute all of $\text{NEXT}(b, I)$ for all b and $I \in SC(b)$.

▶ **Claim 14.** *For all integer b, $|SC(b)|$ is a pseudo-constant. Moreover, it is possible to compute all $NEXT(b, I)$ for all vertex b and set $I \in SC(b)$ in pseudo-linear time.*

Proof. We first prove that for all $b \in G$, $|SC(b)|$ is a pseudo-constant. Then we use Claim 13 in order to prove that we can compute these pointers by induction.

- By $SC_l(b)$ we denote the subset of $SC(b)$ of elements I with $|I| \leq l$. Let d be the degree of our cover. We have that for all $b \in G$, $|SC_1(b)| = d$. For the same reason, we have that $|SC_{l+1}(b)| = O(d \cdot |SC_l(b)|)$. Therefore, we have that for all $b \in G$, $|SC(b)| = |SC_k(b)| \leq O(d^k)$. Since d is pseudo-constant, $|SC(b)|$ is also pseudo-constant.
- We compute the pointers for b from b_{max} to b_{min} downwards, respectively the biggest and the smallest element of G. Given a b in G, assume we have computed $\text{NEXT}(c, J)$ for all $c > b$ and $J \in SC(c)$. We then compute $\text{NEXT}(b, I)$ for $I \in SC(b)$ using Claim 13. Here, every pointer was computed in constant time. Since there is only a pseudo-linear number of them, the time required to compute them all is pseudo-linear. ◀

With these two claims, we are now ready to conclude the case where $\bar{x}_1$ is empty. The preprocessing phase consists of the following steps:

- 1^{st} step: compute a $(4r, 8r)$-neighborhood cover and the associated $4r$-kernel partition according to Corollary 9.
- 2^{nd} step: compute the $2r$-neighborhood V_λ of each P_λ. We now have $T := \{(P_1, V_1, U_1), \dots, (P_\omega, V_\omega, U_\omega)\}$ such that $N^G_{2r}(P_\lambda) = V_\lambda$ and $N^G_{2r}(V_\lambda) \subseteq U_\lambda$.

- 3^{rd} step: compute $L := q_2(G)$. This can be done in pseudo-linear time by Lemma 11.
- 4^{th} step: compute $\text{NEXT}(b, I)$ for all b and $I \in SC(b)$. This can be done in pseudo-linear time by Claim 14.
- 5^{th} step: $\forall 1 \leq l \leq k,\ \forall \lambda \leq \omega$, perform on U_λ the preprocessing phase for the formula:

$$\varphi_{\lambda,l}(x_1, \ldots, x_l, y) := V_\lambda(y) \wedge L(y) \wedge \bigwedge_{i \leq l} \text{dist}(x_i, y) > 2r.$$

This can be done in time $O(\|U_\lambda\|)$ by Theorem 3 because $U_\lambda \in \mathcal{C}_{8r}$.
This is the end of the preprocessing. The total time needed is pseudo-linear.

Now we are given $(\overline{a}, a_{k+1})$ a tuple of elements of G, such that $G \models \exists y\ q(\overline{a}, y)$.

- Let $\lambda_1, \ldots, \lambda_k$ be such that $a_i \in P_{\lambda_i}$ and $I := \{\lambda_1, \ldots, \lambda_k\}$.
- Let $b_0 = \text{NEXT}(a_{k+1}, I)$.
- For all $1 \leq i \leq k$, let $\bar{c}_i$ be the elements of $\overline{a}$ that falls into U_{λ_i} and $m_i = |\bar{c}_i|$. Using Theorem 3 we compute in constant-time

$$b_i = \min\{y \mid y \in V_{\lambda_i} \wedge y \geq a_{k+1} \ \wedge\ \varphi_{\lambda_i, m_i}(\bar{c}_i, y)\}\,.$$

- We return the minimum element among the b_i, $0 \leq i \leq k$.

This is clearly constant time. To see that this is correct, let c be the correct answer.

If $c \in V_{\lambda_i}$ for some $i \in I$ then φ_{λ_i, m_i} gives us b_i that is the smallest $y \in V_{\lambda_i}$ satisfying q_2 and is at distance greater than $2r$ from the elements of $\overline{a}$ present in U_λ. Those that are not in V_{λ_i} must be at distance greater than $2r$ from b_i since $N^G_{2r}(V_{\lambda_i}) \subseteq U_{\lambda_i}$. Hence $c = b_i$.

Otherwise, i.e. $c \notin V_{\lambda_i}$ for all $i \in I$. Then the $\text{NEXT}(\,\cdot\,, I)$ pointers give us b_0 that is the smallest element satisfying q_2 that is not in one of the V_{λ_i}. Therefore $\text{dist}(b_0, \overline{a}) > 2r$. Hence $c = b_0$.

This concludes the first case.

Elements nearby

Assume now that $\overline{x}_1$ contains at least one variable, say x_1. Therefore, for all tuples $(\overline{a}, b)$ such that $G \models q(\overline{a}, b)$, we have that $\text{dist}(a_1, b) < 2kr$. This makes the second case much easier.

The preprocessing phase contains several steps.

- 1^{st} step: compute a $(4rk, 8rk)$-neighborhood cover and the associated $4rk$-kernel partition according to Corollary 9.
- 2^{nd} step: compute the $2rk$ neighborhood V_λ of each P_λ. We now have $T := \{(P_1, V_1, U_1), \ldots, (P_\omega, V_\omega, U_\omega)\}$ such that $N^G_{2rk}(P_\lambda) = V_\lambda$ and $N^G_{2rk}(V_\lambda) \subseteq U_\lambda$.
- 3^{rd} step: $\forall 1 \leq l < k,\ \forall \lambda \leq \omega$, perform the preprocessing phase on U_λ of the formula:

$$\varphi_{\lambda,l}(\bar{x}_1, y, x'_1, \ldots, x'_l) := V_\lambda(y) \ \wedge\ q_1(\bar{x}_1, y) \ \wedge \bigwedge_{i<l} \text{dist}(x'_i, y) > 2r.$$

This can be done in time $O(\|U_\lambda\|)$ by Theorem 3 because $U_\lambda \in \mathcal{C}_{8kr}$.
This is the end of the preprocessing. The total time needed is pseudo-linear.

Now we are given $(\overline{a}, a_{k+1})$ a tuple of elements of G, such that $G \models \exists y\ q(\overline{a}, y)$. Let $\overline{a}_1$ be the elements of $\overline{a}$ corresponding to $\bar{x}_1$ and a_1 the first of them.

- Let λ be such that $a_1 \in P_\lambda$.
- Let $\bar{c} := (a_j \in \bar{a} \mid a_j \in U_\lambda \wedge a_j \notin \overline{a}_1)$ and $m = |\bar{c}|$

- We return $b := \min\{b' \mid b' \in U_\lambda \wedge b' \geq a_{k+1} \ \wedge \ U_\lambda \models \varphi_{\lambda,m}(\bar{a}_1, b', \bar{c})\}$. This can be done in constant time by Theorem 3.

All this is done in constant time. It is correct because by assumption the answer b must be in V_λ, satisfy $q_1(\bar{a}_1, b)$ and be at distance greater than $2r$ from the elements of $\bar{w}$.

The elements of $\bar{w}$ that are not in U_λ must satisfy this last condition since $b \in V_\lambda$ and $N^G_{2rk}(V_\lambda)$
$\subseteq U_\lambda$. It remains to consider the elements of $\bar{w}$ that are in U_λ, that is $\bar{c}$. Then $\varphi_{\lambda,m}(\overline{a}_1, y, \bar{c})$ gives us exactly what we want.

This concludes the second case and therefore the proof.

Besides constant delay enumeration, Theorem 4 has another interesting immediate corollary. Modulo a pseudo-linear time preprocessing we can test, given a tuple $\overline{a}$ in constant time, whether it belong to $q(G)$ or not:

▶ **Corollary 15.** *Let $\mathcal{C}$ be a class of databases with local bounded expansion. Then for all graph G in $\mathcal{C}$, after a pseudo-linear time preprocessing, we can, given a tuple $\overline{a}$, decide in constant time whether it belongs to $q(G)$ or not.*

7 Counting

In this section we consider the counting problem which is to compute, given G and q, the size of $q(G)$, denoted by $\#q(G)$. We aim at computing $\#q(G)$ in time pseudo-linear.

▶ Remark 2. Assume q is $q_1 \vee q_2$ and that q_1 and q_2 have no common solution, i.e. the disjunction is strict. Then $\#q(G) = \#q_1(G) + \#q_2(G)$. Hence if q is a strict disjunction of queries, it is enough to prove Theorem 6 on each of the disjunct to get the result for q.

Again we will build on the bounded expansion case:

▶ **Theorem 16** (Kazana, Segoufin [12])**.** *Let $\mathcal{C}$ be a class of graphs with bounded expansion and $q(\overline{x})$ be a FO query. Then, for all graph G in $\mathcal{C}$, we can compute $\#q(G)$ in linear time.*

The rest of the section is dedicated to the proof of Theorem 6.

Thanks to Gaifman Normal Form, we can assume that the query is a boolean combination of r-local formulas and sentences. By Theorem 10 the sentences can be precomputed during the preprocessing phase. We are then left with a r-local query (any boolean combination of r-local queries is a r-local query). Moreover, in view of Remark 2 and Gaifman Theorem for local queries, we can assume without loss of generality that our query q has the form:

$$q(\overline{x}) = \alpha_1(\overline{x}_1) \wedge \ldots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_1; \ldots; \overline{x}_p)$$

where the α_i and τ_r satisfies the conditions described in Section 2.

The proof goes by induction on p, which is the number of connected components of the distance -type τ.

We first give a small example in order to give a hint of how the induction works.

Consider again the query returning the pairs of blue-red nodes that are far apart:

$$q(x, y) := \text{dist}(x, y) > 2r \wedge B(x) \wedge R(y).$$

In this case, there are two connected components. In order to count the number of solutions, we multiply the number of blue nodes by the number of red nodes and we subtract

from the result the number of blue-red nodes that are at distance smaller than $2r$. Those three numbers correspond to the number of solutions of three queries with only one connected component in their distance type each, hence we can proceed by induction. This is essentially what we do in the general case.

We now give the details. We start with the base case followed by the inductive case.

Let $G \in \mathcal{C}$, and $q(\overline{x}) = \alpha_1(\overline{x}_1) \wedge \ldots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_1; \ldots; \overline{x}_p)$

If $p = 1$. In this case, if $\overline{a} \in q(G)$, then $N_r^G(\overline{a}) \subseteq N_{2rk}^G(a_1)$.

- 1^{st} step: compute a $(2rk, 4rk)$-neighborhood cover and the associated $2rk$-kernel partition according to Corollary 9. We now have:
  $T := \{(P_1, U_1), \ldots, (P_\omega, U_\omega)\}$ such that $N_{2rk}^G(P_\lambda) \subseteq U_\lambda$.
- 2^{nd} step: for all $\lambda \leq \omega$, let $\varphi_\lambda(\overline{x}) := q(\overline{x}) \wedge x_1 \in P_\lambda$.
  We have that $q(G) = \bigcup_{1 \leq i \leq \omega} \varphi_\lambda(U_\lambda)$. Moreover, the union is disjoint. Therefore:

$$\#q(G) = \sum_{i=1}^{\omega} \#\varphi_\lambda(U_\lambda)\,.$$

  Since for all λ, $U_\lambda \in \mathcal{C}_{4kr}$, we can compute $\#\varphi_\lambda(U_\lambda)$ in time $\|U_\lambda\|$ using Theorem 16. Therefore, we can compute $\#q(G)$ in total time $O(\sum_{i=1}^{\omega}(\|U_\lambda\|)) = O(\|T\|)$, that is pseudo-linear in the size of G.

If $p > 1$. Let $\bar{w} = (\overline{x}_2, \ldots, \overline{x}_p)$. Consider the following three queries:

$$\begin{aligned} q_1(\overline{x}_1) &:= \alpha_1(\overline{x}_1) \wedge \tau_r(\overline{x}_1), \\ q_2(\bar{w}) &:= \alpha_2(\overline{x}_2) \wedge \ldots \wedge \alpha_p(\overline{x}_p) \wedge \tau_r(\overline{x}_2; \ldots; \overline{x}_p), \\ q_3(\bar{x}, \bar{w}) &:= q_1(\overline{x}_1) \wedge q_2(\bar{w}) \wedge \mathrm{dist}(\overline{x}_1, \bar{w}) \leq 2r. \end{aligned}$$

We have that

$$G \models q(\overline{a}\overline{b}) \iff q_1(\overline{a}) \wedge q_2(\overline{b}) \wedge \mathrm{dist}(\overline{a}, \overline{b}) > 2r,$$

hence

$$q(G) = q_1(G) \times q_2(G) \setminus \{\overline{a}, \overline{b} \in G \mid q_1(\overline{a}) \wedge q_2(\overline{b}) \wedge \mathrm{dist}(\overline{a}, \overline{b}) \leq 2r\},$$

which is

$$q(G) = q_1(G) \times q_2(G) \setminus q_3(G).$$

Since

$$q_3(G) \subseteq q_1(G) \times q_2(G),$$

it follows that

$$\#q(G) = \#q_1(G) \cdot \#q_2(G) - \#q_3(G).$$

It is easy to see that both q_1 and q_2 have less than p connected components in their distance type. Therefore, by the induction assumption we can compute $\#q_1(G)$ and $\#q_2(G)$ in pseudo linear time. We now have to compute $\#q_3(G)$.

We say that $(\overline{x}'_1; \ldots; \overline{x}'_{p'}) \in \Pi(\overline{x}_1; \ldots; \overline{x}_p)$ if and only if:

- $(\overline{x}'_1, \dots, \overline{x}'_{p'})$ is a partition of $\overline{x}$ with $p' < p$,
- $\overline{x}_1 \subsetneq \overline{x}'_1$,
- $\forall 1 < j \leq p'$, there is a $i > 1$ such that $\overline{x}'_j = \overline{x}_i$.

Basically, $\overline{x}'_1$ is the collapse of $\overline{x}_1$ and at least one of the $\overline{x}_i$. The other $\overline{x}_i$ remain unaltered.

Given $(\overline{x}'_1; \dots; \overline{x}'_{p'})$, we define:
$\alpha'_1(\overline{x}'_1) = \bigwedge_{i \in I} \alpha_i(\overline{x}_i)$ where $I := \{i \leq p \mid \overline{x}_i \subset \overline{x}'_1\}$,
$\alpha'_j(\overline{x}'_j) = \alpha_i(\overline{x}_i)$ where $\overline{x}_i = \overline{x}'_j \quad \forall 1 < j \leq p'$.
It follows from those definitions that:

$$q_3(\overline{x}) = \bigvee_{(\overline{x}'_1;\dots;\overline{x}'_{p'}) \in \Pi(\overline{x}_1;\dots;\overline{x}_p)} \alpha'_1(\overline{x}_1) \wedge \dots \wedge \alpha'_{p'}(\overline{x}'_{p'}) \wedge \tau_r(\overline{x}'_1; \dots; \overline{x}'_{p'}).$$

Moreover these disjunctions are strict. Therefore, with Remark 2:

$$\#q_3(G) = \sum_{(\overline{x}'_1;\dots;\overline{x}'_{p'}) \in \Pi(\overline{x}_1;\dots;\overline{x}_p)} \#\big(\alpha'_1(\overline{x}_1) \wedge \dots \wedge \alpha'_{p'}(\overline{x}'_{p'}) \wedge \tau_r(\overline{x}'_1; \dots; \overline{x}'_{p'})\big).$$

Since every query present here has less than p connected components in its distance type, we can by induction count the number of solutions for each of them in pseudo-linear time. There is only a constant number of queries involved in this sum, therefore $\#q_3(G)$ is computable in pseudo-linear time.

As $\#q_1(G)$ and $\#q_2(G)$ are already computed, we can compute $\#q(G) = \#q_1(G) \cdot \#q_2(G) - \#q_3(G)$.

The total time needed was pseudo-linear. This concludes the proof.

8 Conclusion

We have shown how to efficiently process first-order queries over classes of graphs with locally bounded expansion. We did not explicitly mention the constant factors. These are not very good. Even in the bounded expansion case the constant factor is a tower of exponentials which height depends on the size of the query. Moreover, an elementary constant factor is not reachable (unless FPT = AW[$*$]) even for unranked trees [8].

The results state the existence of an enumeration procedure for all ϵ. A uniform version of this statement would require that the procedure is computable from ϵ. It is indeed the case if the class of local bounded expansion is "effective", see [16] for the precise definition.

An improvement of our work will be to extend the results for the counting and enumeration problems to nowhere-dense structures. On those structures, the model checking can be done in pseudo-linear time [9]. There is therefore hope to find good algorithms for the other problems. However, this remains future work.

References

1 Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995. URL: `http://webdam.inria.fr/Alice/`.

2 Guillaume Bagan. MSO queries on tree decomposable structures are computable with linear delay. In *Computer Science Logic (CSL'06)*, 2006.

3 Arnaud Durand and Etienne Grandjean. First-order queries on structures of bounded degree are computable with constant delay. *ACM Trans. Comput. Log.*, 8(4), 2007. `doi:10.1145/1276920.1276923`.

4 Arnaud Durand, Nicole Schweikardt, and Luc Segoufin. Enumerating answers to first-order queries over databases of low degree. In *Symp. on Principles of Database Systems (PODS'14)*, 2014. doi:10.1145/2594538.2594539.

5 Zdenek Dvorak, Daniel Král, and Robin Thomas. Testing first-order properties for subclasses of sparse graphs. *J. ACM*, 60(5):36, 2013. doi:10.1145/2499483.

6 Markus Frick. Generalized model-checking over locally tree-decomposable classes. *Theory Comput. Syst.*, 37(1):157–191, 2004. doi:10.1007/s00224-003-1111-9.

7 Markus Frick and Martin Grohe. Deciding first-order properties of locally tree-decomposable structures. *J. ACM*, 48(6):1184–1206, 2001. doi:10.1145/504794.504798.

8 Markus Frick and Martin Grohe. The complexity of first-order and monadic second-order logic revisited. *Ann. Pure Appl. Logic*, 130(1-3):3–31, 2004. doi:10.1016/j.apal.2004.01.007.

9 Martin Grohe, Stephan Kreutzer, and Sebastian Siebertz. Deciding first-order properties of nowhere dense graphs. In *Symp. on Theory of Computing (STOC)*, 2014. doi:10.1145/2591796.2591851.

10 Mamadou Moustapha Kanté. *Graph Structurings: Some Algorithmic Applications. (Structurations de Graphes: Quelques Applications Algorithmiques)*. PhD thesis, University of Bordeaux, France, 2008. URL: https://tel.archives-ouvertes.fr/tel-00419301.

11 Wojciech Kazana and Luc Segoufin. First-order query evaluation on structures of bounded degree. *Logical Methods in Computer Science*, 7(2), 2011. doi:10.2168/LMCS-7(2:20)2011.

12 Wojciech Kazana and Luc Segoufin. Enumeration of first-order queries on classes of structures with bounded expansion. In *Symp. on Principles of Database Systems (PODS)*, 2013. doi:10.1145/2463664.2463667.

13 Wojciech Kazana and Luc Segoufin. Enumeration of monadic second-order queries on trees. *ACM Trans. Comput. Log.*, 14(4), 2013. doi:10.1145/2528928.

14 Stephan Kreutzer and Anuj Dawar. Parameterized complexity of first-order logic. *Electronic Colloquium on Computational Complexity (ECCC)*, 16:131, 2009. URL: http://eccc.hpi-web.de/report/2009/131.

15 Leonid Libkin. *Elements of Finite Model Theory*. Springer, 2004. doi:10.1007/978-3-662-07003-1.

16 Jaroslav Nešetřil and Patrice Ossona de Mendez. On nowhere dense graphs. *Eur. J. Comb.*, 32(4):600–617, 2011. doi:10.1016/j.ejc.2011.01.006.