Querying the Guarded Fragment

Vince Bárány * Georg Gottlob * Martin Otto †

Abstract

Evaluating a boolean conjunctive query q over a guarded first-order theory φ is equivalent to checking whether $\varphi \wedge \neg q$ is unsatisfiable. This problem is relevant to the areas of database theory and description logic. Since q may not be guarded, well known results about the decidability, complexity, and finite-model property of the guarded fragment do not obviously carry over to conjunctive query answering over guarded theories, and had been left open in general. By investigating finite guarded bisimilar covers of hypergraphs and relational structures, and by substantially generalising Rosati's finite chase, we prove for guarded theories φ and (unions of) conjunctive queries q that (i) $\varphi \models q$ iff $\varphi \models_{fin} q$, that is, iff q is true in each finite model of φ and (ii) determining whether $\varphi \models q$ is 2EXPTIME-complete. We further show the following results: (iii) the existence of polynomial-size conformal covers of arbitrary hypergraphs; (iv) a new proof of the finite model property of the clique-guarded fragment; (v) the small model property of the guarded fragment with optimal bounds; (vi) a polynomial-time solution to the canonisation problem modulo guarded bisimulation, which yields (vii) a capturing result for guarded-bisimulation-invariant PTIME.

1 Introduction

The guarded fragment. The *guarded fragment* of first-order logic (GF), defined through the relativisation of quantifiers by atomic formulas, was originally introduced by Andréka, van Benthem, and Németi [1], who proved that the satisfiability problem for GF is decidable. Grädel [12] proved that every satisfiable guarded first-order sentence has a finite model, i.e., that GF has the *finite model property (FMP)*. In the same paper, Grädel also proved that satisfiability of GF-sentences is complete for 2ExpTime, and is ExpTime-complete for sentences involving relations of bounded arity. GF has since been intensively studied and extended in various ways. For example, the *clique guarded fragment (*CGF*)* [13] properly extends GF but still enjoys the finite model property as shown by Hodkinson [15], see also [16] for a simpler proof. Guardedness has emerged as a main new paradigm for decidability and other benign properties such as the FMP, and has applications in various areas of computer science. While GF was originally introduced to embed and naturally extend propositional modal logics within first-order logic [1], it has various applications and was more recently shown to be relevant to description logics [11] and to database theory [22, 7]. Fragments of GF were recently studied for query-answering in such contexts, see e.g. [7, 9, 8, 22, 5, 6, 21]. The main problems studied in the present paper were motivated by such applications.

Main problems studied In the present paper we study the problem of *querying guarded theories* using conjunctive queries or unions of conjunctive queries. A boolean conjunctive query (BCQ) q consists of an existentially closed conjunction of atoms. A union of (boolean) conjunctive queries (UCQ) is a disjunction of BCQs. If φ is a guarded sentence (or, equivalently, a guarded theory), we say that a query q evaluates to true over φ, iff $\varphi \models q$. In this context, we considered the following non-trivial main questions:

Finite controllability: Is it true that for each GF-sentence φ and each UCQ q, $\varphi \models q \iff \varphi \models_{\mathrm{fin}} q$? Since the query q may not be guarded, the finite model property of the guarded fragment is not sufficient to answer this question positively. Rather, this question amounts to whether for each φ and q as above, whenever $\varphi \wedge \neg q$ is consistent, it also has a finite model. This is equivalent to the finite model property of the extended fragment GF^+ of GF, where universally quantified boolean combinations of negative atoms can be conjoined to guarded sentences. See [20] for another strengthening of the finite model property of GF. The concept of finite controllability was introduced by Rosati [22][1].

Size of finite models: How can we bound the sizes of finite models? In particular, in case $\varphi \not\models q$, how can we bound the size of the smallest finite models $\mathfrak{M}$ of φ for which $\mathfrak{M} \not\models q$? Note that any recursive bound on the size of such models $\mathfrak{M}$ immediately yields the decidability of query-answering. On the other hand, if φ is consistent and $\varphi \models q$, then the existence of a finite model $\mathfrak{M}$ such that

*{vince.barany,georg.gottlob}@comlab.ox.ac.uk, Oxford University Computing Laboratory

†otto@mathematik.tu-darmstadt.de, TU Darmstadt

[1] Rosati's definition is slightly stronger (see Proposition 1); the definition given here is, however, better suited for the full guarded fragment.

$\mathfrak{M} \models q$ follows trivially from the FMP of GF, because every model $\mathfrak{M}$ of φ is also a model of q. However, little was known about the size of the smallest finite models of a satisfiable guarded sentence φ. Grädel's finite-model construction in [12], in case of unbounded arities, first transforms φ into a doubly exponentially sized structure, which is then input to a transformation according to Herwig's theorem [14], requiring a further exponential blow-up in the worst case. This suggests a triple-exponential upper bound. Can we do better?

Size of hypergraph covers: It turns out (and will be made clear further below) that the above problems are closely related to bounds on the size of the smallest guarded bisimilar conformal hypergraph $\mathfrak{A}^+$ that covers a given hypergraph $\mathfrak{A}$, a problem of independent interest. Existence of such covers was established in [16]; their doubly exponential construction being the only known bound. Is it possible to find a better, possibly polynomial bound?

Decidability and complexity: Is UCQ-answering over guarded theories decidable, and if so, what is the complexity of deciding whether $\varphi \models q$ for a BCQ or a UCQ q over a guarded theory φ?

Canonisation and capturing: As a further problem of independent interest, which is closely related to the above questions, we study PTIME canonisation — the problem of providing a unique representative for each guarded bisimulation equivalence class of structures, to be computed in PTIME from any given member of that class. This has implications for capturing the guarded bisimulation invariant fragment of PTIME in the sense of descriptive complexity.

We provide answers to all these questions. Before summarising our results, let us briefly explain how the above questions relate to database theory and description logic. (A more detailed treatment will be contained in the full paper.)

Applications to databases and description logic. In the database area, *query answering under integrity constraints* plays an important role. In this context a relational database D, consisting of a finite set (conjunction) of ground atoms is given, and a set Σ of *integrity constraints* is specified on D. The database D does not necessarily satisfy Σ, and may thus be "incomplete". The problem of answering a BCQ q on D under Σ consists of determining whether $D \cup \Sigma \models q$, also written as $(D, \Sigma) \models q$. An important class of integrity constraints in this context are so-called *tuple-generating dependencies (TGDs)* [4]. Given a relational schema (i.e., signature) $\mathcal{R}$, a *tuple-generating dependency* (*TGD*) σ over $\mathcal{R}$ is a first-order formula of the form $\forall\overline{x}\forall\overline{y}\big(\Phi(\overline{x},\overline{y}) \rightarrow \exists\overline{z}\,\Psi(\overline{x},\overline{z})\big)$, where $\Phi(\overline{x},\overline{y})$ and $\Psi(\overline{x},\overline{z})$ are conjunctions of atoms over $\mathcal{R}$, called the *body* and the *head* of σ, respectively. It is well-known that database query-answering under TGDs is undecidable, see [3], even for very restricted cases [7]. For the relevant class of *guarded TGDs* [7], however, query-answering is decidable and actually 2EXPTIME-complete [7]. A guarded TGD is a TGD σ with an atom in its body that contains all universally quantified variables of σ. For example, the sentence

$$\forall M, N, D\ Emp(M, N, D) \wedge Manages(M, D) \rightarrow \exists E, N'\ Emp(E, N', D) \wedge Reportsto(E, M)$$

is a GTGD stating that if M is a manager named N belonging to and managing department D, then there must be at least one employee E having some name N' in department D reporting to M. In general, GTGDs are, strictly speaking, not guarded sentences, because their heads may be unguarded. However, by using "harmless" auxiliary predicates and splitting up GTD-heads into several rules, each set of GTGDs can be rewritten into a guarded sentence that is (for all relevant purposes) equivalent to the original set.

The class of *inclusion dependencies (IDs)* is a simple subclass of the class of GTGDs. An ID has the logical form $\forall\,\overline{x},\overline{y}\,(\alpha(\overline{x},\overline{y}) \rightarrow \exists\overline{z}\,\beta(\overline{x},\overline{z}))$, where α and β are single atoms. In [17] it was shown that query-answering under IDs is decidable and, more precisely, PSPACE-complete in the general case and NP-complete for bounded arities. One very important problem was left open in [17]: the finite controllability of IDs. Given that in the database world attention is limited to *finite databases*, a boolean query that would be false in infinite models of $D \cup \Sigma$ only, would still be finitely satisfied by $D \cup \Sigma$ and should be answered positively. Do such queries exist? This problem was solved by Rosati [22], who, by using a finite model generation procedure called *Finite Chase*, showed that IDs are finitely controllable. Rosati's result is actually formulated as follows:

Proposition 1 (Rosati [22])**.** *For every finite set of facts D and set $\mathcal{I}$ of IDs and for every N there exists a finite structure $\mathfrak{C}$ extending D and satisfying $\mathcal{I}$ and such that for every boolean conjunctive query q comprised of at most N atoms $\mathfrak{C} \models q$ iff $D, \mathcal{I} \models q$.*

Description logics are used for ontological reasoning in the Semantic Web and in other contexts. Useful description logics such as DL-Lite$_{core}$ and DL-Lite$_{\mathcal{R}}$ [6] are essentially based on IDs, and are thus finitely controllable. The class of GTDS and the yet more expressive class of *weakly guarded TGDS (WGTGDs)* have been introduced and studied in [7, 9] as powerful tools for data integration, data exchange [10], and ontological reasoning. Their finite controllability, however, was left as an open problem. Unfortunately, Rosati's Finite Chase cannot be directly applied to GTGDs and WTGTDs. However, it is easy to see (and will be detailed in the full paper) that the finite controllability of GTGDs and WGTGDs follows from the finite controllability of GF, which is the main result of the present paper.

Summary of results

Finite Controllability. That answering UCQs against guarded theories is finitely controllable was already implicit in the report [19], although not formulated in this terminology. The finite models constructed in [19] are of non-elementary size and do not yield meaningful complexity results. The following central result of our paper, derived by a completely new proof, yields a much better size bound.

Theorem 2. *For every* GF *sentence φ and every UCQ q, $\varphi \models q \iff \varphi \models_{\text{fin}} q$. More specifically, if $\varphi \wedge \neg q$ is satisfiable then it has a model of size $2^{|\varphi||q|^{\mathcal{O}(|q|)}}$, when the signature is taken to be fixed.*

To obtain the above result, we first establish new results on hypergraph covers, which are of independent interest.

Hypergraph Covers. We relate finite controllability to the concept of hypergraph covers. A *hypergraph cover* for a given hypergraph $\mathfrak{A}$ consists of a hypergraph $\mathfrak{B}$ together with a homomorphism $\pi\colon \mathfrak{B} \xrightarrow{\sim} \mathfrak{A}$ that induces a hypergraph bisimulation between $\mathfrak{B}$ and $\mathfrak{A}$. This notion naturally extends to relational structures $\mathfrak{A}, \mathfrak{B}$ on the basis of homomorphism induced guarded bisimulations. The following main technical result is used to derive most other results (for definitions of notions mentioned see Section 2).

Theorem 3 (Main Technical Result)**.** *Given $N \geq 2$ and a hypergraph (relational structure) $\mathfrak{A}$ one can construct an N-conformal hypergraph (structure) $\mathfrak{R}_N$ constituting a weakly N-acyclic hypergraph cover (guarded bisimilar cover) of $\mathfrak{A}$. In particular, $\mathfrak{R}_N$ is conformal whenever $N > w$, where w is the width of $\mathfrak{A}$; moreover, $|\mathfrak{R}_N| = |\mathfrak{A}|^{w^{\mathcal{O}(N)}}$ and, for fixed w and N, $\mathfrak{R}_N$ can be computed in polynomial time. We call $\mathfrak{R}_N$ the* Rosati cover *of $\mathfrak{A}$.*

Let us explain very informally the role of the Rosati covers in establishing Theorem 2. In Lemma 10 we observe that deciding $\varphi \models q$ for guarded φ and arbitrary queries q can be reduced to the equivalent question of $\varphi \models \chi_q$, where χ_q is a disjunction of *acyclic queries* stemming from the original query. It is more difficult to show that this reduction is equally valid over finite models, i.e. that $\varphi \models_{\text{fin}} q \iff \varphi \models_{\text{fin}} \chi_q$. In particular, that given a finite $\mathfrak{A} \models \varphi \wedge \neg\chi_q$ a finite model of $\varphi \wedge \neg q$ can too be found. The "unraveling" $\mathfrak{A}^*$ of $\mathfrak{A}$ constitutes a tree-like model of $\varphi \wedge \neg\chi_q$ and an acyclic cover of $\mathfrak{A}$. Thus, by virtue of acyclicity, $\mathfrak{A}^* \models \neg q$. However, $\mathfrak{A}^*$ is typically infinite. The challenge is to find a *finite* cover of $\mathfrak{A}$ retaining a "sufficient degree of acyclicity" so as not to render it a model of q. This is captured by the notion of *weakly N-acyclic covers* ensuring "faithful answers to queries of size at most N". Theorem 3 shows that such finite covers can be constructed.

Conformal covers. Hodkinson and Otto showed in [16] that all hypergraphs admit guarded bisimilar covers by conformal hypergraphs (for definitions, see Section 2). While the construction in [16] involves a doubly exponential blow-up in size, we here obtain a polynomial construction of conformal covers as a corollary to Theorem 3.

Corollary 4. *Every hypergraph $\mathfrak{A}$ of width w admits a conformal hypergraph cover of size $|\mathfrak{A}|^{w^{\mathcal{O}(w)}}$. For bounded width, we thus obtain polynomial size conformal covers.*

Translated into logic, Theorem 3 actually proves Rosati's Theorem in a more compact and more general form. It will emerge from our technical discussion that the finite controllability of GF can therefore be reduced to Theorem 3.

Finite model property of the clique-guarded fragment. As it happens, our construction used for Theorem 2 can also be applied to obtain finite models of any satisfiable clique-guarded formula. In fact, our construction yields more compact finite models. We thus obtain a new proof of the FMP for the clique-guarded fragment.

Small model property. Through our new method of finite-model construction, we are able to improve the bounds implicit in [12] for GF and the overhead for CGF implicit in [15, 16] on the size of the smallest finite model of a satisfiable (clique-)guarded sentence.

Theorem 5. *Every satisfiable formula of* CGF *(and thus of* GF*) has a finite model of size exponential in the length and doubly exponential in the width of the formula. Moreover, for every $k \geq 2$, the k-variable fragment of* CGF *(*GF*) has finite models of exponential size in the length of the formula.*

Complexity of query answering. In [12] Grädel proved that satisfiability of GF-sentences is complete for 2ExpTime, and ExpTime-complete in case of bounded arity. We prove that exactly the same bounds actually hold for answering BCQs and UCQs over guarded theories, which solves the initially posed complexity question about query-answering over guarded theories. The first step consists in reducing $\varphi \models q$ to $\varphi \models \chi_q$, as mentioned above. The formula χ_q may, however, be of exponential size. Our results then follow by analysing the complexity of checking the (un)satisfiability of the guarded theory $\varphi \wedge \neg\chi_q$. We also investigate the problem of query answering over models of a fixed guarded sentence, and provide a number of useful bounds. Our bounds for fixed sentences φ are not all tight and leave space for future research.

Canonisation and capturing. As a further consequence of Theorem 3, we find a polynomial solution to the canonisation problem for guarded bisimulation equivalence $\sim_{\text{g}}$. This allows us to capture the $\sim_{\text{g}}$-invariant fragment of PTime in the sense of descriptive complexity, i.e., to provide effective syntax with PTime model checking for the PTime queries closed under guarded bisimulation equivalence. Canonisation is achieved through inversion of the natural game invariant $I(\mathfrak{A})$ that uniquely characterises the

guarded bisimulation class, or the complete GF-theory, of a given structure $\mathfrak{A}$. A PTIME reconstruction of a model from the abstract specification of its equivalence class yields PTIME canonisation.

Theorem 6. *For every relational signature τ there exists a* PTIME *algorithm computing from a given invariant $I(\mathfrak{A})$ of an unspecified τ-structure $\mathfrak{A}$ a finite τ-structure $\mathrm{can}(\mathfrak{A})$ such that $I(\mathrm{can}(\mathfrak{A})) = I(\mathfrak{A})$; hence $can(\mathfrak{A}) \sim_{\mathrm{g}} \mathfrak{A}$ and $can(\mathfrak{A}') = can(\mathfrak{A})$ whenever $\mathfrak{A} \sim_{\mathrm{g}} \mathfrak{A}'$.*

Corollary 7. *The class of all those* PTIME *boolean queries that are invariant under guarded bisimulation,* PTIME$/\sim_{\mathrm{g}}$*, can be captured in the sense of descriptive complexity.*

Organisation. Section 2 defines the main concepts and introduces guarded bisimilar hypergraph covers as a main tool. It also states the above-mentioned Lemma 10. Section 3 presents the construction of the Rosati cover. From this and Lemma 10, the finite controllability of GF is proven in Section 4. Section 5 establishes our new complexity results. Section 6 deals with canonisation and capturing.

2 Hypergraphs and guarded fragments

We work with finite relational signatures possibly also admitting constants. Let us fix such a signature τ and let $\mathrm{width}(\tau)$ denote the the maximal arity of any of the predicate symbols in τ.

The guarded fragment of first order logic (GF), introduced by Andréka et al. [1], is the collection of first-order formulas with some syntactic restrictions in the quantification pattern, which is analogous to the relativised nature of modal logic. The set of $\mathsf{GF}(\tau)$ formulas is the smallest set

(i) containing all atomic τ-formulas and equalities;

(ii) closed under boolean connectives: $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$;

(iii) and such that whenever $\psi(\overline{x},\overline{y}) \in \mathsf{GF}(\tau)$ with all free variables indicated and $\alpha(\overline{x},\overline{y})$ is a τ-atom, or equality $x = y$, involving all free variables of ψ then the following are in $\mathsf{GF}(\tau)$ as well:

$$(\forall\, \overline{x}.\, \alpha)\, \psi := \forall \overline{x}(\alpha \rightarrow \psi) \;\text{ and }\; (\exists\, \overline{x}.\, \alpha)\, \psi := \exists \overline{x}(\alpha \wedge \psi).$$

In a τ-structure $\mathfrak{A}$ a set X of elements is said to be *guarded* if there is an atom $R^{\mathfrak{A}}(\overline{a})$ such that every member of X occurs in $\overline{a}$. A *maximal guarded set* is one not properly included in any other guarded set. A tuple $\overline{b}$ of elements is guarded if the set of its components is guarded.

An *atomic τ-type* $t(x_1,\ldots,x_n)$ is a maximal consistent set of τ-literals (atoms or negated atoms) whose constituent terms are among the variables $x_1,\ldots,x_n$ and the constants from τ. An atomic type $t(x_1,\ldots,x_n)$ determines, for every choice of indices $\bar{\imath} = (i_1,\ldots,i_k)$, its *restriction* to components $\bar{\imath}$, which is an atomic type in k variables $(x_{i_1},\ldots,x_{i_k})$ denoted $t|_{\bar{\imath}}$; conversely we say that t is an *extension* of $t|_{\bar{\imath}}$. In a τ-structure $\mathfrak{A}$ the atomic type $\mathrm{atp}_{\mathfrak{A}}(\overline{a})$ of a tuple $\overline{a}$ is the unique atomic type $t(\overline{x})$ such that $\mathfrak{A} \models t(\overline{a})$. One says that t is realised by $\overline{a}$ in $\mathfrak{A}$. Over a signature of r many relational symbols of maximal arity w and k constants there are $2^{\mathcal{O}(r(n+k)^w)}$ many atomic types in n variables. We identify each atomic type with the conjunction of its literals.

Guarded bisimulation $\sim_{\mathbf{g}}$ can be defined either in terms of the guarded bisimulation game, a variant of the Ehrenfeucht-Fraïssé game in which the set of pebbles must at any given time be guarded, or as a back-and-forth system of partial isomorphisms whose domain and image are both guarded. GF is preserved under guarded bisimulation:

$$\mathfrak{A} \sim_{\mathrm{g}} \mathfrak{B} \quad\Longrightarrow\quad \text{for all } \varphi \in \mathsf{GF}:\ \mathfrak{A} \models \varphi \;\Leftrightarrow\; \mathfrak{B} \models \varphi.$$

Given a relational structure $\mathfrak{A}$, its *guarded-bisimulation game graph*, denoted $G(\mathfrak{A})$, has as its vertices the maximal guarded tuples of $\mathfrak{A}$, each labelled by its atomic type. Two such tuples $\overline{a}$ and $\overline{b}$ are linked by an edge labelled by a partial bijection $\rho \subseteq \{1..k\} \times \{1..k\}$ whenever $a_i = b_j$ for all $(i,j) \in \rho$. Note that structures $\mathfrak{A}$ and $\mathfrak{B}$ are guarded bisimilar iff $G(\mathfrak{A})$ and $G(\mathfrak{B})$ are bisimilar (in the modal sense).

The *guarded bisimulation invariant* $I(\mathfrak{A})$ of $\mathfrak{A}$ is defined as the bisimulation quotient of $G(\mathfrak{A})$. Vertices of $I(\mathfrak{A})$ correspond to $\sim_{\mathrm{g}}$ classes of maximal guarded tuples of $\mathfrak{A}$, labelled by their atomic types. A ρ-labelled edge links vertices v and w if there are guarded tuples $\overline{a}$ and $\overline{b}$ in $\mathfrak{A}$ realising the $\sim_{\mathrm{g}}$-classes represented by v and by w, respectively, and such that $a_i = b_j$ for all $(i,j) \in \rho$.

While GF provides an important extension of the modal fragment, guarded quantification is too restrictive to express some basic temporal operators. To remedy this shortcoming various relaxations of the notion of guardedness and corresponding fragments have been introduced, chief among them the clique-guarded fragment.

The Clique Guarded Fragment (CGF) relaxes the constraints on guards α in GF to allow existentially quantified conjunctions of atoms as guards that guarantee that the tuple of free variables is clique-guarded. A set X of elements of a structure $\mathfrak{A}$ is *clique-guarded* if every pair of elements of X is guarded, equivalently, if X induces a clique in the Gaifman graph of $\mathfrak{A}$. A tuple $\overline{a}$ is clique-guarded whenever the set of its components is. Observe that while guarded sets are bounded in size by the width of the signature, there can be arbitrarily large clique-guarded sets whenever the width is at least 2. Recall that the *width* of a formula φ, $\mathrm{width}(\varphi)$ is the maximal number of free variables in any of its subformulas. In a clique-guarded formula φ the maximal size of a clique-guarded set quantified over is bounded by $\mathrm{width}(\varphi)$.

Scott normal form and satisfiability criterion Grädel's analysis of decidability for GF [12] uses the following Scott normal form corresponding to a relational Skolemisation.

Lemma 8. *To every (clique-)guarded τ-sentence φ one can associate a companion (clique-)guarded $\tau \cup \sigma$-sentence*

$$\psi = \bigwedge_j (\forall \overline{x}.\alpha_j)\, \vartheta_j(\overline{x}) \wedge \bigwedge_i (\forall \overline{x}.\beta_i)\, (\exists \overline{y}.\gamma_i)\, \psi_i(\overline{x},\overline{y})$$

such that $\psi \models \varphi$ and every $\mathfrak{A} \models \varphi$ has a $\tau \cup \sigma$-expansion $\mathfrak{B} \models \psi$. Here $|\sigma| \leq |\varphi|$, $\mathrm{width}(\psi) = \mathrm{width}(\varphi)$ and the ϑ_j, ψ_i are quantifier-free.

A guarded bisimulation game graph G or invariant I is said to satisfy the formula ψ in normal form if

(i) its vertices are labelled by guarded atomic types in the signature of ψ and that satisfy its universal conjuncts;

(ii) for each vertex v with label $t(\overline{x}\overline{z})$ and each conjunct $(\forall \overline{x}.\beta_i)\,(\exists \overline{y}.\gamma_i)\,\psi_i(\overline{x},\overline{y})$ such that $t(\overline{x}) \models \beta_i(\overline{x})$ there exists a vertex w labelled with some type $s(\overline{x}'\overline{y}) \models \psi_i(\overline{x}',\overline{y})$ such that $s|_{\overline{x}'} = t|_{\overline{x}}$ and v and w are linked by an edge labelled with the mapping $\rho : \overline{x} \to \overline{x}'$.

Proposition 9 (cf. [12, Lemma 3.4])**.** *Let ψ be the normal form of φ. Then φ is satisfiable iff there exists a guarded bisimulation invariant $\mathfrak{I}$ satisfying ψ and such that vertices of $\mathfrak{I}$ are labelled by distinct guarded atomic types.*

Hypergraphs, acyclicity and covers

A *hypergraph* is a pair $H = (V, S)$ with V its set of elements and $S \subseteq \mathcal{P}(V)$ a set of subsets of V, called hyperedges. For a set of hyperedges S, let $S{\downarrow}$ stand for the closure of S under subsets. A set X of elements of H is guarded if $X \in S{\downarrow}$. The Gaifman graph $\Gamma(H)$ of H is the undirected graph having vertex set V and, as edges, all guarded pairs of H. The maximal size of any hyperedge is referred to as the *width* of H. To every τ-structure $\mathfrak{A}$ one associates in a natural way a hypergraph $H[\mathfrak{A}]$ with V the universe of $\mathfrak{A}$ and S the collection of maximal guarded subsets of $\mathfrak{A}$. The width of $H[\mathfrak{A}]$ is then bounded by $\mathrm{width}(\tau)$. The Gaifman graph of $\mathfrak{A}$ is $\Gamma[\mathfrak{A}] := \Gamma(H[\mathfrak{A}])$.

A *homomorphism* $h\colon H \to H'$ between hypergraphs $H = (V, S)$ and $H' = (V', S')$ is a map from V to V' such that $h(s) \in S'{\downarrow}$ for all $s \in S$. Obviously any homomorphism $h\colon \mathfrak{A} \to \mathfrak{A}'$ between relational structures induces a hypergraph homomorphism from $H[\mathfrak{A}]$ to $H[\mathfrak{A}']$.

$G(H)$ and $I(H)$ are defined similarly for hypergraphs H, where instead of guarded bisimulation we use the natural notion of *hypergraph bisimulation* – it is safe to think of hypergraph bisimulation as of guarded bisimulation stripped of all atomic relational information. Vertices in the game graph are hyperedges and edges are labelled by partial bijections compatible with the actual overlaps.

A hypergraph H is *(N-)conformal* if every clique in $\Gamma(H)$ (of size at most N) is covered by a hyperedge of H. A structure $\mathfrak{A}$ is (N-)conformal whenever $H[\mathfrak{A}]$ is, i.e., if every k-clique ($k \leq N$) in its Gaifman graph is covered by a ground atom. Over conformal structures guarded quantification is as powerful as clique-guarded quantification.

A hypergraph H is *(N-)chordal* if all cycles in $\Gamma(H)$ of length greater than 3 (and at most N) have a chord in $\Gamma(H)$. An analogous notion for relational structures $\mathfrak{A}$ is similarly defined in terms of the Gaifman graph $\Gamma(H(\mathfrak{A}))$.

A hypergraph is *(N-)acyclic* if it is both (N-)chordal and (N-)conformal. For finite hypergraphs acyclicity is equivalent to tree decomposability. A finite hypergraph is *tree decomposable* if it can be reduced to the empty hypergraph by iteratively deleting some non-maximal hyperedge or some vertex contained in at most one hyperedge (cf. Graham's algorithm) [2]. We say that a relational structure $\mathfrak{A}$ is *guarded tree decomposable* if $\mathfrak{A}$ allows a tree decomposition in the sense of Robertson–Seymour with guarded bags. This is equivalent to $H[\mathfrak{A}]$ being tree decomposable, i.e. acyclic.

A guarded bisimilar cover $\pi\colon \mathfrak{B} \xrightarrow{\sim} \mathfrak{A}$ is an onto homomorphism $\pi\colon \mathfrak{B} \to \mathfrak{A}$ inducing a guarded bisimulation $\{(\bar{b}, \pi(\bar{b})) \mid \bar{b} \text{ guarded in } \mathfrak{B}\}$. It is *weakly N-conformal* if the image under π of any clique of size up to N in the Gaifman graph of $\mathfrak{B}$ is guarded in $\mathfrak{A}$; similarly it is *weakly N-chordal* if the image under π of every cycle of length at most N in the Gaifman graph of $\mathfrak{B}$ has a chordal decomposition (triangulation) in the Gaifman graph of $\mathfrak{A}$; and it is *weakly N-acyclic* if it is both weakly N-conformal and weakly N-chordal. Analogous notions of hypergraph covers are defined mutatis mutandis. Note that the restrictions of a cover homomorphism to hyperedges are necessarily bijections onto hyperedges.

A homomorphism $h\colon H \to H'$ into the hypergraph $H' = (V', S')$ is called *(guarded) tree decomposable* if there is some $S'' \subseteq S'{\downarrow}$ s.t. $h\colon H \to H''$ is a homomorphism into $H'' = (V', S'')$ and H'' is tree decomposable. This extends to relational structures in the usual manner. Every homomorphism into a (guarded) tree decomposable hypergraph (structure) is trivially (guarded) tree decomposable. Note that a hypergraph cover (guarded bisimilar cover) $\pi\colon \mathfrak{B} \xrightarrow{\sim} \mathfrak{A}$ is weakly N-acyclic iff for every homomorphism $h\colon \mathcal{Q} \to \mathfrak{B}$ from a hypergraph (structure) $\mathcal{Q}$ on at most N elements $\pi \circ h$ is (guarded) tree decomposable.

Conjunctive queries (CQ) are formulas of the form $\exists \overline{x} \bigwedge_i \alpha_i$, where the α_i are positive literals. A boolean conjunctive query (BCQ) is one with no free variables. A union of (boolean) conjunctive queries (UCQ) is a disjunction of BCQs. The size $|q|$ of a UCQ is the length of q as a formula.

To every BCQ $Q = \exists \overline{x} \bigwedge_i \alpha_i$ of signature τ one can associate the τ-structure $\mathcal{Q}$ having as universe the set of variables in $\overline{x}$ and atoms as prescribed by the α_i's of Q. Then $\mathfrak{A} \models Q$ iff there exists a homomorphism $h : \mathcal{Q} \to \mathfrak{A}$. We say that Q is *acyclic* if the associated structure $\mathcal{Q}$ is acyclic.

For each BCQ Q we define χ_Q as the disjunction of all *acyclic* BCQs T comprised of at most *three times as many atoms* as Q and such that $T \models Q$. For $q = \bigvee_i Q_i$ a UCQ we set $\chi_q = \bigvee_i \chi_{Q_i}$. It is obvious that $\chi_q \models q$ for every q.

Lemma 10. *Consider a UCQ $q = \bigvee_i Q_i$ in signature τ.*

(i) $\mathfrak{A} \models \chi_q$ *for all* $\mathfrak{A}$ *such that there is some guarded tree decomposable homomorphism* $h : \mathcal{Q}_i \to \mathfrak{A}$*;*

(ii) $\varphi \models q \leftrightarrow \varphi \models \chi_q$ *for all* $\varphi \in \mathsf{GF}$*;*

(iii) $|\chi_q| = |\tau|^{\mathcal{O}(|q|)}(|q|\mathrm{width}(\tau))^{\mathcal{O}(|q|\mathrm{width}(\tau))}$

3 The Rosati cover

Rosati proved Proposition 1 using a "finite chase" [22] procedure that safely reuses variables and results in very compact finite models. However, his proof of correctness of the finite chase with respect to conjunctive query answering is very intricate. We adapt the core idea of his model construction to give a more general guarded bisimilar cover construction for finite models, and a conceptually cleaner and simpler proof of faithfulness with respect to conjunctive queries of bounded size.

Theorem 3. *Given $N \geq 2$ and a (guarded) bisimulation invariant $\mathfrak{I} = I(\mathfrak{A})$ of an unspecified hypergraph (structure) $\mathfrak{A}$ one can construct hypergraphs (finite structures) $\mathfrak{R}_N$ and $\mathfrak{R}$ such that $I(\mathfrak{R}_N) = I(\mathfrak{R}) = \mathfrak{I}$, and $\mathfrak{R}_N$ is N-conformal and is a weakly N-acyclic (guarded) bisimilar cover of $\mathfrak{R}$. We have $|\mathfrak{R}_N| = |\mathfrak{I}|^{w^{\mathcal{O}(N)}}$ where w is the width of $\mathfrak{A}$ (also apparent from $\mathfrak{I}$) and, for fixed w and N, $\mathfrak{R}_N$ can be computed in polynomial time.*

It is not hard to see how this formulation entails the statement of Theorem 3 as given in the introduction. Observe that $I(\mathfrak{A}) = G(\mathfrak{A})$ can be enforced by introducing new predicates to distinguish each individual guarded tuple of $\mathfrak{A}$. Then $\mathfrak{R}_N$ is a weakly N-acyclic (guarded) bisimilar cover, the *Rosati cover of* $\mathfrak{A}$ itself. Then, for $N > \mathrm{width}(\mathfrak{A})$, $\mathfrak{R}_N$ is a conformal cover of $\mathfrak{A}$, hence Corollary 4.

We first define the Rosati cover of a given finite hypergraph (or relational structure), for fixed N. After preliminary observations much resembling some of Rosati's key lemmas we prove its two crucial properties: weak N-chordality and weak N-conformality of $\mathfrak{R}_N$ over $\mathfrak{A}$. Our $\mathfrak{R}_N$ turns out to be N-conformal but only *weakly* N-chordal over $\mathfrak{A}$. In fact, $\mathfrak{R}_N$ will be the top layer of a chain of covers of increasing degrees of weak chordality, similar to [19]; for weak N-chordality we then collect chords of cycles in $\pi_N(H)$ as images of hyperedges that are found in intermediate covering layers.

Definition of $\mathfrak{R}_N$

Let w be the width of $\mathfrak{A}$ (apparent from $\mathfrak{I} = I(\mathfrak{A})$), i.e. the maximal size of any of its hyperedges (guarded sets). We assume throughout that $w > 1$, since width 1 is trivial. For the rest of this section we also fix $m \geq N \geq 2$.

Let e be a vertex of $\mathfrak{I}$ labelled by some atomic type $\tau_e(x_1, \ldots, x_k)$. We associate to e and every $1 \leq i \leq k$, and every edge $\rho : d \to e$ of $\mathfrak{I}$, and every $0 \leq j < w^{m+1}$

- a constant symbol $c^j_{e,i}$, and
- a function symbol $f^j_{\rho,i}(z_1, \ldots, z_l)$ with $l = |\mathrm{dom}\rho|$, provided that x_i is not in the image of ρ.

Elements of the Rosati cover will be well-formed terms built with these constant and function symbols. As shorthand we write $\mathbf{f}^j_\rho(\bar{t})$ for $(f^j_{\rho,i}(\bar{t}))_{i \in e \setminus \mathrm{rng}(\rho)}$, and $\mathbf{c}^j_e$ for $(c^j_{e,i})_{1 \leq i \leq k}$, and for any a tuple $\bar{t} = (t_1, \ldots, t_l)$ we let $\{\bar{t}\}$ stand for $\{t_1, \ldots, t_l\}$. The *truncation* of a term t at depth κ, denoted $t/_\kappa$, is defined by the following recursive rules and is extended to sets and tuples of terms in the obvious way.

$$\begin{array}{lcll} c^j_{e,i}/_\kappa & = & c^j_{e,i} & \\ f^j_{\rho,i}(\bar{t})/_0 & = & c^j_{e,i} & (e \text{ the target of } \rho) \\ f^j_{\rho,i}(\bar{t})/_{\kappa+1} & = & f^j_{\rho,i}(\bar{t}/_\kappa) & \end{array}$$

We define for each $e \in \mathfrak{I}$ a set of hyperedges $\mathcal{H}^r_N(e)$ above e at height r, by mutual recursion:

$$\begin{array}{rcl} \mathcal{H}^0_N(e) & = & \{\, \{\mathbf{c}^j_e\} \mid j < w^{m+1} \,\} \\ \mathcal{H}^{r+1}_N(e) & = & \mathcal{H}^r_N(e) \cup \{\, \rho^j(h|_{\mathrm{dom}(\rho)}) \mid h \in \mathcal{H}^r_N(d), \\ & & \qquad \rho : d \to e,\ j < w^{m+1} \,\} \\ \mathcal{H}_N(e) & = & \bigcup_r \mathcal{H}^r_N(e) \\ \mathcal{H}_N & = & \bigcup_{e \in \mathfrak{I}} \mathcal{H}_N(e) \end{array}$$

where $\rho^j(\{\bar{t}\}) = \{\mathbf{f}^j_\rho(\bar{t}/_{N-1})\} \cup \{\bar{t}\}$ if j does not occur in $\bar{t}/_{N-1}$ and is undefined otherwise.

Observe that all terms appearing in any hyperedge in $\mathcal{H}_N$ have depth at most N and that the function symbol at the root of any subterm does not occur anywhere else within that subterm. By definition every $h \in \mathcal{H}_N$ is either of the form $\{\mathbf{c}^j_e\} \in \mathcal{H}^0_N(e)$ or $\{\mathbf{f}^j_\rho(\bar{t}/_{N-1})\} \cup \{\bar{t}\} \in \mathcal{H}^{r+1}_N(e)$ for some r and e the target of ρ. Crucially, $N \geq 2$ ensures that the latter partitioning of h is unique and we say that h is obtained by *ρ-extension* of some (not necessarily unique) hyperedge in $\mathcal{H}^r_N(d)$ for d the source of ρ. In particular the sets $\mathcal{H}(e)$ partition $\mathcal{H}$. Henceforth we often omit the subscript N writing $\mathcal{H}$, $\mathcal{H}(e)$, etc.

A hyperedge h is a *primary guard of* X if it is a guard of X, i.e. $X \subseteq h$, and is not the ρ-extension of some h' also guarding X.

Lemma 11. *For every guarded set X of terms there is an $e_X \in \mathfrak{I}$ such that all primary guards of X belong to $\mathcal{H}(e_X)$.*

Proof. If X is guarded by some $\{\mathbf{c}^j_e\}$, in which case $\{\mathbf{c}^j_e\}$ is the only primary guard of X, then set $e_X = e$.

If X is guarded by some $\{\mathbf{f}^j_\rho(\bar{t}/_{N-1})\} \cup \{\bar{t}\} \in \mathcal{H}^{r+1}(e)$ then either $X \subseteq \{\bar{t}\}$, in which case X is already guarded by some hyperedge in $\mathcal{H}^r(d)$, with $\rho : d \to e$ in $\mathfrak{I}$ so h is not a primary guard of X, or there is some $f^j_{\rho,i}(\bar{t}/_{N-1})$ in X, and we set $e_X = e$ to be the target of ρ.

Suppose a hyperedge $\{\mathbf{f}^{j'}_\sigma(\bar{s}/_{N-1})\} \cup \{\bar{s}\} \in \mathcal{H}^{r+1}(e')$ with $e' \neq e$ was also a primary guard of X. Then some

$f^{j'}_{\sigma,i'}(\overline{s}_{/N-1})$ would have to be in X. This, however, would imply that $f^{j}_{\rho,i}(\overline{t}_{/N-1})$ had to be among $\overline{s}$ and vice versa $f^{j'}_{\sigma,i'}(\overline{s}_{/N-1})$ among $\overline{t}$ contradicting the requirement that j and j' do not repeat in these terms, given that $N \geq 2$. □

It follows that we can lift all hyperedges of $\mathfrak{A}$ to $\mathfrak{R}^m_N$, not just the maximal ones. Let $\overline{\mathcal{H}_N}$ be comprised of the hyperedges in $\mathcal{H}_N$ together with sub-hyperedges $h' \subseteq h$ for each $h \in \mathcal{H}(e)$ precisely as specified by the type τ_e associated to $e \in \mathfrak{I}$. By Lemma 11, we may assume that h is a primary guard of h' since for every $\rho : d \to e$ the types $\tau_d|_{\mathrm{dom}\rho}$ and $\tau_e|_{\mathrm{rng}\rho}$ are identical. This definition is therefore sound in the sense that it does not depend on the choice of h.

Definition 1 (Rosati cover)**.** *We define $\mathfrak{R}^m_N$ as having universe $\bigcup \overline{\mathcal{H}_N}$ and hyperedges $\overline{\mathcal{H}_N}$, and set $\mathfrak{R}_N = \mathfrak{R}^N_N$.*

Using similar reasoning as in Lemma 11 one can verify that $\mathfrak{I}$ is indeed the guarded bisimulation invariant of $\mathfrak{R}_N$, i.e., that $\mathfrak{R}_N \sim_{\mathrm{g}} \mathfrak{A}$ for (any) $\mathfrak{A}$ (with $\mathfrak{I} = I(\mathfrak{A})$).

Lemma 12. *$I(\mathfrak{R}^m_N) = \mathfrak{I}$. In particular, for each $e \in \mathfrak{I}$ all hyperedges in $\mathcal{H}_N(e)$ realise the same guarded bisimulation type represented by e.*

Proof. Consider $h_0 \in \mathcal{H}(e_0)$ and $g_0 \in \mathcal{H}(d_0)$ such that $X = h_0 \cap g_0 \neq \emptyset$. As in Lemma 11 we can find primary guards $h_r, g_s \in \mathcal{H}(e_X)$ of X by tracing backward from h_0 and from g_0, respectively, through a sequence of extension

$$h_0 \xleftarrow{\rho_1} h_1 \xleftarrow{\rho_2} h_2 \cdots \xleftarrow{\rho_r} h_r \in \mathcal{H}(e_X) \text{ and}$$
$$g_0 \xleftarrow{\sigma_1} g_1 \xleftarrow{\sigma_2} g_2 \cdots \xleftarrow{\sigma_s} g_s \in \mathcal{H}(e_X)$$

ignoring the particular j-values. Let $h_i \in \mathcal{H}(e_i)$ for all $0 \leq i < r$ and $g_l \in \mathcal{H}(d_l)$ for all $0 \leq l < s$. Then in $\mathfrak{I}$ we have paths $e_0 \xleftarrow{\rho_1} e_1 \cdots e_{r-1} \xleftarrow{\rho_r} e_X \xrightarrow{\sigma_1} d_{s-1} \cdots d_1 \xrightarrow{\sigma_s} d_0$.

Given the nature of edges in (guarded) bisimulation game graphs $G(\mathfrak{A})$ as representing partial isomorphisms they are invertible and compositional in the sense that for each $v \xrightarrow{\rho} w$ there is $w \xrightarrow{\rho^{-1}} v$ and then for every $w \xrightarrow{\sigma} u$ there is also $v \xrightarrow{\sigma\circ\rho} u$ as long as $\sigma \circ \rho \neq \emptyset$. These properties are inherited by all (guarded) bisimulation invariants.

In this instance, this means that for any partial isomorphism $\emptyset \neq \pi \subseteq \sigma_1 \circ \cdots \circ \sigma_s \circ \rho_r^{-1} \circ \cdots \circ \rho_1^{-1}$ we have $e_0 \xrightarrow{\pi} d_0$ in $\mathfrak{I}$ and there is such a π for which $X = h_0|_{\mathrm{dom}\pi}$.

It follows that all moves made from any $h \in \mathcal{H}_N(e)$ to any $g \in \mathcal{H}_N(d)$ in the guarded bisimulation game on $\mathfrak{R}^m_N$ have corresponding edges from e to d in $\mathfrak{I}$. The converse of this being enforced by the very definition of $\mathfrak{R}^m_N$ we can establish that the (guarded) bisimulation invariant of $\mathfrak{R}^m_N$ is no other than $\mathfrak{I}$. □

Lemma 13. *$\mathcal{H}^r_N(e)/_{N-1} = \mathcal{H}^r_{N-1}(e)$ for all $m \geq N \geq 2$ and r. Truncation of terms at depth k thus acts, for all $m \geq N > k \geq 2$, as a homomorphic projection from $\mathfrak{R}^m_N$ onto $\mathfrak{R}^m_k$ inducing a guarded bisimulation. In other words, we have the following chain of covers:*

$$\mathfrak{R}^N_N \overset{\sim}{\to} \mathfrak{R}^N_{N-1} \overset{\sim}{\to} \cdots \mathfrak{R}^N_3 \overset{\sim}{\to} \mathfrak{R}^N_2$$

The size of $\mathfrak{R}_N$ can be bounded as follows. Let w be the width of $\mathfrak{I}$, assume that $w \geq 2$ and let $J = w^{m+1}$. Then there are $|\mathfrak{A}|^{\mathcal{O}(w)}J$ many constants $c^j_{e,i}$ and function symbols $f^j_{\rho,i}$ altogether and each term of depth at most N is built up from at most w^{N+1} many such symbols. For $m = N$, therefore, the total number of terms in $\mathfrak{R}_N$ is, as stated in Theorem 3, at most $(|\mathfrak{I}|^{\mathcal{O}(w)}w^{N+1})^{w^{N+1}} = |\mathfrak{I}|^{w^{\mathcal{O}(N)}}$.

Auxiliary notions

Consider a hyperedge $h = \{\mathbf{f}^j_\rho(\overline{t}_{/N-1})\} \cup \{\overline{t}\} \in \mathcal{H}^{r+1}_N(e)$. The elements of $\{\mathbf{f}^j_\rho(\overline{t}_{/N-1})\}$ will be referred to as *siblings*, and denoted as $f^j_{e,i}(\overline{t}_{/N-1}) \equiv f^j_{e,l}(\overline{t}_{/N-1})$. We will also say that these terms are *introduced in the hyperedge h* and that h *is a ρ-extension*. Furthermore, elements of $\{\overline{t}\}$ are said to be *predecessors* of those in $\{\mathbf{f}^j_\rho(\overline{t}_{/N-1})\}$, and denoted as $t_l \to f^j_{e,i}(\overline{t}_{/N-1})$, for l and i as appropriate. Constants covered by a hyperedge $\{\mathbf{c}^j_e\} \in \mathcal{H}^0_N(e)$ will also be regarded as siblings introduced in that hyperedge. Compare Lemmas 6 and 7 in the long version of [22] for the following.

Lemma 14. *(i) The relations $\equiv$, $\to$, and its inverse $\leftarrow$ form a partition of guarded pairs of $\mathfrak{R}^m_N$.*

(ii) $\equiv$ is an equivalence relation having guarded classes.

(iii) Whenever $t^0 \to t^1 \equiv t^2$ then also $t^0 \to t^2$.

(iv) In $\mathfrak{R}^m_N$ there are no directed $\to$-cycles of length $\leq N$.

Remark 15. *(i) Lemma 14 implies that the predecessor relation is transitive on any guarded set of terms.*

(ii) If h is a primary guard of $\overline{t}$ then the $\to$-maximal element of h must be among $\overline{t}$.

(iii) Assume $m \geq N \geq 3$. If $h \in \mathfrak{R}^m_N$ is a (primary) guard of $X \subseteq \mathfrak{R}^m_N$ then $h/_{N-1}$ is a (primary) guard of $X/_{N-1} \subseteq \mathfrak{R}^m_{N-1}$. In particular: $e_X = e_{X/_{N-1}}$ for any guarded set $X \subseteq \mathfrak{R}^m_N$.

N-conformality of $\mathfrak{R}_N$

Consider $3 \leq l \leq N$ and an l-clique $\{t^0, \ldots, t^{l-1}\}$ in $\mathfrak{R}_N$, i.e., such that all pairs $\{t^i, t^j\}$ are guarded. By Lemma 14 there are no predecessor-cycles in $\{t^0, \ldots, t^{l-1}\}$ but there is a term, w.l.o.g. t^0, such that every one of $t^1, \ldots, t^{l-1}$ is either a predecessor or a sibling of t^0.

Observe that the projection of any primary guard of t^0 to $\mathfrak{R}^m_{N-1}$ guards $\{t^0_{/N-1}, \ldots, t^{l-1}_{/N-1}\}$. This would already be sufficient to establish a weaker form of Theorem 3 still yielding Theorem 2 for GF. However, we wish to show that the entire l-clique is guarded already in $\mathfrak{R}_N$.

Proposition 16. *Assume that for some $2 \leq l \leq N$ there are $t^0, \ldots, t^{l-1}$ in $\mathfrak{R}_N$ such that all pairs $\{t^i, t^j\}$ are guarded. Then the entire set $\{t^0, \ldots, t^{l-1}\}$ is guarded in $\mathfrak{R}_N$.*

Proof. The case $l = 2$ being trivial we proceed by induction on l. By the preceding observation we may assume w.l.o.g. that each of $t^1, \ldots, t^{l-1}$ is either a predecessor or a sibling of t^0. By the induction hypothesis $X = \{t^1, \ldots, t^{l-1}\}$ is guarded. Let $g_0 \in \mathcal{H}(e_X)$ be a primary guard of X.

Consider first the case when $t^0 \equiv t^i$ for some $i \neq 0$. Then t^i is a $\rightarrow$-maximal element of X and as such is necessarily introduced in g_0. Then t^0, being a sibling of t^i, is also introduced in g_0, which therefore guards the entire clique.

Assume now that $t^i \rightarrow t^0$ for all $0 < i < l$ and let $h^{(0)}$ be some hyperedge in which t^0 is introduced. In this case t^0 takes the form $f^{j_0}_{\rho_0, i_0}(\overline{u}/_{N-1})$ for some $\rho_0 : d \rightarrow e$ and appropriate i_0 and $h^{(0)} = \rho_0^{j_0}(\overline{u}) = \{\mathbf{f}^{j_0}_{\rho_0}(\overline{u}/_{N-1})\} \cup \{\overline{u}\} \in \mathcal{H}(e)$ where $\overline{u} = h^{(1)}|_{\mathrm{dom}\rho}$ for some $h^{(1)} \in \mathcal{H}(d)$.

Note that each $t^i/_{N-1}$ is a subterm of t^0 at depth one, i.e., is among those in $\overline{u}/_{N-1}$. For each i, let u^i denote the component of $\overline{u}$ such that $u^i/_{N-1} = t^i/_{N-1}$ and let $Y = \{u^1, \ldots, u^{l-1}\}$. Repeating this kind of analysis in a backward tracing manner, we can find a chain of expansions

$$h^{(0)} \xleftarrow{\rho_0^{j_0}} h^{(1)} \xleftarrow{\rho_1^{j_1}} \cdots \xleftarrow{\rho_r^{j_r}} h^{(r+1)}$$

with $h^{(\lambda)} = \rho^{j_\lambda}(h^{(\lambda+1)}|_{\mathrm{dom}\rho_\lambda})$ a guard of Y, for all $\lambda \leq r$, until we reach some $h^{(r+1)} \in \mathcal{H}(e_Y)$ a primary guard of Y.

Let $\{\overline{v}\} = h^{(r+1)}$ and consider some $\{\overline{w}\} = g^{(0)} \in \mathcal{H}(e_X)$ a primary guard of X. Given that $N \geq 3$ and $X/_{N-1} = Y/_{N-1}$, according to Remark 15 (iii), we have $e_X = e_{X/_{N-1}} = e_{Y/_{N-1}} = e_Y$. By Lemma 12, $g^{(0)} \sim_{\mathrm{g}} h^{(r+1)}$. Moreover, by Remark 15 (ii), every $w_l \in g^{(0)}$ is either a sibling or a predecessor of some t^i and, similarly, every $v_l \in h^{(r+1)}$ is a sibling or a predecessor, respectively, of the same u^i. In other words the exact relationships within $g^{(0)}$ are mirrored in $h^{(r+1)}$ and therefore $h^{(r+1)}/_{N-1} = g^{(0)}/_{N-1}$. It follows that the extension sequence $\rho_r^{j_r}; \ldots; \rho_1^{j_1}; \rho_0^{j_0}$ is applicable – with the very same j_λ values – to $g^{(0)}$ producing an analogous derivation:

$$\begin{gathered}
\rho_0^{j_0}(g^{(r)}|_{\mathrm{dom}\rho_0}) = g^{(r+1)} \sim_{\mathrm{g}} h^{(0)} = \rho_0^{j_0}(h^{(1)}|_{\mathrm{dom}\rho_0}) \\
\rho_1^{j_1}(g^{(r-1)}|_{\mathrm{dom}\rho_1}) = g^{(r)} \sim_{\mathrm{g}} h^{(1)} = \rho_1^{j_1}(h^{(2)}|_{\mathrm{dom}\rho_1}) \\
\vdots \\
\rho_r^{j_r}(g^{(0)}|_{\mathrm{dom}\rho_r}) = g^{(1)} \sim_{\mathrm{g}} h^r = \rho_r^{j_r}(h^{(r+1)}|_{\mathrm{dom}\rho_r}) \\
g^{(0)} \sim_{\mathrm{g}} h^{(r+1)}
\end{gathered}$$

ending in some $g^{(r+1)} \in \mathcal{H}(e)$. Note that because each $h^{(r+1-\lambda)}$ is a guard of Y also each $g^{(\lambda)}$ is a guard of X. Moreover, a simple induction shows that $h^{(r+1-\lambda)}/_{N-1} = g^{(\lambda)}/_{N-1}$ for all $\lambda \leq r+1$. Therefore, $g^{(r+1)}$ introduces

$$f^{j_0}_{\rho_0, i_0}(g^{(r)}|_{\mathrm{dom}\rho_0}/_{N-1}) = f^{j_0}_{\rho_0, i_0}(h^{(1)}|_{\mathrm{dom}\rho_0}/_{N-1}) = t^0$$

and thus guards the entire clique. □

Weak N-chordality of $\mathfrak{R}_N$ over $\mathfrak{A}$

Proposition 17. *For every $3 \leq l \leq N \leq m$ and l-cycle $C = \{\{t^0, t^1\}, \{t^1, t^2\}, \ldots, \{t^i, t^{i+1}\}, \ldots, \{t^{l-1}, t^0\}\}$ in the Gaifman graph of $\mathfrak{R}^m_N$ the projection $C/_{N-l+3}$ of C into $\mathfrak{R}^m_{N-l+3}$ admits a guarded triangulation in $\mathfrak{R}^m_{N-l+3}$.*

Proof. By Proposition 16 all 3-cycles are guarded in $\mathfrak{R}^m_N$, proving the case of $N = 3$. We proceed by induction on N.

Given an l-cycle in $\mathfrak{R}^m_N$ as above, by Lemma 14 (i) we know that for every i either $t^i \equiv t^{i+1}$ or $t^i \rightarrow t^{i+1}$ or $t^{i+1} \rightarrow t^i$. According to Lemma 14 (iv) there are no predecessor cycles of length $\leq N$ in $\mathfrak{R}^m_N$, hence it cannot be the case that $t^i \rightarrow t^{i+1}$ for all i, nor that $t^{i+1} \rightarrow t^i$ for all i. Then for some i one of the following cases must hold:

- $t^{i-1} \equiv t^i \equiv t^{i+1}$: then, by Lemma 14 (ii), $\{t^{i-1}, t^i, t^{i+1}\}$ is guarded and $t^{i-1} \equiv t^{i+1}$;
- $t^{i-1} \rightarrow t^i \equiv t^{i+1}$ (or $t^{i+1} \rightarrow t^i \equiv t^{i-1}$): then, by Lemma 14 (iii), the triple $\{t^{i-1}, t^i, t^{i+1}\}$ is guarded, and $t^{i-1} \rightarrow t^{i+1}$ (or $t^{i+1} \rightarrow t^{i-1}$);
- $t^{i-1} \rightarrow t^i \leftarrow t^{i+1}$: then both $t^{i-1}_{/N-1}$ and $t^{i+1}_{/N-1}$ are maximal proper subterms of t^i, therefore, the projection $h/_{N-1}$ of any hyperedge h of $\mathfrak{R}^m_N$ in which t^i was introduced guards $\{t^{i-1}, t^i, t^{i+1}\}/_{N-1}$ in $\mathfrak{R}^m_{N-1}$.

In either case we have found some i, such that in $\mathfrak{R}^m_{N-1}$ $\{t^0, \ldots, t^{i-1}, t^{i+1}, \ldots, t^{l-1}\}/_{N-1}$ constitutes a cycle of length $l-1$ and $\{t^{i-1}, t^i, t^{i+1}\}/_{N-1}$ a guarded triangle. By the induction hypothesis the claim follows. □

4 Finite controllability and small models

Next we show that answering UCQs under GF is finitely controllable. Theorems 2 & 5 provide optimal upper bounds on the minimal size of finite models for various guarded fragments. Matching lower bounds are implicit in [12].

Theorem 2. *For every GF-sentence φ and every UCQ q, $\varphi \models q \iff \varphi \models_{\mathrm{fin}} q$. More specifically, if $\varphi \wedge \neg q$ is satisfiable then it has a model of size $2^{|\varphi||q|^{\mathcal{O}(|q|)}}$ when the signature is taken to be fixed.*

Proof. Recall the properties of χ_q from Lemma 10. We establish the claim by proving the following equivalences.

$$\varphi \models q \text{ iff } \varphi \models \chi_q \text{ iff } \varphi \models_{\mathrm{fin}} \chi_q \text{ iff } \varphi \models_{\mathrm{fin}} q$$

The first equivalence was proved in Lemma 10 (ii) and the second equivalence follows from the finite model property of the guarded fragment. Also $\varphi \models_{\mathrm{fin}} \chi_q \Rightarrow \varphi \models_{\mathrm{fin}} q$ is a trivial consequence of $\chi_q \models q$. It remains to be seen that $\varphi \not\models_{\mathrm{fin}} \chi_q$, or what is the same $\varphi \not\models \chi_q$, implies $\varphi \not\models_{\mathrm{fin}} q$.

So assume that $\varphi \wedge \neg\chi_q$ is satisfiable. Then, by Proposition 9, there is a certain invariant $\mathfrak{I}$ satisfying its Scott normal form ψ from Lemma 8. Applying Theorem 3 on input $\mathfrak{I}$

and with $N = |q|$ we obtain $\mathfrak{R}_2^N$ and $\mathfrak{R}_N^N$ a guarded bisimilar cover of $\mathfrak{R}_2^N$ with both structures satisfying $\varphi \wedge \neg\chi_q$. Furthermore, $\mathfrak{R}_N^N$ is a weakly N-acyclic cover of $\mathfrak{R}_2^N$, and hence from Lemma 10 (i) it follows that $\mathfrak{R}_N^N \models \varphi \wedge \neg q$. This concludes the proof of finite controllability.

According to Theorem 3, $|\mathfrak{R}_N| = |\mathfrak{I}|^{w^{\mathcal{O}(N)}}$, where w is the width of the signature. From Proposition 9 it follows that $|\mathfrak{I}|$ is bounded by the number of atomic types in the signature of ψ, which is of the order $2^{\mathcal{O}((r+|\varphi|+|\chi_q|)w^w)}$. Lemma 10 (iii) gives $|\chi_q| = r^{\mathcal{O}(|q|)}(|q|w)^{\mathcal{O}(|q|w)}$, which simplifies to $(r|q|)^{\mathcal{O}(|q|)}$ for w bounded.

Combining these figures yields the following bounds on $|\mathfrak{R}_N|$: (i) $2^{(|\varphi||q|)^{\mathcal{O}(|\varphi||q|)}}$ in general; (ii) $2^{(|\varphi||q|)^{\mathcal{O}(|q|)}}$ when w is bounded; (iii) $2^{|\varphi||q|^{\mathcal{O}(|q|)}}$ for a fixed signature. □

Corollary 18. *For a finite set F of τ-structures let $\mathcal{C}_F$ denote the class of those τ-structures not allowing a homomorphic image of any member of F. If a guarded sentence φ has a model in $\mathcal{C}_F$ then it also has one of size $2^{\mathcal{O}(|\varphi|)}$.*

Theorem 5. *Every satisfiable formula of* CGF *(and thus of* GF*) has a finite model of size exponential in the length and doubly exponential in the width of the formula. Moreover, for every $k \geq 2$, the k-variable fragment of* CGF *(*GF*) has finite models of exponential size in the length of the formula.*

Proof. If $\varphi \in$ GF with Scott normal form ψ is satisfiable then, by Proposition 9, there is an invariant $\mathfrak{I}$ satisfying ψ. The size of $\mathfrak{I}$ can be bounded by the number of atomic types in the signature of ψ, which is $2^{\mathcal{O}(r+|\varphi|)w^w}$. Applying Theorem 3 with input $\mathfrak{I}$ and parameter $N = 2$ yields a model of φ of size $|\mathfrak{I}|^{w^{\mathcal{O}(1)}} = 2^{|\varphi|w^{\mathcal{O}(w)}}$.

For CGF we appeal to [16, Section 3.3] for a simple polynomial reduction from (finite) satisfiability of CGF to (finite) satisfiability of GF. This reduction maps a sentence $\varphi \in \mathsf{CGF}[\tau]$ to $\varphi^* \in \mathsf{GF}[\tau \cup \{R_C\}]$, R_C a new relation of arity $w := \mathrm{width}(\varphi)$, such that (i) every model of φ can be expanded to a model of φ^* and (ii) φ^* implies φ *over conformal structures*. Hence, given a finite model $\mathfrak{A} \models \varphi^*$, its Rosati cover $\mathfrak{R}_N$ with parameter $N = w+1$, being conformal, is a model of φ. We have $|\mathfrak{R}_N| = |\mathfrak{A}|^{w^{\mathcal{O}(w)}}$. By the above we can assume $|\mathfrak{A}| = 2^{|\varphi^*|w^{\mathcal{O}(w)}} = 2^{|\varphi|^{\mathcal{O}(1)}w^{\mathcal{O}(w)}}$ therefore also $|\mathfrak{R}_N| = 2^{|\varphi|^{\mathcal{O}(1)}w^{\mathcal{O}(w)}}$. □

5 Complexity of query answering

Query answering is the problem of deciding $\varphi \models q$ for a given $\varphi \in$ GF and q a UCQ. By Lemma 10 (ii) this amounts to testing unsatisfiability of the guarded sentence $\varphi \wedge \neg\chi_q$, known to be 2ExpTime-complete and in DTime($2^{\mathcal{O}((r+|\varphi|+|\chi_q|)w^w)}$), where r is the size and w the width of τ [12]. Query answering is thus 2ExpTime-complete and in DTime($2^{|\varphi|w^w+(|q|rw)^{\mathcal{O}(|q|w)}}$). Note that 2ExpTime-completeness holds even for a fixed query q.

Also note that for q an ACQ (union of *acyclic* BCQs) the exponential blow-up in passing from q to χ_q can be avoided. Regarding query answering against a fixed $\varphi \in$ GF we thus find that for ACQs the complexity reduces to ExpTime. In fact, it can also be show to be ExpTime-complete.

For a fixed $\psi \in \mathsf{GF}[\tau \cup \sigma]$ *target query answering* is the problem of deciding $D \wedge \psi \models q$ on input q a UCQ and D a τ-structure (given as a conjunction of ground atoms with elements of D as individual constants). The next theorem summarises our observations on query answering and some results on subproblems of target query answering.

Theorem 19.

1. *Deciding $\varphi \models q$, on input $\varphi \in$ GF and q a UCQ, is* 2ExpTime*-complete (even if the query is fixed).*
2. *For each $\varphi \in$ GF, deciding $\varphi \models q$ on input q an ACQ, is in* ExpTime*; and* ExpTime*-complete for certain φ. (Hence for certain signatures τ satisfiability for $\mathsf{GF}[\tau]$ is* ExpTime*-complete, strengthening a result of [12].)*
3. *There is a* GF*-sentence ψ such that deciding $D \wedge \psi \models Q$, on input Q a BCQ and D a conjunction of atoms* of bounded width*, is* PSpace*-hard.*
4. *For all* universal $\psi \in$ GF*, deciding $D \wedge \psi \models q$, on input q a UCQ and D a conjunction of atoms, is in Π_2^P; for certain* universal ψ *it is Π_2^P-complete already for conjunctive queries q.*
5. *For all $\psi \in$ GF and q a UCQ, deciding $D \wedge \psi \models q$ on input D, is in co-NP and co-NP-complete already for $q = \bot$ and certain* universal ψ.
   Hence, satisfiability of $D \wedge \psi$ on input D is in NP and NP-complete for certain universal $\psi \in$ GF.

6 Canonisation and capturing

As explained above, the *guarded bisimulation invariant* $I(\mathfrak{A})$ of $\mathfrak{A}$, defined as the bisimulation quotient of the guarded bisimulation game graph $G(\mathfrak{A})$, is an abstraction of the complete GF-theory of $\mathfrak{A}$ or, equivalently, of its $\sim_{\mathrm{g}}$-equivalence class. $I(\mathfrak{A})$ can, in polynomial time, be computed via an inductive refinement procedure, which successively classifies maximal guarded tuples up to $\sim_{\mathrm{g}}^i$ for increasing levels of i until a fixed point ($\sim_{\mathrm{g}}^i = \sim_{\mathrm{g}}^{i+1} = \sim_{\mathrm{g}}$ in restriction to $\mathfrak{A}$) is reached. Furthermore, fixing an arbitrary linear ordering on the finite set of atomic τ-types, i.e. the equivalence classes w.r.t. $\sim_{\mathrm{g}}^0$, and extending this inductively to higher $\sim_{\mathrm{g}}^i$-classes by lexicographic refinement, one can even generate, along with $I(\mathfrak{A})$, a linear ordering of $I(\mathfrak{A})$.

The problem of *inverting the invariant* concerns the reconstruction of an actual model $\mathfrak{A}$ from a given invariant. The construction of the Rosati cover $\mathfrak{R}_2$ on the basis of a given invariant $I(\mathfrak{A})$ as an input solves this problem in PTime for every fixed signature τ.

Theorem 6. *For every relational signature τ there exists a* PTIME *algorithm computing from a given invariant $I(\mathfrak{A})$ of an unspecified τ-structure $\mathfrak{A}$ a finite τ-structure $\mathrm{can}(\mathfrak{A})$ such that $I(\mathrm{can}(\mathfrak{A})) = I(\mathfrak{A})$; hence $can(\mathfrak{A}) \sim_{\mathrm{g}} \mathfrak{A}$ and $can(\mathfrak{A}') = can(\mathfrak{A})$ whenever $\mathfrak{A} \sim_{\mathrm{g}} \mathfrak{A}'$.*

Proof. Starting with the $\sim_{\mathrm{g}}$-invariant $\mathfrak{I} = I(\mathfrak{A})$, the Rosati cover with parameter $N = 2$ yields a canonical structure $\mathfrak{R}_2(\mathfrak{I})$ such that, by Lemma 12, $I(\mathfrak{R}_2(\mathfrak{I})) = \mathfrak{I}$. In other words $\mathfrak{R}_2(\mathfrak{I}) \sim_{\mathrm{g}} \mathfrak{A}'$ for any $\mathfrak{A}'$ with $I(\mathfrak{A}') = \mathfrak{I}$, i.e., for any $\mathfrak{A}' \sim_{\mathrm{g}} \mathfrak{A}$. It follows from the bounds given there that, for fixed τ, $|\mathfrak{R}_2(\mathfrak{I})| = |\mathfrak{I}|^{O(1)}$. Set $\mathrm{can}(\mathfrak{A}) := \mathfrak{R}_2(I(\mathfrak{A}))$. Note that the concrete polynomial complexity and size bounds will depend on τ. □

As pointed out above, the invariant $I(\mathfrak{A})$ can, even as a linearly ordered structure, be obtained in an inductive fixpoint process of lexicographic refinement. A closer analysis shows that this translates into uniform interpretability in inductive fixpoint logic IFP of $I(\mathfrak{A})$ over $\mathfrak{A}$ itself (as a linearly ordered set of equivalence classes of tuples over $\mathfrak{A}$). Since $I(\mathfrak{A})$ is linearly ordered and IFP-interpretable over $\mathfrak{A}$, the Immerman–Vardi theorem tells us that $can(\mathfrak{A})$ is itself uniformly IFP-interpretable over $\mathfrak{A}$.

The class of all PTIME (or IFP) queries against these IFP-interpreted canonical representatives thus captures precisely the class of *all* PTIME queries that are invariant under guarded bisimulation equivalence. Intuitively, we may think of the evaluation of the canonisation, $\mathfrak{A} \mapsto can(\mathfrak{A})$, as a pre-processing step that acts as a filter to enforce invariance under $\sim_{\mathrm{g}}$. This entails Corollary 7 on capturing.

References

[1] H. Andréka, J. van Benthem and I. Németi. Modal languages and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27:217–274, 1998.

[2] C. Beeri, R. Fagin, D. Maier and M. Yannakakis. On the desirability of acyclic database schemes. *Journal of the ACM*, 30(3):479–513, 1983.

[3] C. Beeri and M. Y. Vardi. The implication problem for data dependencies. Rroc. ICALP'81pp. 73-85, 1981.

[4] C. Beeri, M. Y. Vardi. A proof procedure for data dependencies. JACM 31(4) pp.218-741, 1984. 741.

[5] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. Proc. KR'06, 2006.

[6] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati: Tractable reasoning and efficient query answering in description logics: the DL-Lite family. J. Autom. Reasoning 39(3) pp. 385-429, 2007.

[7] A. Calí, G. Gottlob, and M. Kifer. Taming the infinite chase: query answering under expressive relational constraints. Proc. KR'08, pp. 70-80, 2008.

[8] A. Calí, G. Gottlob, T. Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. . Proc. PODS 2009, 2009.

[9] A. Calí, G. Gottlob, T. Lukasiewicz. Datalog$^{\pm}$: a unified approach to ontologies and integrity constraints. Proc. ICDT'09 pp. 14-30, 2009.

[10] R. Fagin, Ph. Kolaitis, R. J. Miller and L. Popa. Data exchange: semantics and query answering. *Theor. Comp. Sci.*, 336(1):89-124, 2005.

[11] E. Grädel. Description logics and guarded fragments of first-order logic. In *Proc. DL'98*, CEUR Electronic Workshop Proc., 1998.

[12] E. Grädel. On the restraining power of guards. *Journal of Symbolic Logic*, 64(4):1719–1742, 1999.

[13] E. Grädel. Decision procedures for guarded logics. In *Proc. CADE'99*, pp. 31-51, 1999.

[14] B. Herwig. Extending partial isomorphisms on finite structures. Combinatorica, n.155, pp. 365-371, 2005.

[15] I. Hodkinson. Loosely guarded fragment of first-order logic has the finite model property. *Studia Logica*, 70:205–240, 2002.

[16] I. Hodkinson and M. Otto. Finite conformal hypergraph covers and Gaifman cliques in finite structures. *Bulletin of Symbolic Logic*, 9:387–405, 2003.

[17] D. S. Johnson and A. C. Klug. Testing containment of conjunctive queries under functional and inclusion dependencies. JCSS 28(1):167-189, 1984.

[18] M. Otto. Bisimulation-invariant Ptime and higher-dimensional μ-calculus. *TCS*, 224:237-265, 1999.

[19] M. Otto. Avoiding incidental homomorphisms into guarded covers. Techn. rep., TU Darmstadt, 2009.

[20] M. Otto. Highly Acyclic Groups, Hypergraph Covers and the Guarded Fragment. In *Proc. LICS'10*, 2010.

[21] H. Pérez-Urbina, B. Motik, and I. Horrocks. Tractable query answering and rewriting under description logic constraints. J. Applied Logic, 2009 (to appear).

[22] R. Rosati. On the decidability and finite controllability of query processing in databases with incomplete information. In *Proc. PODS'06*, pp. 356–365, 2006. Full version submitted, available from author.