Guarded negation

Vince Barany, Balder Ten Cate, Luc Segoufin

Guarded negation

Vince Bárány, LogicBlox Inc.
Balder ten Cate, LogicBlox Inc. and UC Santa Cruz
Luc Segoufin, INRIA and ENS-Cachan

Abstract

We consider restrictions of first-order logic and of fixpoint logic in which all occurrences of negation are required to be guarded by an atomic predicate. In terms of expressive power, the logics in question, called GNFO and GNFP, extend the guarded fragment of first-order logic and the guarded least fixpoint logic, respectively. They also extend the recently introduced unary negation fragments of first-order logic and of least fixpoint logic.

We show that the satisfiability problem for GNFO and for GNFP is 2ExpTime-complete, both on arbitrary structures and on finite structures. We also study the complexity of the associated model checking problems. Finally, we show that GNFO and GNFP are not only computationally well behaved, but also model theoretically: we show that GNFO and GNFP have the tree-like model property and that GNFO has the finite model property, and we characterize the expressive power of GNFO in terms of invariance for an appropriate notion of bisimulation.

Our complexity upper bounds for GNFO and GNFP hold true even for their "clique-guarded" extensions CGNFO and CGNFP, in which clique guards are allowed in the place of guards.

1 Introduction

Modal logic is well known for its "robust decidability": not only are basic decision problems such as satisfiability, validity and entailment decidable, but the decidability of these problems is preserved under various natural variations and extensions to the syntax and semantics of modal logic (e.g., addition of fixpoint operators, backward modalities, nominals; restriction to finite structures). As observed by Vardi [Var96], this robust decidability is intimately linked to the fact that modal logic has a combination of three properties, namely (i) the *tree model property* (if a formula has a model, it has a model which is a tree), (ii) *translatability into tree automata* (each formula can be transformed into a tree automaton, or equivalently, an MSO formula, recognizing its tree models) and, (iii) the *finite model property* (every satisfiable modal formula is satisfied in a finite structure). The decidability of satisfiability for modal logic, both on arbitrary structures and on finite structures, follows immediately from these three properties. Similar arguments can be used to show the good behavior of many extensions of modal logic, although we should note here that the two-way μ-calculus (the extension of modal logic with fixpoint operators and backward modalities) lacks the finite model property, and hence the decidability of satisfiability on finite structures for this logic involves a separate (non-trivial) argument [Boj03].

The properties (i), (ii) and (iii) described above can be viewed as a semantic explanation for the robust decidability of modal logic. Given that modal logic can be viewed as a syntactic fragment of first-order logic, it is also natural to ask for syntactic explanations: *what syntactic features of modal formulas (viewed as first-order formulas) are responsible for their good behavior? And can we generalize modal logic, preserving these features, while at the same time dropping inessential restrictions inherent in modal logic (such as the fact that it can only describe structures with unary and binary relations)?*

Several answers to these questions have been proposed. The first one is to consider the two variable fragment of first-order logic, which is decidable and has the finite model property [Mor75]. Unfortunately, this observation does not go very far towards explaining the robust decidability of modal logic, since it seems impossible to extend the two variable fragment with a fixpoint mechanism while maintaining decidability [GOR99].

The second proposal is to consider logics with guarded quantifications. The *guarded fragment* of first-order logic (GFO), introduced in [ABN98], consists of FO formulas in which all quantifiers are "guarded" by atomic predicates. It has a natural extension with fixpoint operators (GFP) that extends the two-way μ-calculus [GW99]. Both GFO and GFP have the tree-like model property (if a formula has a model, it has one of bounded tree width), they can be translated into tree automata (each formula can be transformed into a tree automaton recognizing tree decompositions of its models of bounded tree width) and GFO has the finite model property [ABN98, Grä01]. Finite satisfiability of GFP was only recently proved decidable in [BB12].

The third, and most recent proposal is based on unary negation. Unary negation first-order logic (UNFO) restricts first-order logic by constraining the use of negation to subformulas having at most one free variable (and viewing universal quantification as a defined connective). Unary negation fixpoint (UNFP) is the natural extension of UNFO using monadic fixpoints. Again, UNFO generalizes modal logic, and UNFP generalizes the two-way μ-calculus. Both UNFO and UNFP have the tree-like model property, they can be translated into tree automata and UNFO has the finite model property [CS13]. Decidability of finite satisfiability for UNFP was also established in [CS13].

The three extensions of modal logics presented above are incomparable in terms of expressive power. In particular there are properties expressible in UNFO that are not expressible in GFO and vice-versa. In this paper we unify the unary negation and guarded quantification approaches by introducing guarded-negation logics.

Guarded-negation first-order logic (GNFO) restricts FO by requiring that all occurrences of negation are of the form $\alpha \wedge \neg\phi$ where the "guard" α is an atomic formula (possibly an equality statement) containing all the free variables of ϕ. For instance, GNFO cannot express $x \neq y$ but it can express $R(x, y, z) \wedge x \neq y$. We also disallow universal quantification as a primitive connective (though a limited form of universal quantification can be expressed using existential quantification and guarded negation). For instance, GNFO cannot express $\forall \bar{x}\, R(\bar{x})$ but it can express $\forall \bar{x}\, S(\bar{x}) \rightarrow R(\bar{x})$ as $\exists y\, y = y \wedge \neg\big(\exists \bar{x}\, S(\bar{x}) \wedge \neg R(\bar{x})\big)$. Guarded-negation fixpoint logic (GNFP) extends GNFO with a guarded fixpoint mechanism. In terms of expressive power, GNFO forms a strict extension of both UNFO and GFO.

We show that our guarded-negation logics have the same desirable properties as modal logics, unary negation logics and guarded logics: Both GNFO and GNFP have the tree-like model property, they can be translated into tree automata and GNFO has the finite model property.

More precisely, we show that the satisfiability problem for GNFO and GNFP is decidable, both on arbitrary structures and on finite structures. These two problems are both 2ExpTime-complete, even for a fixed finite signature (in contrast the satisfiability of GFO decreases from 2ExpTime to ExpTime when the signature is fixed).

We also study the (combined) complexity of the model checking problem of GNFO and GNFP. The problem is $\mathrm{P}^{\mathrm{NP}[O(\log^2 n)]}$-complete for GNFO. In the case of GNFP, it is hard for P^{NP} and contained in $\mathrm{NP}^{\mathrm{NP}} \cap \mathrm{coNP}^{\mathrm{NP}}$. These results are obtained using a simple polynomial time reduction to their unary negation variants, UNFO and UNFP, whose model checking was solved in [CS13]. Recall that the model checking problem of GFO is PTime-complete [BG01] and that a similar gap between the upper bound and the lower bound exists for GFP and the μ-calculus, where the complexity of model checking is known to lie between PTime and NP $\cap$ coNP [BG01].

Next, we explore the model theory of GNFO. We define a guarded-negation variant of bisimulation suitable for guarded-negation logics and most of our results build on the fact that guarded-negation logics are invariant under guarded-negation bisimulations. The appropriateness of guarded-negation bisimulation is illustrated by showing that GFO is exactly the fragment of first-order logic that is invariant under guarded-negation bisimulation.

Finally, we show that our complexity results can be lifted to the *clique-guarded* extensions of GNFO and GNFP, which provide a further generalization of GNFO and GNFP that subsume the *clique-guarded fragment* (as well as the closely related *loosely guarded fragment* and *packed fragment*) [vB97, Mar99, Gr].

The most involved result is the decidability of satisfiability on finite structures. For GNFO, we give a reduction to testing whether a union of conjunctive queries is implied by a guarded formula, recently shown decidable in [BGO14]. In the case of GNFP, we make a reduction to the decidability of finite satisfiability of GFP, recently proved in [BB12].

An extended abstract of this paper was published in [BtCS11]. The present paper provides detailed proofs of the results presented there. In particular we have clarified and fixed several issues concerning certain definitions. In addition, it contains new material concerning syntactic variants of guarded-negation fixpoint logic, and concerning clique-guarded negation logics.

Outline of the paper Guarded-negation first-order logic, GNFO, is presented in Section 2 and its satisfiability is shown decidable in Section 3. The fixpoint extension of GNFO, GNFP, is introduced in Section 4 where it is shown to be decidable via a reduction to GFP. The same reduction also implies the finite model property of GNFO. The model checking problems of GNFO and GNFP are studied in Section 5. A variant of bisimulation suitable for guarded-negation formulas is introduced in Section 6, where it is also shown that GNFO is exactly those first-order formulas closed under guarded-negation bisimulation. The tree-like model property of GNFO and GNFP is derived from this notion. Finally, in Section 7, we extend our results to a generalization of GNFO and GNFP with clique-guards.

2 Preliminaries

Structures and formulas We restrict our attention to relational structures. However, as we will explain in Section 7.2, all complexity results presented in this paper generalize to the case with constant symbols.

A (relational) *signature* τ is a finite set of relation symbols, each having an associated arity. By the *arity of a signature*, we mean the maximal arity of its relations. A *structure* M over a relational signature τ consists of a set $dom(M)$, the *domain* of M, together with an interpretation R^M of each relation symbol $R \in \tau$, which is a k-ary relation over $dom(M)$, where k is the arity of R according to τ. A structure M is said to be *finite* if $dom(M)$ is finite. An *expansion* of a structure M over τ is a structure M' over a signature $\sigma \supseteq \tau$ such that M and M' agree on their domain and on the interpretation of all relation symbols in τ. If a tuple of elements $\bar{a}$ from $dom(M)$ belongs to the interpretation of a relation symbol R, then we say that $R(\bar{a})$ is a *fact* of M. A set of elements of M is *guarded* (in M) if it is either a singleton set or there is a fact of M containing all its elements. A tuple of elements of M is guarded if the set of elements occurring in the tuple is guarded. We denote by $guarded(M)$ the set of all guarded tuples of M. If M and N are structures and $\bar{a}$ and $\bar{b}$ are tuples of elements from $dom(M)$ and $dom(N)$, respectively, then we say that $(M, \bar{a})$ and $(N, \bar{b})$ are *locally isomorphic* if there is a partial isomorphism $f : M \to N$ such that $f(\bar{a}) = \bar{b}$.

We assume familiarity with first-order logic, FO, and least fixpoint logic, LFP, over relational structures. We use classical syntax and semantics for FO and LFP. The *size of a formula* ϕ, denoted by $|\phi|$, is the number of symbols needed to write down the formula. We use the notation $\phi(\bar{x})$ to indicate that the free variables of ϕ are exactly the variables in $\bar{x}$. A *sentence* is a formula with no free variable. We say that a structure M is a *model* of a sentence ϕ if $M \models \phi$. We also write $M \models \phi(\bar{u})$ or $(M, \bar{u}) \models \phi(\bar{x})$ when a tuple $\bar{u}$ of elements of the structure M makes the formula $\phi(\bar{x})$ true in M. Finally we write $\models \varphi$ if φ is true in all structures.

Conjunctive queries A conjunctive query (CQ) is a first-order formula of the form

$$\exists y_1 \cdots y_l \, (\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n)$$

where each α_i is an atomic formula, possibly an equality statement. A union of conjunctive queries (UCQ) is a disjunction of CQs. A positive-existential query is a first-order formula built using disjunction, conjunction and existential quantification only. Every positive-existential query can be transformed in a UCQ at the cost of a possible exponential blow-up. The *width* of a CQ is the number of variables occurring in it, and the width of a UCQ is the maximum width of its CQs. The *height* of a UCQ is the maximum size of its CQs. In particular the height of a CQ is it size.

GNFO We define GNFO, *guarded-negation first-order logic,* as the fragment of FO given by the following grammar, where R ranges over predicate symbols, and $\alpha(\bar{x}\bar{y})$ is an atomic formula (possibly an equality statement).

$$\varphi ::= \quad R(\bar{x}) \mid x = y \mid \exists x \, \varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \alpha(\bar{x}\bar{y}) \wedge \neg\varphi(\bar{y}) \tag{1}$$

Hence the logic can only negate a subformula if all its free variables are "guarded" by some fact, or if the subformula has at most one free variable (in which case one can use an equality statement of the form $x = x$ or $y = y$ as the guard). For example, $x \neq y$ is not a formula of GNFO but $R(x, y, z) \wedge x \neq y$ is. Notice that all positive-existential queries belong to GNFO. We write GNFO$[\tau]$ for the set of formulas using relation symbols in the signature τ.

We will refer to formulas of the form $\alpha(\bar{x}\bar{y}) \wedge \varphi'(\bar{x})$ (where α is an atomic formula) as *answer-guarded formulas*. In addition, we consider every atomic formula by itself to be an answer-guarded formula (motivated by the fact that $\alpha(\bar{x})$ is equivalent to $\alpha(\bar{x}) \wedge \alpha(\bar{x})$.

We say that a formula of GNFO is in *GN-normal form* if, in its syntax tree, no disjunction is directly below an existential quantifier or a conjunction, and no existential quantifier is directly below a conjunction sign. Every GNFO formula can be brought into GN-normal form, at the cost of an exponential increase in length and linear increase in the number of variables, by pushing out disjunction and pushing in conjunctions using the following rewriting rules (where x' is a variable not occurring in $\psi \wedge \phi$ and $\phi[x'/x]$ is the formula constructed from ϕ by replacing all occurrences of x by x'):

$$\begin{aligned} \exists x\ (\phi \vee \psi) &\rightarrow\ \exists x\ \phi \vee \exists x\ \psi \\ \phi \wedge (\psi \vee \chi) &\rightarrow\ (\phi \wedge \psi) \vee (\phi \wedge \chi) \\ (\exists x\ \phi) \wedge \psi &\rightarrow\ \exists x'\ (\phi[x'/x] \wedge \psi) \end{aligned}$$

The appeal of the GN-normal form is that it highlights the fact that GNFO formulas can be naturally viewed as being built up from atomic formulas using guarded negation, and unions of conjunctive queries. Indeed, the GNFO formulas in GN-normal form are precisely generated by the following recursive definition:

$$\varphi ::= \quad R(\bar{x}) \mid x = y \mid \alpha(\bar{x}\bar{y}) \wedge \neg\varphi(\bar{y}) \mid q[\varphi_1/U_1, \ldots, \varphi_s/U_s] \tag{2}$$

where q is a UCQ using relation symbols $U_1, \ldots, U_s$, and $\varphi_1, \ldots, \varphi_s$ are answer-guarded formulas generated by the same recursive definition with the appropriate number of free variables corresponding to the relation symbols they replace. Here, $q[\varphi_1/U_1, \ldots, \varphi_s/U_s]$ is the result of replacing in q all subformulas of the form $U_i(\bar{x})$ with $i \leq s$ by $\varphi_i(\bar{x})$.

A formula of GNFO is said to be *of width k* if, when brought into GN-normal form in the way described above, it uses at most k variables (or equivalently, is built up using UCQs q of width at most k). We denote by GNFO^k all GNFO formulas of width k.

Example 1. *Consider for example the existential positive formula*

$$\exists xy\ \big(R(x) \wedge \exists z\ T(x, y, z) \wedge \exists z'\ S(x, z')\ \big).$$

When brought into GN-normal form it gives

$$\exists xyzz'\ R(x) \wedge T(x, y, z) \wedge S(x, z').$$

and has width 4. Notice that it is also equivalent to the GNFO *formula*

$$\exists xyz\ R(x) \wedge T(x, y, z) \wedge \neg\neg(\exists z\ S(x, z))$$

that is in GN-normal form and has width 3. Recall that $\neg\neg(\exists z\ S(x, z))$ is indeed a formula of GNFO *as it negates unary formula and those could be seen as guarded by an equality atom. In other words it is equivalent to $x = x \wedge \neg(x = x \wedge \neg(\exists z\ S(x, z)))$*

GNFO extends GFO and UNFO GNFO generalizes the unary negation logic, UNFO, studied in [CS13], which only allows the negation of formulas having at most one free variable. It also generalizes the *guarded fragment of first-order logic* (GFO). The logic GFO is the fragment of FO defined by the following grammar, where, again, $\alpha(\bar{x}\bar{y}\bar{z})$ is an atomic formula (possibly an equality statement):

$$\varphi ::= R(\bar{x}) \mid x = y \mid \varphi \vee \varphi | \varphi \wedge \varphi | \neg\varphi \mid \exists\bar{x}\ \alpha(\bar{x}\bar{y}\bar{z}) \wedge \varphi(\bar{x}\bar{y}) \mid \forall\bar{x}\ \alpha(\bar{x}\bar{y}\bar{z}) \rightarrow \varphi(\bar{x}\bar{y})$$

It is straightforward to check that:

Proposition 2. *Every* GFO *sentence is equivalent to a* GNFO *sentence, via a polynomial time transformation.*

This result extends to answer-guarded formulas, however $\neg R(xy)$ is in GFO but not expressible in GNFO.

Proof. Consider a GFO-sentence φ. Because φ is closed, every subformula $\vartheta(\bar{x})$ of φ with free variables $\bar{x}$ falls in the scope of an (innermost) guarded quantifier with some guard, let us refer to it as $\alpha_\vartheta(\bar{x}\bar{u})$. We can therefore safely replace in φ each negated subformula $\vartheta(\bar{x})$ of the form $\neg\psi(\bar{x})$ with $(\alpha_\vartheta(\bar{x}\bar{u}) \wedge \vartheta(\bar{x}))$. Each universally quantified subformula $\vartheta(\bar{x}\bar{z})$ of the form $\forall\bar{y}\ \alpha(\bar{x}\bar{y}\bar{z}) \rightarrow \psi(\bar{x}\bar{y})$ is equivalent to $\neg\exists\bar{y}\ (\alpha(\bar{x}\bar{y}\bar{z}) \wedge \neg\psi(\bar{x}\bar{y}))$. It can therefore be replaced in φ by $\alpha_\vartheta(\bar{x}\bar{z}\bar{u}) \wedge \neg\,(\exists\bar{y}\ \alpha(\bar{x}\bar{y}\bar{z}) \wedge \neg\psi(\bar{x}\bar{y}))$. Let $\hat{\varphi}$ be the sentence so obtained from φ. By construction, $\hat{\varphi}$ is a sentence of GNFO, and it is easy to see that $\hat{\varphi}$ is logically equivalent to φ. □ □

The following example shows that GNFO is strictly more expressive than GFO and UNFO.

Example 3. *The* GNFO *sentence δ defined as*

$$\exists xy\big(E(x,y) \wedge \neg\exists uvw(E(x,u) \wedge E(u,v) \wedge E(v,w) \wedge E(w,y))\big)$$

is not equivalent to any GFO *sentence or to any* UNFO *sentence, even on undirected graphs. This is because δ defines a property that is not invariant under guarded bisimulation (which, incidentally, amounts to ordinary bisimulation in case of simple graphs), as can be easily verified, nor is it invariant under "UN-bisimulation" as befits* UNFO *formulas, cf. [CS13].*

3 The satisfiability problem for GNFO

In this section we show that (finite) satisfiability for GNFO is 2ExpTime-complete. The 2ExpTime lower bound follows immediately from the fact that satisfiability for UNFO is already hard for 2ExpTime [CS13, Grb]. It holds even if the signature is fixed (recall that when the signature is fixed the complexity of satisfiability for GFO is ExpTime-complete).

The upper bound is proved using a reduction to the problem of testing whether a GFO formula entails (on finite structures) a UCQ. The latter problem is also known as the problem of query answering against a GFO theory, and it has been solved in [BGO14]. To streamline the presentation, we will allow the possibility of zero-ary relation symbols.

The reduction is obtained by rewriting the formula by adding new relational symbols in order to simplify it while preserving its satisfiability status. The first step is the following lemma.

Lemma 4. *Given any formula $\varphi(\bar{x}) \in \mathsf{GNFO}[\tau]$ we can construct in polynomial time a companion formula $\psi(\bar{x}) \in \mathsf{GNFO}[\tau \cup \sigma]$ of the form*

$$\psi(\bar{x}) = \underbrace{S(\bar{x}) \wedge \bigwedge_j \forall\,\bar{z}\bar{u}\ R_j(\bar{z}\bar{u}) \rightarrow q_j(\bar{z})}_{\psi^+} \quad \wedge \quad \underbrace{\bigwedge_i \forall\,\bar{z}\bar{u}\ T_i(\bar{z}\bar{u}) \rightarrow \neg p_i(\bar{z})}_{\psi^-} \tag{3}$$

where σ is the signature (disjoint from τ) consisting of the relation symbols S and T_i[1], where the R_j are atomic formulas, the q_j's and p_i's are positive-existential first-order formulas, $\mathrm{width}(\psi) = \mathrm{width}(\varphi)$ *and such that*

$$\models \varphi \leftrightarrow \exists\sigma\,\psi$$

where $\exists\sigma$ is a shorthand for the existential second-order quantification of all the symbols in σ.

[1] In (3) the R_j's and T_i's are not necessarily distinct; moreover the size of the tuples $\bar{z}$ and $\bar{u}$ will in general vary from predicate to predicate and are denoted here uniformly by "$\bar{z}$" and "$\bar{u}$" only for sake of a simpler illustration.

Proof. Given a GNFO-formula φ consider an inner-most occurrence of a guarded negation $R(\bar{z}\bar{u}) \wedge \neg q(\bar{z})$ as a subformula of φ. Then $q(\bar{z})$ is necessarily positive existential. Let T be a new predicate symbol of the same arity as R. We substitute $T(\bar{z}\bar{u})$ in the input formula for the subformula $R(\bar{z}\bar{u}) \wedge \neg q(\bar{z})$, and add the following as conjuncts to ψ^+ and ψ^-, according to their kind.

$$\begin{array}{l} \forall\, \bar{z}\bar{u}\; T(\bar{z}\bar{u}) \rightarrow \neg q(\bar{z}) \\ \forall\, \bar{z}\bar{u}\; T(\bar{z}\bar{u}) \rightarrow R(\bar{z}\bar{u}) \\ \forall\, \bar{z}\bar{u}\; R(\bar{z}\bar{u}) \rightarrow T(\bar{z}\bar{u}) \vee q(\bar{z}) \end{array}$$

Inner-most *equality-guarded* negations $z = u \wedge \neg q(z, u)$ are handled in a similar fashion. Again, $q(z, u)$ must be positive-existential. We choose a new unary relation symbol T, replace the subformula in question by $z = u \wedge T(z)$, and add $\forall\, z\; T(z) \rightarrow \neg q[u/z]$ and $\forall\, z\; T(z) \vee q[u/z]$ as conjuncts to the normal form, where $q[u/z]$ is the formula constructed from $q(z, u)$ by replacing all occurrences of u by z.

Proceeding in this manner from the inside-out we eliminate all guarded negations until the original input formula is reduced to a single positive-existential formula $p(\bar{x})$ (in the extended signature). Finally we replace $p(\bar{x})$ with $S(\bar{x})$ where S is an appropriate new predicate symbol and add $\forall\, \bar{x}.S(\bar{x}) \rightarrow p(\bar{x})$ as conjunct to the normal form, which is thus finalized. It is now easy to verify the correctness of this transformation. □ □

In view of Lemma 4, it remains to reduce the satisfiability problem of formulas in the form of (3) to the query answering problem against a GFO theory.

We may assume without loss of generality that the positive-existential formulas q_j of (3) are in prenex normal form, i.e. $q_j(\bar{z}) = \exists \bar{u}\, \xi_j(\bar{z}, \bar{v})$ for some quantifier-free positive formula $\xi_j(\bar{z}, \bar{v})$. Also note that each conjunct $\forall\, \bar{z}\bar{u}\; T_i(\bar{z}\bar{u}) \rightarrow \neg p_i(\bar{z})$ of (3) is the negation of a positive-existential sentence $\exists\, \bar{z}\bar{u}\; T_i(\bar{z}\bar{u}) \wedge p_i(\bar{z})$. Therefore, the entire ψ^- of (3) can be conceived as the negation of a single positive-existential sentence q. This leads us to the following equivalent formula.

$$\underbrace{S(\bar{x}) \;\wedge\; \bigwedge_j \left(\forall\, \bar{z}\bar{u}\; R_j(\bar{z}\bar{u}) \rightarrow \exists \bar{v}\; \xi_j(\bar{z}\bar{v}) \right)}_{\psi^+} \;\wedge\; \neg \underbrace{\bigvee_i \left(\exists\, \bar{z}\bar{u}\; T_i(\bar{z}\bar{u}) \wedge p_i(\bar{z}) \right)}_{q} \tag{4}$$

Observe next that without affecting satisfiability of (4) we may introduce new atoms guarding the existential quantifiers in ψ^+ thus obtaining, from ψ^+, a GFO-formula

$$\psi^* = S(\bar{x}) \;\wedge\; \bigwedge_j \left(\forall\, \bar{z}\bar{u}\; R_j(\bar{z}\bar{u}) \rightarrow \exists \bar{v}\; Q_j(\bar{z}\bar{v}) \wedge \xi_j(\bar{z}\bar{v}) \right)$$

where the Q_j's are distinct new relation symbols of appropriate arity. Then, $\models \psi^* \rightarrow \psi^+$ and, conversely, every model of ψ^+ has an expansion that is a model of ψ^*.

The entire transformation of an input GNFO-formula φ to the equi-satisfiable $\psi^* \wedge \neg q$, with ψ^* in GFO and q positive existential, can be performed in polynomial time and only results in a polynomial blowup in the signature of the latter formula. In a final transformation step, which may require at most exponential time, the positive-existential sentence q can be converted to an equivalent Boolean UCQ q^*. In general q^* may be comprised of exponentially many CQs each of size at most $|q|$. Summing up all the reduction steps we obtain:

Proposition 5. *Given any formula $\varphi(\bar{x}) \in \mathsf{GNFO}[\tau]$ one can compute in exponential time a* GFO*-formula $\psi^*(\bar{x})$ and* UCQ *q^*, both over a signature $\tau \uplus \{\bar{T}\}$, such that*

$$\models \varphi \;\longleftrightarrow\; \exists \bar{T}\, (\, \psi^* \wedge \neg\, q^*\,) \tag{5}$$

and such that $|\psi^*|$ and $\mathrm{height}(q^*)$ are polynomial in $|\varphi|$.

We now summarize the main results of [BGO14]. Later we will build on key elements of the construction of [BGO14], stated below as Lemmas 13 and Theorem 20, from which the following Theorem 6 can be directly derived.

Theorem 6 ([BGO14]). *Given a* GFO*-formula ψ and a* UCQ *q of height h it is decidable in time $|q| \cdot 2^{(h|\psi|)^{\mathcal{O}(h|\psi|)}}$ whether or not $\psi \wedge \neg q$ is satisfiable; and if $\psi \wedge \neg q$ has a model then it has a finite model of size $2^{(h|\psi|)^{\mathcal{O}(h^2|\psi|)}}$*

By combining Theorem 6 with the estimates of Proposition 5 we derive the complexity of satisfiability for GNFO, as well as its finite model property.

Theorem 7. *1. The satisfiability problem for* GNFO *is 2*ExpTime*-complete.*

2. Every satisfiable GNFO*-sentence φ has a finite model of size $2^{2^{|\varphi|^{\mathcal{O}(1)}}}$.*

4 Adding fixpoints: the satisfiability problem for GNFP

In this section, we introduce and study GNFP, which, in a nutshell, is the extension of GNFO with guarded fixpoints. We show that both satisfiability and finite satisfiability are decidable for GNFP.

GNFP Guarded-negation fixpoint logic, GNFP, is a syntactic fragment of least fixpoint logic LFP, from which it inherits the semantics, cf. [DG02]. Recall that the syntax of LFP assumes an infinite supply of (second-order) fixpoint variables (denoted $X, Y, Z, \ldots$), of arbitrary arity. These fixpoint variables are distinct from the relation symbols that are in the relational signature (denoted $P, Q, R, S, \ldots$), and they serve for expressing fixpoints. Syntactically, they are treated on a par with the ordinary relation symbols: they can be used in the same way in formulas. LFP then extends FO with the ability to construct formulas of the form

$$[\mu_{Z,\bar{z}}\ \phi(\bar{Y}, Z, \bar{z}, \bar{u})](\bar{x})$$

where Z is a fixpoint variable whose arity matches the length of sequences $\bar{z}$ and $\bar{x}$, such that Z occurs only positive in ϕ (i.e., all occurrences are under an even number of negation signs).

The guarded-negation fragment of LFP, called GNFP, is defined as follows:

Definition 8. *Formulas of* GNFP$[\tau]$*, we omit the signature τ when it is clear from the context, pertain to the following syntax:*

$$\begin{array}{rcl} \phi & ::= & R(\bar{x}) \mid x = y \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \exists x\, \phi \mid \alpha(\bar{x}\bar{y}) \wedge \neg\phi(\bar{x}) \mid \\ & & \beta(\bar{u}\bar{w}) \wedge Z(\bar{u}) \mid \mu_{Z,\bar{z}}[\,\phi(\bar{Y}, Z, \bar{z})\,](\bar{x}) \end{array}$$

where R is any relational symbol in τ, $\alpha(\bar{x}\bar{y})$ and $\beta(\bar{u}\bar{w})$ are atomic τ-formulas (possibly equality statements) and, in the last clause of the definition, the fixpoint variable Z occurs only positively in $\phi(\bar{Y}, Z, \bar{z})$, i.e. always under an even number of negations.

Note that

- no first-order parameters (i.e., free variables other than those $\bar{z}$ bound by the fixpoint operator) are permitted in the matrix of a fixpoint operator,
- free fixpoint variables $\bar{Y}$ other than Z are still allowed, enabling nesting and alternation of fixpoint definitions;
- fixpoint variables cannot be used as guards, and in fact, all atomic formulas involving fixpoint variables must be guarded by atomic τ-formulas or equalities.

As we mentioned before, the semantics of GNFP is inherited from the logic LFP, of which it is a fragment. We briefly recall here the semantics of the fixpoint operator. Take a formula of the form

$$\mu_{Z,\bar{z}}[\,\phi(\bar{Y}, Z, \bar{z})\,](\bar{x})$$

and consider any structure $(M, \bar{S})$, where M is a structure over the relational signature and $\bar{S}$ is a collection of relations over the domain of M (of suitable arity) that form the interpretation for the second-order

variables $\bar{Y}$. Since Z occurs in ϕ only positively, $\phi(\bar{Y}, Z, \bar{z})](\bar{x})$ induces a monotone operation $\mathcal{O}_\phi$ on n-ary relations over the domain of M, where n is the arity of the fixpoint variable Z, and where $\mathcal{O}_\phi(R) = \{\bar{a} \mid (M, \bar{S}, R) \models \phi(\bar{a})\}$. By the Knaster-Tarski fixpoint theorem, this monotone operation has a unique least-fixpoint. By definition, an n-tuple $\bar{b}$ of elements of M satisfies the formula $[\mu_{Z,\bar{z}}\ \phi(\bar{Y}, Z, \bar{z})](\bar{x})$ in $(M, \bar{S})$ if and only if $\bar{b}$ belongs to this least fixpoint. The least fixpoint of the monotone operation $\mathcal{O}_\phi$ is known to be the intersection of all its pre-fixpoints, i.e., $\bigcap\{R \mid R \supseteq \mathcal{O}_\phi(R)\}$, and it can be equivalently characterized as ${\mathcal{O}_\phi}^\kappa(\emptyset)$ with $\kappa = |dom(M)|$, where ${\mathcal{O}_\phi}^0(\emptyset) = \emptyset$; for all successor ordinals $\lambda + 1$, ${\mathcal{O}_\phi}^{\lambda+1}(\emptyset) = \mathcal{O}_\phi({\mathcal{O}_\phi}^\lambda(\emptyset))$; and for all limit ordinals $\lambda \leq \kappa$, ${\mathcal{O}_\phi}^\lambda(\emptyset) = \bigcup_{\lambda' < \lambda} {\mathcal{O}_\phi}^{\lambda'}(\emptyset)$.

It is worth noting that, although the relation $\mathcal{O}_\phi^\lambda(\emptyset)$ (for λ an ordinal) may contain unguarded tuples, the syntax of GNFP guarantees that the relation $\mathcal{O}_\phi^{\lambda+1}(\emptyset)$ depends only on the τ-guarded tuples in $\mathcal{O}_\phi^\lambda(\emptyset)$, and similarly for limit ordinals. In this sense, the fixpoint variables can be taken to range over guarded relations (i.e., relations consisting of guarded tuples only).

Example 9. *The fixpoint formula*

$$\mu_{Z,x,y}[\, E(x,y) \vee \exists z\ (Z(x,z) \wedge E(z,y))\,](u,v)$$

computing the transitive closure of E is not a formula of GNFP *as the matrix formula does not guard the variables x, z occurring in $Z(x, z)$.*

The fixpoint formula

$$\mu_{Z,z}[\, y = z \vee \exists y'(Z(y') \wedge E(y', y))\,](x)$$

computing the connected component of y is also not a formula of GNFP *as the matrix formula has y as a parameter.*

However the fixpoint formula

$$\mu_{Z,z}[\, B(z) \vee \exists y'(Z(y') \wedge E(y', z))\,](x)$$

computing the set of nodes reachable from a node in B is in GNFP *as singletons are guarded by definition.*

Notes on syntax We could have defined GNFP with different alternative syntaxes of equivalent expressive power and varying degrees of succinctness. The variations concern how guardedness of fixpoint predicates is enforced. Let us briefly discuss the alternatives.

An ostensibly more restrictive syntax is obtained if we require that every fixpoint formula is to be guarded by a single atom of the base signature.

$$\mu_{Z,\bar{z}}[\, \alpha(\bar{z}) \wedge \phi(\bar{Y}, Z, \bar{z})\,](\bar{x}) \tag{6}$$

Notice how every formula of this form can be promptly rewritten in our syntax of choice as $\mu_{Z,\bar{z}}[\, \phi^*(\bar{Y}, Z, \bar{z})\,](\bar{x})$ where ϕ^* is obtained from ϕ by replacing in it each atom $Z(\bar{u})$ with the conjunction $\alpha(\bar{u}) \wedge Z(\bar{u})$. This transformation is obviously linear in the number of atoms.

In [BtCS11] we presented GNFP using a similar pattern of fixpoint definitions

$$\mu_{Z,\bar{z}}[\, \mathtt{guarded}_\tau(\bar{z}) \wedge \phi(\bar{Y}, Z, \bar{z})\,](\bar{x}) \tag{7}$$

where the clause $\mathtt{guarded}_\tau(\bar{z})$ is understood as a shorthand formula signifying guardedness in the signature τ without expressly declaring any concrete guard. Notice that adding the special guardedness atom to a fixpoint definition in either of the two earlier forms does not affect the meaning of the formula but ensures compliance with this most liberal syntax. Conversely, to transcribe a formula adhering to syntax (7) one can replace each occurrence of an atom $Z(\bar{z})$ involving a fixpoint predicate variable Z with the disjunction $\bigvee_i \exists \bar{w}^i \alpha_i(\bar{w}^i \bar{z}) \wedge Z(\bar{z})$ where $\bigvee_i \exists \bar{w}^i \alpha_i(\bar{w}^i \bar{z})$ spells out the definition of $\mathtt{guarded}_\tau(\bar{z})$. This translation too is linear for any fixed signature, but is exponential in the maximum arity of relation symbols in the signature.

In order to show that syntax (6) does not restrict the expressive power of GNFP we shall temporarily avail ourselves of a further syntactic enhancement: that of *simultaneous fixpoint definitions*. Let again

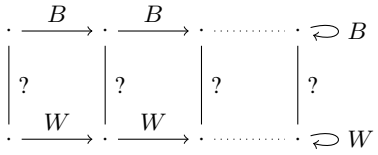

Figure 1: Illustration of the example fixpoint formula

$\bigvee_{1\leq i\leq n} \exists \bar{w}^i \alpha_i(\bar{w}^i \bar{z})$ be the definition of guardedness of the tuple $\bar{z}$ in signature τ. To translate a fixpoint definition of the form (7) into (6) notation consider first the simultaneous fixpoint schema

$$\mu \left\{ \begin{array}{rcl} Z_1(\bar{w}^1\bar{z}) & \leftarrow & \alpha_1(\bar{w}^1\bar{z}) \wedge \phi'(\bar{Y}, \bar{Z}, \bar{z}) \\ & \vdots & \\ Z_n(\bar{w}^n\bar{z}) & \leftarrow & \alpha_n(\bar{w}^n\bar{z}) \wedge \phi'(\bar{Y}, \bar{Z}, \bar{z}) \end{array} \right\}$$

where now Z has been replaced with a tuple $\bar{Z} = Z_1, \ldots, Z_n$ of fixpoint variables, one for each disjunct in the definition of guardedness, with each Z_i of arity $|\bar{w}^i| + |\bar{z}|$, and where ϕ' is obtained by replacing in ϕ every occurrence of an atom $Z(\bar{u})$ with the disjunction $\bigvee_i \exists \bar{w}^i Z_i(\bar{w}^i \bar{u})$. Notice how the definition of each fixpoint predicate variable Z_i is guarded according to the pattern (6). This simultaneous fixpoint can be transformed using the Bekič principle, cf. e.g. [AN01], with a formula for each of the alternative Z_i

$$\mu_{Z_i, \bar{w}^i, \bar{z}} \left[\alpha_i(\bar{w}^i\bar{z}) \wedge \phi_i^*(\bar{Y}, \bar{Z}, \bar{z})\right]$$

pertaining to (6). Finally, the original fixpoint formula as in (7) is rewritten to

$$\bigvee_i \exists \bar{u}^i \; \mu_{Z_i, \bar{w}^i, \bar{z}} \left[\alpha_i(\bar{w}^i\bar{z}) \wedge \phi_i^*(\bar{Y}, \bar{Z}, \bar{z})\right] (\bar{u}^i \bar{x})$$

and this rewriting is to be applied recursively to formulas with nested fixpoint definitions. Notice how this translation of formulas from dialect (7) to dialect (6) of GNFP will thus produce exponentially long formulas in terms of the number of nested fixpoints in the formula started with, even if we fix the signature.

To illustrate this translation consider the example, over graphs with black and white edges, provided by the formula $\rho(p, q)$ defined as

$$\mu_{Z,x,y} \left[\texttt{guarded}_{\{B,W\}}(x, y) \wedge (\exists uv \left(B(x, u) \wedge W(y, v) \wedge Z(u, v)\right) \vee \left(B(x, x) \wedge W(y, y)\right))\right] (p, q)$$

expressing the existence of two parallel paths, one black the other white, forming a reel of conjoined cells ending in adjacent vertices with a black and a white self-loop, respectively (cf. Figure 1). First observe that $\texttt{guarded}_{\{B,W\}}(x, y)$ is equivalent to the disjunction $B(x, y) \vee W(x, y) \vee B(y, x) \vee W(y, x) \vee x = y$. Correspondingly, we construct a simultaneous fixpoint schema consisting of five equations:

$$\mu \left\{ \begin{array}{l} Z_1(x, y) \leftarrow B(x, y) \wedge \exists uv\big(B(x, u) \wedge W(y, v) \wedge (\bigvee_i Z_i(u, v))\big) \vee \big(B(x, x) \wedge W(y, y)\big) \\ Z_2(x, y) \leftarrow W(x, y) \wedge \exists uv\big(B(x, u) \wedge W(y, v) \wedge (\bigvee_i Z_i(u, v))\big) \vee \big(B(x, x) \wedge W(y, y)\big) \\ \vdots \end{array} \right\}$$

To simplify the notation and keep the example comprehensible, we will (against better knowledge) pretend that the fixpoint schema we constructed consists only of the two formulas for Z_1 and Z_2 spelled out above. Then, eliminating the simultaneous fixpoint definition results in two nested fixpoint formulas, first

$$\begin{array}{rc} \zeta_1(p, q) = & \mu_{Z_1,x,y}[B(x, y) \wedge \exists uv\big(B(x, u) \wedge W(y, v) \wedge (Z_1(u, v) \vee \xi_2(Z_1, u, v))\big) \\ & \vee \big(B(x, x) \wedge W(y, y)\big)] \; (p, q) \end{array}$$

where

$$\begin{array}{rc} \xi_2(Z_1, u, v) = & \mu_{Z_2,\hat{x},\hat{y}}[W(\hat{x}, \hat{y}) \wedge \exists \hat{u}\hat{v}\big(B(\hat{x}, \hat{u}) \wedge W(\hat{y}, \hat{v}) \wedge (Z_1(\hat{u}, \hat{v}) \vee Z_2(\hat{u}, \hat{v})\big) \\ & \vee \big(B(\hat{x}, \hat{x}) \wedge W(\hat{y}, \hat{y})\big)] \; (u, v) \end{array}$$

and, symmetrically,

$$\zeta_2(p,q) = \quad \mu_{Z_2,x,y}[W(x,y) \wedge \exists uv\big(B(x,u) \wedge W(y,v) \wedge (Z_1(u,v) \vee \xi_1(Z_2,u,v)\big) \\ \vee \big(B(x,x) \wedge W(y,y)\big)]\ (p,q)$$

where

$$\xi_1(Z_W,u,v) = \quad \mu_{Z_1,\hat{x},\hat{y}}[B(\hat{x},\hat{y}) \wedge \exists \hat{u}\hat{v}\big(B(\hat{x},\hat{u}) \wedge W(\hat{y},\hat{v}) \wedge (Z_1(\hat{u},\hat{v}) \vee Z_2(\hat{u},\hat{v})\big) \\ \vee \big(B(\hat{x},\hat{x}) \wedge W(\hat{y},\hat{y})\big)]\ (u,v)$$

corresponding to the cases of reels starting in a black and in a white cross edge, respectively.

Finally, the original black and white reel formula $\rho(p,q)$ is transcribed into the equivalent formula $\bigvee_i \zeta_i(p,q)$, which adheres to the most restrictive of the three alternative syntaxes for GNFP discussed above, i.e., the syntax given in (6).

This example also illustrates the fact that allowing simultaneous fixpoint schemas in GNFP formulas does not increase the expressive power of the logic (although it can facilitate more succinct definitions).

We have opted for the syntax of definition 8 because it allows for more intuitive formulas than (6) while also avoiding a penalty of increased complexity of model checking that the most liberal syntax (7) entails, as discussed in Section 5.

The definition of GN-normal form that we gave for GNFO formulas applies to GNFP as well. Formulas of GNFP in GN-normal form can be naturally thought of as being built up from atomic formulas using (i) guarded negation, (ii) unions of conjunctive queries, and (iii) fixpoint operators. As in the case of GNFO, the *width* of a GNFP-formula is the number of variables it contains after being put in GN-normal form and we let GNFP^k denote the set of GNFP-formulas of width k.

GNFP extends GFP and UNFP Syntactically, GNFP generalizes the logic UNFP studied in [CS13], which only allows the negation of formulas having at most one free variable, and only unary fixpoints. GNFP also generalizes the *guarded fragment of fixpoint logic* (GFP) [GW99], in the sense that every sentence of GFP is equivalent to a sentence of GNFP.

Recall that GFP is the fragment of LFP obtained by extending GFO with least fixpoints: given a formula $\phi(\bar{Y},Z,\bar{z})$ that is positive in Z, has no free first-order variables other than $\bar{z}$ and Z has arity the number of variables in $\bar{z}$, the formula $\mu_{Z,\bar{z}}[\,\phi(\bar{Y},Z,\bar{z})\,]$ is also a formula of GFP. Although the occurrences of fixpoint variables are not required to be guarded, in the context of a GFP *sentence*, every occurrence of an atom using a fixpoint relation is implicitly guarded, namely by the atom guarding the closest quantifier whose scope includes the occurrence in question). This implies:

Proposition 10. *Every sentence of* GFP *is equivalent to a sentence of* GNFP*, via a polynomial time transformation.*

GNFP is decidable The aim of this section is to establish the following main result.

Theorem 11. *It is decidable whether a sentence of* GNFP *has a model and whether it has a finite model. Both of these problems are* 2ExpTime*-complete.*

The proof of Theorem 11 is a reduction to the (finite) satisfiability of GFP: given a formula of GNFP we construct a formula of GFP whose (finite) satisfiability is equivalent to the one of the initial formula and we then apply known results on (finite) satisfiability for GFP, namely [GW99] for the infinite case and [BB12] for the finite case. Before we describe the reduction, we start with some useful notation and some preliminary results taken from [BGO14].

Acyclic structures, acyclic queries, and treeifications A structure M is said to be *acyclic* if it admits a *guarded tree decomposition*, that is, a tree decomposition each bag of which belongs to *guarded*(M) [Yan81, FFG02]. We omit here the definition of tree decomposition, which can be found e.g. in [FFG02], as it turns out not to be important in what follows. The above definition of acyclicity extends also to *conjunctive queries*. Formally, we associate to each query $q(\bar{x})$ a structure $[q]$, called the *canonical structure*

of q, whose nodes are the variables occurring in q, and whose facts are the atoms of q. A conjunctive query is *acyclic* if its canonical structure has a guarded tree decomposition. Acyclic conjunctive queries have been studied extensively in database theory, and have been shown to have many desirable properties [Yan81]. Every acyclic conjunctive query can be equivalently rewritten (in polynomial time) as a formula of GFO built with only conjunction and existential quantification, and, conversely, every such GFO formula can be rewritten (in polynomial time) as an acyclic conjunctive query [FFG02]. For instance the query $\exists yzw\, T(x,y,z) \wedge T(x,w,z) \wedge E(x,y)$ is acyclic because it is equivalent to the guarded formula $\exists yz\, T(x,y,z) \wedge E(x,y) \wedge (\exists w\, T(x,w,z))$.

Definition 12 (Treeification)**.** *Given a signature τ, the τ-*treeification *$\Lambda_q^\tau(\bar{x})$ of a positive existential query $q(\bar{x})$ over τ is the* UCQ *consisting of the disjunction of all those* acyclic CQ*s over τ (modulo renaming of bound variables) that imply q and that are minimal (in the sense that removing any atomic formula would render it non-acyclic or not implying q).*

Lemma 13 below will justify this definition by showing that there are only finitely many minimal acyclic conjunctive queries (up to logical equivalence) that imply a given query q.

First we give an example. Consider the conjunctive query $q(x)$ defined as

$$\exists yzw\, (E(x,y) \wedge E(y,z) \wedge E(z,w) \wedge E(w,x)).$$

Then its $\{E\}$-treeification $\Lambda_q^{\{E\}}$ is the formula

$$E(x,x) \vee \exists y\, (E(x,y) \wedge E(y,x)).$$

Indeed, the only minimal acyclic queries implying $q(x)$ are obtained by identifying some of its variables resulting in either a reflexive edge on x or a pair of inverse edges. If the signature is $\{E,T\}$, where T is a ternary predicate, the treeification of $q(x)$ has a number of additional disjuncts corresponding to various triangulations of $q(x)$, such as

$$\exists yzw\, \big(T(x,y,z) \wedge T(x,w,z) \wedge E(x,y) \wedge E(y,z) \wedge E(z,w) \wedge E(w,x)\big)$$

(which is acyclic because it is equivalent to $\exists yz\, \big(T(x,y,z) \wedge E(x,y) \wedge E(y,z) \wedge \exists w\, \big((T(x,w,z) \wedge E(z,w) \wedge E(w,x))\big)\big)$. It can be shown that each disjunct in the treeification of any CQ in whatever signature contains at most three times as many atoms as the CQ itself [BGO14] leading to the following observations.

Lemma 13. *Consider a signature τ having r many predicate symbols of maximal arity w. Let $q(\bar{x})$ be a* UCQ *of height h over τ. Then $\Lambda_q^\tau(\bar{x})$ has width w, size $r^{\mathcal{O}(h)}(hw)^{\mathcal{O}(hw)}$, height $\mathcal{O}(hw)$, and can be constructed in time $|q|r^{\mathcal{O}(h)}(hw)^{\mathcal{O}(hw)}$.*

These figures constitute a slight refinement of those offered in [BGO14, Lemma 10], where it was shown that every disjunct T in the treeification of Q has at most $c = 3$ times as many atoms as Q, whence the prenex normal form of T involves at most $c|Q|w$ many variables and has size $\mathcal{O}(|Q|w)$. In fact, one can check that $c = 2$ suffices and that the number of CQs of these dimensions is at most $(r(c|Q|w)^w)^{c|Q|}$. Note that when writing T as a guarded formula only w many variables are needed and the resulting formula is still of size $\mathcal{O}(|Q|w)$. Because each acyclic CQ as above may occur (modulo renaming of bound variables) at most once in Λ_Q^τ, we find that $|\Lambda_Q^\tau| = (r(c|Q|w)^w)^{c|Q|}\mathcal{O}(|Q|w) = r^{\mathcal{O}(|Q|)}(|Q|w)^{\mathcal{O}(|Q|w)}$, the width of Λ_Q^τ is w, and its height is $\mathcal{O}(|Q|w)$.

Remark 14. *Over acyclic structures, q and Λ_q^τ are equivalent. Indeed consider a conjunctive query q and an acyclic structure M such that $M \models q(\bar{b})$. Then there is an homomorphism f from q to M. The image $f(q)$ can be viewed as an acyclic conjunctive query that clearly implies q.*

In general q and Λ_q^τ are not equivalent but we will use structures that are locally sufficiently acyclic such that equivalence is achieved for small queries q. This suggests the following definition:

Definition 15 (Allowing for treeifications)**.** *A τ-structure M* allows for treeifications of width n *if for every conjunctive query q of width n and tuple $\bar{b}$ guarded in M, we have*

$$M \models q(\bar{b}) \leftrightarrow \Lambda_q^\tau(\bar{b}).$$

We now discuss ways of obtaining structures that allow for treeifications using *guarded bisimulations*. Guarded bisimulations [ABN98] form a fundamental tool in the study of of guarded logics. In particular, the existence of a guarded bisimulation implies GFP-indistinguishability [Gr]. We briefly review guarded bisimulations and some of their applications here. Later, in Section 6, we will introduce *guarded-negation bisimulations* in order to capture, in the same way, the expressive power of GNFP.

Recall the notion of a guarded tuple and the notation *guarded*(M) from Section 2.

Definition 16 (Guarded bisimulation)**.** *Let M, N be two structures. A guarded bisimulation between M and N is a binary relation $Z \subseteq$ guarded$(M) \times$ guarded(N) such that, for every pair $(\bar{a}, \bar{b}) \in Z$, where $\bar{a} = a_1, \ldots, a_m$ and $\bar{b} = b_1, \ldots, b_n$, the following conditions hold:*

- *$(M, \bar{a})$ and $(N, \bar{b})$ are locally isomorphic (that is, $m = n$ and the relation $\{(a_i, b_i) \mid 1 \leq i \leq n\}$ is the graph of a partial isomorphism between M and N).*
- *[**Forward clause**] For every tuple $\bar{a}'$ in guarded(M) there is a tuple $\bar{b}'$ in guarded(N) such that $(\bar{a}', \bar{b}')) \in Z$ and, whenever $a'_i = a_i$ for some $i \leq n$ then also $b'_i = b_i$.*
- *[**Backward clause**] For every tuple $\bar{b}'$ in guarded(N) there is a tuple $\bar{a}'$ in guarded(M) such that $(\bar{a}', \bar{b}')) \in Z$ and, whenever $b'_i = b_i$ for some $i \leq n$ then also $a'_i = a_i$.*

Theorem 17 ([ABN98, Gr])**.**
If Z is a guarded bisimulation between structures M and N, and $(\bar{a}, \bar{b}) \in Z$, then, for all GFP*-formulas $\phi(\bar{x})$, $M \models \phi\,[\bar{a}]$ iff $N \models \phi\,[\bar{b}]$.*

One important consequence of Theorem 17 is that GFP has the *tree-like model property*: every satisfiable GFP-formula has an acyclic model. This follows from Theorem 17, because every structure is guarded bisimilar to an acyclic structure [ABN98, Grä01].

Even though every structure is guarded bisimilar to an acyclic structure, the latter is in general infinite even if the original structure was finite. For example, let M be the structure that consists of a directed R-cycle of length 3 (where R is a binary relation symbol). It is easy to see that every acyclic structure M' that is guarded bisimilar to M must contain an infinite R-path, and no reflexive or symmetric R-edges. It follows that, if M' is finite, then it must contain some minimal directed R-cycle of length at least 3. This shows that M' cannot be finite and acyclic at the same time.

To address this problem, in [BGO14], a construction was presented, parametrized by a natural number n, that takes any finite structure M and produces a "weakly n-acyclic" finite companion structure $M^{(n)}$ that is guarded bisimilar to M and that allows for treeifications of conjunctive queries of width at most n. To state the result formally, we need the concept of *guarded bisimilar covers* due to [Ott04], cf. [BGO14, Definition 1].

Definition 18 (Guarded bisimilar cover)**.** A guarded bisimilar cover $\pi\colon N \xrightarrow{\sim} M$ *is a surjective homomorphism $\pi\colon N \to M$ such that the induced map $\{(\bar{b}, \pi(\bar{b})) \mid \bar{b}$ guarded in $N\}$ is a guarded bisimulation. A cover $\pi\colon N \xrightarrow{\sim} M$ is* weakly k-acyclic *if for every homomorphism $h\colon Q \to N$ with $|Q| \leq k$ the composition $\pi \circ h$ factors as $g \circ f$ for some homomorphisms $f\colon Q \to T$ and $g\colon T \to M$ where T is acyclic.*

Note that, in the above, if Q is the *canonical structure* $[q]$ of a CQ q and t is the acyclic CQ such that $T = [t]$ then we have $t \models q$. Therefore, in the above T can wlog. be chosen with t corresponding to a disjunct of the treeification of q. The following is thus a straighforward corollary of (and motivation for) the definitions involved.

Fact 19. *If $\pi\colon N \xrightarrow{\sim} M$ is a weakly k-acyclic guarded bisimilar cover of M then (i) M and N have identical* GFP *theories and (ii) N allows for treeifications of width k.*

Proof. The first claim is a corollary of guarded bisimulation invariance of GFP. For the second claim, consider $N \models q(\bar{b})$ for some guarded tuple $\bar{b}$ and a CQ $q(\bar{x})$ of width at most k. Let $h\colon [q] \to N$ be the homomorphism witnessing this. Then, by definition of weak k-acyclicity we have $\pi \circ h = g \circ f$ for homomorphisms $f\colon [q] \to [t]$ and $g\colon [t] \to M$ where t is some acyclic CQ. In particular, we have that $t(\bar{x}) \models q(\bar{x})$ and that $M \models t(\pi(\bar{b}))$. Note that as such t can be chosen to be minimal (in terms of number of atoms) and thus a disjunct of the treeification $\Lambda_q(\bar{x})$ of $q(\bar{x})$. Also, by guarded bisimulation invariance of acyclic conjunctive queries we get that $N \models t(\bar{b})$, whence $N \models \Lambda_q(\bar{b})$. □

The following technical result will play a key role in our argument.

Theorem 20 ([BGO14, Theorem 4])**.** *For every finite relational structure M and every $n \in \mathbb{N}$ one can effectively construct a weakly n-acyclic guarded bisimilar cover $\pi\colon M^{(n)} \xrightarrow{\sim} M$ of size $|M^{(n)}| = |M|^{w^{\mathcal{O}(n^2)}}$, where w is the maximal arity of the relations of M. Furthermore, $M^{(n)}$ is n-*conformal*, meaning that every clique-guarded tuple of $M^{(n)}$ of size at most n is guarded. $M^{(n)}$ is called the n-th* Rosati cover *of M.*

The last assertion of Theorem 20 requires some explanation. We say that a tuple $\bar{b}$ of elements of a structure is *clique-guarded* if for every pair $b_i, b_j \in \bar{b}$, b_i and b_j co-occur in an atomic fact. Note that every guarded tuple is clique-guarded. The n-conformality expresses a restricted form of the converse direction, which will be put to use later on, in Section 7.1.

Fact 19 asserts that weakly k-acyclic covers allow for treeifications of width k. The next lemma identifies conditions under which this key property extends to suitable expansions of the base structure and of the cover. Below we use this observation in our inductive argument for Lemma 23 concerning GNFP-formulas with free fixpoint variables.

Say that $Z \subseteq M^r$ is a *guarded relation* over M if every tuple $\bar{a} \in Z$ is guarded in M. Given a cover $\pi\colon N \xrightarrow{\sim} M$, say that a guarded relation $W \subseteq N^r$ is *π-saturated* if $W = \pi^{-1}(\pi(W)) \cap \mathrm{guarded}(N)$, viz. if $W = \pi^{-1}(Z) \cap \mathrm{guarded}(N)$ for some guarded relation Z over M.

Lemma 21. *Consider a weakly (wn^w)-acyclic guarded bisimulation cover $\pi : N \xrightarrow{\sim} M$ of some relational τ-structure M, where w is the maximal arity of the relation symbols in the signature τ.*[2] *Let $Z_1, \ldots, Z_t$ be guarded relations over M and for each $1 \leq i \leq t$ let $W_i = \pi^{-1}(Z_i) \cap \mathrm{guarded}(N)$. Then $(N, W_1, \ldots, W_t)$ allows for treeifications of width n.*

Proof. We write $\widehat{N}$ for $(N, W_1, \ldots, W_t)$ and $\widehat{M}$ for $(M, Z_1, \ldots, Z_t)$. Let τ denote the signature of N and M and σ the signature of $\widehat{N}$ and $\widehat{M}$. We first observe that π remains a guarded bisimilar cover from $\widehat{N}$ to $\widehat{M}$. This is an immediate consequence of the fact that the Z_i are guarded relations and the W_i are π-saturated.

Consider now a conjunctive query $q(\bar{x})$ of width n in the signature σ and a guarded tuple $\bar{b}$ of elements of $\widehat{N}$ such that $\widehat{N} \models q(\bar{b})$. We need to show that $\widehat{N} \models \Lambda_q^{\sigma}(\bar{b})$.

Let $h\colon [q], \bar{x} \to \widehat{N}, \bar{b}$ be a homomorphism witnessing $\widehat{N} \models q(\bar{b})$. Consider an atom α of q whose symbol is in $\sigma \setminus \tau$. Its image by h is a tuple $\bar{c}$ from W_l for l as specified by α. Since by assumption the relations W_l are guarded in N, $\bar{c}$ is guarded by a tuple $\bar{d}$ occurring in some relation R of N.

Let $R(\bar{z})$ be a new atom such that for all i, z_i is the smallest variable y_j of α such that $h(y_j) = d_i$ or let z_i be a fresh new variable if there is no such y_j. We denote this atom by $\alpha[h, \tau]$. Let Q be the query in the signature τ constructed from q by omitting all its atoms α whose symbol is in $\sigma \setminus \tau$, replacing each by $\alpha[h, \tau]$, and by quantifying existentially all the fresh new variables. By construction, h can be extended to a homomorphism $H\colon ([Q], \bar{x}) \to (N, \bar{b})$.

Wlog. we may assume that the atoms α of q being replaced are pairwise incomparable, so that no one is contained inside another (otherwise we may freely omit the replacement of the smaller one from Q). Also note that the assumptions of the lemma imply that each W_i is of arity at most w. The number of (maximal) atoms α of q thus being replaced is no more than n^w, so that Q has at most that many replacement atoms $\alpha[h, \tau]$, each contributing at most $w - 1$ new variables. The width of Q is therefore not greater than the original n plus $(w - 1)n^w$, not more than wn^w.

[2] A slightly more meticulous argument operating at the level of the underlying hypergraphs, as in [BGO14], would allow one to accurately establish the tight weak acylicity bound of n in place of the (wn^w) stated here.

Let q' be the conjunctive query in the signature σ obtained by adding to q the conjunct $\alpha[h,\tau]$ for all (maximal) $\sigma \setminus \tau$-atoms α and quantifying existentially all the fresh new variables. Note that q' implies Q and also implies q, that $[q']$ has the same domain as $[Q]$ and that H also witnesses $\widehat{N} \models q'(\bar{b})$. Our aim is to show $\widehat{N} \models \Lambda^{\sigma}_{q'}(\bar{b})$ from which $\widehat{N} \models \Lambda^{\sigma}_{q}(\bar{b})$ will follow trivially.

Let $\bar{a} = \pi(\bar{b})$. By construction we have $H' = \pi \circ H$ maps $[Q], \bar{x}$ into $M, \bar{a}$. By virtue of Fact 19 it holds that $M \models \Lambda^{\tau}_{Q}(\bar{a})$, witnessed by a homomorphism $G' \colon [T], \bar{x} \to M, \bar{a}$, with T a disjunct of Λ^{τ}_{Q} such that $H' = G' \circ F$ for some $F \colon [Q(\bar{x})] \to [T(\bar{x})]$.

Let t' be constructed from T by adding to T the conjunct $F(\alpha)$ for every atom $\alpha \in \sigma \setminus \tau$ in q (hence also in q'). By construction t' remains acyclic and $F \colon [q'(\bar{x})] \to [t'(\bar{x})]$, which means that t' implies q'. Recall that H also witnesses the fact that $\widehat{N} \models q'(\bar{b})$ and therefore H' witnesses the fact that $\widehat{M} \models q'(\bar{a})$. As $H' = G' \circ F$ and $F \colon [q'(\bar{x})] \to [t'(\bar{x})]$, it must be the case that G' witnesses the fact that $\widehat{M} \models t'(\bar{a})$. Altogether we have shown that $t'(\bar{x})$ is an acyclic CQ that implies $q'(\bar{x})$ and that is satisfied in $\widehat{M}$ by $\bar{a}$. Therefore $G' \circ F$ witnesses $\widehat{M} \models \Lambda^{\sigma}_{q'}(\bar{a})$. As $\Lambda^{\sigma}_{q'}$ is acyclic it is invariant under guarded bisimulation and, therefore, from $\widehat{M} \models \Lambda^{\sigma}_{q'}(\bar{a})$ we get $\widehat{N} \models \Lambda^{\sigma}_{q'}(\bar{b})$ as desired. □

Reduction to (finite) satisfiability for GFP Let φ be any given GNFP sentence. As a first step, we compute its GN-normal form $\tilde{\varphi}$. Note that $\tilde{\varphi}$ has the following dimensions: $|\tilde{\varphi}| = 2^{\mathcal{O}(|\varphi|)}$, $\mathrm{width}(\tilde{\varphi}) = \mathcal{O}(|\varphi|)$, and $\tilde{\varphi}$ is built up using only UCQs of height at most $|\varphi|$ (as well as guarded negations and fixpoint operators) as in (2).

Next, essentially, our reduction transforms all UCQs occurring in $\tilde{\varphi}$ to their treeifications. For every $k \geq 1$, and for every relational signature τ consisting of at most k-ary relations, we define a translation η from GNFP$^k[\tau]$ formulas in GN-normal form to GFP$^k[\tau \uplus \{C_k\}]$ formulas, where C_k is a new symbol of arity k, by structural recursion, using the following rules.

$$
\begin{aligned}
\eta(\, R(\bar{x})\,) &= R(\bar{x}) && (a)\\
\eta(\, \alpha(\bar{x}\bar{y}) \wedge Z(\bar{x})\,) &= \alpha(\bar{x}\bar{y}) \wedge Z(\bar{x}) && (b)\\
\eta(\, \alpha(\bar{x}\bar{y}) \wedge \neg\psi(\bar{x})\,) &= \alpha(\bar{x}\bar{y}) \wedge \neg\, \eta(\psi(\bar{x})) && (c)\\
\eta(\, \mu_{Z,\bar{z}}[\,\psi(\bar{Y}, Z, \bar{z})\,]\,) &= \mu_{Z,\bar{z}}[\,\eta(\psi(\bar{Y}, Z, \bar{z}))\,] && (d)\\
\eta(\, q[\phi_1/U_1, \ldots, \phi_s/U_s]\,) &= \Lambda_q^{\tau \uplus \{U_1,\ldots,U_s,C_k\}}[\eta(\phi_1)/U_1, \ldots, \eta(\phi_s)/U_s] && (e)
\end{aligned}
$$

where in (e) q is a UCQ of signature $\{U_1, \ldots, U_s\}$ disjoint from $\tau \uplus \{\bar{Y}, C_k\}$ and $\phi_1, \ldots, \phi_s \in$ GNFP$^k[\tau \uplus \{\bar{Y}\}]$, where $\bar{Y}$ enumerates the free fixpoint variables occurring in any of the ϕ_i's, each ϕ_i being a guarded formula.

By (2) all formulas in GN-normal form can be decomposed as in (a)–(e) and we have the following bounds on the translation η.

Lemma 22. *For every* GNFP^k*-formula φ with GN-normal form $\tilde{\varphi}$ we have $|\eta(\tilde{\varphi})| = 2^{(k|\varphi|)^{\mathcal{O}(1)}}$ and $\eta(\tilde{\varphi})$ can be computed within this time bound and its width remains k.*

Proof. To establish the bound $|\eta(\tilde{\varphi})| = 2^{(k|\varphi|)^{\mathcal{O}(1)}}$ we proceed via structural induction on φ following the definition of the translation η according to the cases (a)-(e) and using as invariant the claim $|\eta(\tilde{\psi})| \leq 2^{c|\tau|k^2|\psi|^3}$, where c is an appropriate constant to be fixed later. The claim of the lemma follows assuming wlog. that $|\tau| \leq |\varphi|$.

Irrespective of c this bound trivially holds for all atomic formulas whether based on a τ-predicate (a) or a fixpoint variable (b). It is also plain to see that assuming $|\eta(\tilde{\psi})| = 2^{c|\tau|k^2|\psi|^3}$ the same bound holds for (c) all guarded-negation formulas of the form $\alpha(\bar{x}\bar{y}) \wedge \neg\psi(\bar{x})$, as well as for (d) least fixpoint formulas $\mu_{Z,\bar{z}}[\,\psi(\bar{Y}, Z, \bar{z})\,]$. In each of these cases $\eta(\tilde{\psi})$ is computable in $2^{(k|\psi|)^{\mathcal{O}(1)}}$-time, assuming the same for the relevant subformulas of ψ.

The remaining case is when $\tilde{\varphi}$ is $q[\tilde{\phi}_1/U_1, \ldots, \tilde{\phi}_s/U_s]$ for q a UCQ and $\phi_1, \ldots, \phi_s$ subformulas of φ. We have already noted that the height h of q is no more than $|\varphi|$, and the same holds for s too. Further, by design we know that the maximum arity of the predicates among $\tau \uplus \{U_1, \ldots, U_s, C_k\}$ is k. Therefore,

using Lemma 13, we have $|\Lambda_q^{\tau \uplus \{U_1,\ldots,U_s,C_k\}}| = (|\tau|+s+1)^{\mathcal{O}(h)}(kh)^{\mathcal{O}(kh)} \leq 2^{c|\tau|(k|\varphi|)^2}$ for some constant c. By the induction hypothesis for each $1 \leq i \leq s$ we have $|\eta(\tilde{\phi}_i)| \leq 2^{ck^2|\phi_i|^3}$. Let $l = \max_i|\phi_i| < |\psi|$. Then the size of $\eta(\tilde{\varphi})$, obtained by substituting $\tilde{\phi}_i$ for U_i for every $1 \leq i \leq s$ in $\Lambda_q^{\tau \uplus \{U_1,\ldots,U_s,C_k\}}$, is bounded by $2^{c|\tau|(k|\varphi|)^2} \cdot 2^{c|\tau|k^2l^3} \leq 2^{c|\tau|k^2|\varphi|^2(l+1)} \leq 2^{c|\tau|k^2|\varphi|^3}$. That $\eta(\tilde{\varphi})$ can be computed in the stated time bound also for φ of type (e) follows similarly using Lemma 13. □ □

The following key lemma attests to the correctness of our reduction. It is proved by structural induction on formulas, while relying on Theorem 20 and Lemma 21 to deal with the cases (d) and (e) of the translation, respectively.

Lemma 23. *Let $\pi : N \mathrel{\widetilde{\to}} M$ be a weakly (wk^w)-acyclic guarded bisimulation cover of a $\tau \uplus \{C_k\}$-structure M, where w is the maximal arity of the symbols of $\tau \uplus \{C_k\}$ and let $\phi(\bar{Y}, \bar{x}) \in \mathrm{GNFP}^k[\tau]$ be a formula in GN-normal form with free fixpoint variables $\bar{Y}$. Then for every interpretation of $\bar{Y}$ by π-saturated[3] guarded relations $\bar{W}$ on N and for every guarded tuple $\bar{b}$ in N we have:*

$$(N, \bar{W}) \models \eta(\phi)(\bar{b}) \leftrightarrow \phi(\bar{b}) \; .$$

Proof. We proceed by induction on the structure of ϕ, wlog. in GN-normal form. The base case is trivial since $\phi = \eta(\phi)$ for all atomic formulas. Moreover, the claim trivially distributes over positive Boolean combinations. It is equally clear that if the claim holds for some ψ then it also holds for $\phi(\bar{x}\bar{y}) = \alpha(\bar{x}\bar{y}) \wedge \neg\psi(\bar{x})$ (note that the guard $\alpha(\bar{x}\bar{y})$ ensures that the equivalence $\eta(\psi)(\bar{x}) \leftrightarrow \psi(\bar{x})$ is only ever used for guarded instantiations of $\bar{x}$.)

Consider the case of $\phi = q[\phi_1/U_1, \ldots, \phi_s/U_s](\bar{Y}, \bar{x})$ where q is a UCQ of width k and each of the ϕ_i is an answer-guarded formula. According to Lemma 21, $\pi : (N, \bar{W}) \mathrel{\widetilde{\to}} (M, \pi(\bar{W}))$ is a guarded bisimilar cover allowing for treeifications of width k. For each $1 \leq i \leq s$ let $T_i = \{\bar{b} \mid (N, \bar{W}) \models \eta(\phi_i)(\bar{b})\}$ be the relation defined by $\eta(\phi_i)$ on $(N, \bar{W})$. As ϕ_i is answer-guarded, T_i is a guarded relation on $(N, \bar{W})$ hence also on N and, by guarded-bisimulation invariance of $\eta(\phi_i)$, for all guarded tuples $\bar{b}$ of N we have $(N, \bar{W}) \models \eta(\phi_i)(\bar{b}) \iff (M, \pi(\bar{W})) \models \eta(\phi_i)(\pi(\bar{b}))$. It follows that each T_i is a π-saturated guarded relation and so, by Lemma 21 again, $\pi : (N, \bar{T}) \mathrel{\widetilde{\to}} (M, \pi(\bar{T}))$ is a guarded bisimilar cover allowing for treeifications of width k. Therefore, as q has width k, for every guarded tuple $\bar{b}$ of $(N, \bar{T})$ we have that $(N, \bar{T}) \models q(\bar{b}) \leftrightarrow \Lambda_q^{\tau'}(\bar{b})$, where $\tau' = \tau \uplus \{U_1, \ldots, U_s, C_k\}$. All in all we have $(N, \bar{W}) \models q[\eta(\bar{\phi})/\bar{U}](\bar{b}) \leftrightarrow \Lambda_q^{\tau'}[\eta(\bar{\phi})/\bar{U}](\bar{b})$ for all guarded $\bar{b}$. Finally, by the induction hypothesis, for each ϕ_i and $\bar{b}$ a guarded tuple $(N, \bar{W}) \models \phi_i(\bar{b}) \leftrightarrow \eta(\phi_i)(\bar{b})$, hence $(N, \bar{W}) \models \phi(\bar{b}) \leftrightarrow q[\bar{\phi}/\bar{U}](\bar{b}) \leftrightarrow q[\eta(\bar{\phi})/\bar{U}](\bar{b}) \leftrightarrow \Lambda_q^{\tau'}[\eta(\bar{\phi})/\bar{U}](\bar{b}) \leftrightarrow \eta(\phi)(\bar{b})$, as needed.

Consider now the case of $\phi = \mu_{Z,\bar{z}}[\psi(\bar{Y}, Z, \bar{z})]$ and, accordingly, $\eta(\phi) = [\mu_{Z,\bar{z}}; \eta(\psi)(\bar{Y}, Z, \bar{z})]$. Let $(U^\alpha)_\alpha$ and $(V^\alpha)_\alpha$ be the relations obtained at the respective transfinite stages of the inductive fixpoint computation of $\psi(\bar{z})$ and by $\eta(\psi)(\bar{z})$, respectively. In other words, $U^0 = V^0 = \emptyset$,

$$U^{\alpha+1} = \{\bar{a} \mid (N, \bar{W}, \breve{U}^\alpha) \models \psi(\bar{a})\} \quad \text{and} \quad V^{\alpha+1} = \{\bar{a} \mid (N, \bar{W}, \breve{V}^\alpha) \models \eta(\psi)(\bar{a})\}$$

for all ordinals α, moreover, $U^\lambda = \bigcup_{\alpha<\lambda} U^\alpha$ and $V^\lambda = \bigcup_{\alpha<\lambda} V^\alpha$ for limit ordinals λ. Here $\breve{U}^\alpha = U^\alpha \cap \mathrm{guarded}(N)$ and $\breve{V}^\alpha = V^\alpha \cap \mathrm{guarded}(N)$ denote the *guarded interior* of U^α and V^α, respectively.

Observe that guarded-bisimulation invariance of $\eta(\psi)$ implies that $\breve{V}^\alpha = \pi^{-1}(\pi(\breve{V}^\alpha)) \cap \mathrm{guarded}(N)$ for all ordinals α. So assuming $\breve{U}^\alpha = \breve{V}^\alpha$ for some α, $\breve{U}^\alpha$ is π-saturated and thus we can apply the induction hypothesis of the claim of this lemma for the structurally simpler formula ψ and the assignment mapping $\bar{Y}$ to $\bar{\breve{W}}$ and Z to $\breve{U}^\alpha$ to establish $\breve{U}^{\alpha+1} = \breve{V}^{\alpha+1}$. Thus it follows by transfinite induction (the limit case being entirely trivial) that $\breve{U}^\alpha = \breve{V}^\alpha$ for all ordinals α, whence also $\bigcup_\alpha \breve{U}^\alpha = \bigcup_\alpha \breve{V}^\alpha$. In other words, the guarded interiors of the least fixpoints defined by ϕ and by $\eta(\phi)$ on $(N, \bar{W})$ do coincide as claimed. □ □

Theorem 24. *A GNFP^k-sentence $\tilde{\varphi}$ in GN-normal form is satisfiable (in the finite) if, and only if, $\eta(\tilde{\varphi}) \in \mathrm{GFP}^k$ is satisfiable (in the finite).*

[3] recall that a guarded relation W on the cover N is π-saturated if $W = \pi^{-1}(\pi(W)) \cap \mathrm{guarded}(N)$

Proof. It is easy to see that for every model M of $\tilde{\varphi}$ its expansion (M, C_k) is a model of $\eta(\tilde{\varphi})$, where C_k is the complete k-ary relation on M. Indeed, for every positive subformula $\psi(\bar{x})$ as in (e) above and for every

$$M \models \psi[\phi_1/U_1, \ldots, \phi_s/U_s](\bar{a}) \tag{8}$$

there is a CQ $\exists \bar{y}\, Q(\bar{x}\bar{y})$, a disjunct of the CNF of ψ, such that Q is a conjunction of $\{U_1, \ldots, U_s\}$-atoms and $|\bar{x}| + |\bar{y}| \leq k$ and $M \models Q[\phi_1/U_1, \ldots, \phi_s/U_s](\bar{a}\bar{b})$ for some $\bar{b} \in M^{|\bar{y}|}$. Although Q need not be acyclic $\Lambda_{\psi}^{\tau \uplus \{U_1, \ldots, U_s, C_k\}}(\bar{x})$ does contain as a disjunct the acyclic conjunctive query $\exists \bar{y}\, C_k(\bar{x}\bar{y}) \wedge Q(\bar{x}\bar{y})$. Therefore, given the interpretation of C_k, we have

$$(M, C_k) \models \Lambda_{\psi}^{\tau \uplus \{U_1, \ldots, U_s, C_k\}}[\eta(\phi_1)/U_1, \ldots, \eta(\phi_s)/U_s](\bar{a}) \tag{9}$$

and the converse implication (9)⇒(8) holds by definition of treeification. Using the equivalence of (8) and (9) it is straightforward to verify by induction on formulas that $(M, C_k) \models \eta(\varphi)(\bar{a})$ iff $M \models \varphi(\bar{a})$ for all $\bar{a}$.

Conversely, consider some M a model of $\eta(\tilde{\varphi})$ and its (wk^w)-th Rosati cover $M^{(wk^w)}$, equally a model of $\eta(\tilde{\varphi})$, where w is the maximum of the width of τ and k. Lemma 23 proves that $M^{(wk^w)}$ is, in fact, a model of $\tilde{\varphi}$, and we know from Theorem 20 that if M is finite then so is $M^{(wk^w)}$. □ □

Both satisfiability [GW99] and finite satisfiability [BB12] of GFP sentences have been shown decidable in time $2^{\mathcal{O}(nw^w)}$, where n is the length of the input formula and w is its width[4]. Starting with a GNFP^k sentence φ whose GN-normal form is $\tilde{\varphi}$, we get from Lemma 22 that $|\eta(\tilde{\varphi})| = 2^{(k|\varphi|)^{\mathcal{O}(1)}}$ and that $\eta(\tilde{\varphi})$ is computable within that same time bound, but its width remains k. Theorem 11 now follows from these bounds via Theorem 24.

5 Model checking for GNFO and GNFP

In this section we study the combined complexity of the model checking problems for GNFO and GNFP, where the input consists of a sentence and a structure and the goal is to decide whether the sentence is true on the structure. For the *unary negation* cases, it was shown in [CS13] that the model checking problem for UNFO is $\mathrm{P}^{\mathrm{NP}[O(\log^2 n)]}$-complete, and that the model checking problem for UNFP is in $\mathrm{NP}^{\mathrm{NP}} \cap \mathrm{coNP}^{\mathrm{NP}}$ and P^{NP}-hard. We show that these upper-bounds also apply to GNFO and GNFP. The proof is a reduction to formulas with unary negations by constructing an incidence structure.

Theorem 25. *The model checking problem for* GNFO *is* $\mathrm{P}^{\mathrm{NP}[O(\log^2 n)]}$*-complete.*
For GNFP *it is in* $\mathrm{NP}^{\mathrm{NP}} \cap \mathrm{coNP}^{\mathrm{NP}}$ *and hard for* P^{NP}.

Proof. The lower bounds are immediate as UNFO is a fragment of GNFO and UNFP is a fragment of GNFP. For the upper bounds we reduce the model checking problem for GNFO (resp. GNFP) to the model checking problem for UNFO (resp. UNFP).

Given a relational structure M and a sentence ϕ of GNFP we construct in polynomial time a relational structure M' and a sentence ϕ' of UNFP such that ϕ' is in UNFO if ϕ is in GNFO and

$$M \models \phi \text{ iff } M' \models \phi'. \tag{10}$$

The structure M' is an extension of M essentially representing M together with its incidence structure. M' contains one new element per fact of M. For each relation symbol R of the signature of M, we add a new unary symbol P_R interpreted as the set of all facts of M involving the relation R. Finally we add a new binary relation symbol $E_{R,i}$ for each relation R of the signature of M and number i between 1 and the arity of R, interpreted as the binary relation that relates each new element y denoting a fact $R(\bar{x})$ of M to x_i. The construction of M' is clearly in polynomial time.

[4]The *width* as defined in [GW99, BB12] is the maximal number of free variables occurring in a subformula. This number is of course bounded by the width as defined in this paper.

When ϕ is in GNFO, the formula ϕ' is constructed from ϕ by first replacing each subformula $R(\bar{z}) \wedge \neg\psi(\bar{x})$, where $\bar{x} \subseteq \bar{z}$ with:

$$\exists y P_R(y) \wedge \bigwedge_i E_{R,i}(y, z_i) \wedge \neg\big(\exists \bar{x} \bigwedge_i E_{R,\alpha_i}(y, x_i) \wedge \psi(\bar{x})\big)$$

where α_i is such that $x_i = z_{\alpha_i}$.

In the case where ϕ is in GNFP we further extend the structure constructed above by adding the following for each subformula $\xi(\bar{z})$ occurring in ϕ and of the form $\beta(\bar{z}) \wedge Z(\bar{x})$, where Z is a fixpoint predicate variable (recall that according to our syntactic restrictions this means $\bar{x} \subseteq \bar{z}$): we have a new unary predicate P_ξ interpreted with new elements, one per fact of M in β. Finally, for each i between 1 and the arity of Z, we have a new binary relation $E_{\xi,i}$ interpreted as the pairs (v, u_j) where v represents the fact $\beta(\bar{u})$ and j is such that z_j is the variable in position i within $Z(\bar{x})$. This concludes the construction of M'.

For the construction of ϕ', we first do as in the GNFO case. Moreover, we have one extra unary fixpoint predicate P_z per fixpoint predicate Z occurring in ϕ and we replace each subformula $\xi(\bar{z})$ of the form $\beta(\bar{z}) \wedge Z(\bar{x})$ by

$$\exists y P_\xi(y) \wedge P_Z(y) \wedge \bigwedge_i E_{\xi,i}(y, x_i)$$

and each fixpoint subformula $\mu_{Z,\bar{z}}[\,\phi(\bar{Y}, Z, \bar{z})\,](\bar{x})$ by

$$\exists y \;\; \mu_{P_Z,z}[\,\phi(\overline{P_Y}, P_Z, z)\,](y) \wedge \bigvee_\xi \big(P_\xi(y) \wedge \bigwedge_i E_{\xi,i}(y, x_i)\big).$$

In both cases the construction of ψ' and M' are clearly in polynomial time. The reader can now verify that (10) holds. □

Remark 26. *If we had opted for the alternative syntax that does not explicitly declare any concrete guard but instead uses the clause* $\texttt{guarded}_\tau$ *as in* (7)*, then the complexity of the model checking problem would be higher. Indeed it is not difficult to show that in this case it becomes* ExpTime*-complete. The reason is that if the maximal arity of the relational predicates of the signature is k then there are exponentially many, in k, potential guards (in other words predicates P_ξ in the construction above, accounting for the number of permutations of variables in a k-ary atom).*

With this in mind we encode the model checking problem for a single-rule Datalog program (SIRUP) into the model checking problem of this alternative syntax of GNFP*. The former is known to be* ExpTime*-complete [GP03]. Given a structure M and a SIRUP ϕ we introduce a new extra relation whose arity is the number of elements of M and containing a single tuple that enumerates all elements of M. The fixpoint formula ϕ is then trivially guarded by this new relation (no need to guard negations because there is no negation in Datalog).*

For the upper-bound, given a structure M and a sentence ϕ of GNFP*, we first compute in a new relation, in exponential time, all guarded tuples of M and then evaluate ϕ as in Theorem 25. The total time is exponential, for the algorithm underlying Theorem 25 is polynomial in the size of M.*

6 Expressive power of GNFO and GNFP

In this section, we develop an appropriate notion of bisimulation for GNFO and GNFP, and use it to characterize the expressive power of GNFO.

Recall the notions of guarded tuples and the notation *guarded*(M) from Section 2. For a number k, we say that a tuple is *k-guarded* if it is guarded by a fact of M using at most k elements of M. We denote by *guarded*$^k(M)$ the set of all k-guarded tuples of M.

Definition 27. *Let M, N be two structures. A GN-bisimulation (resp. a GN-bisimulation of width $k \geq 1$) is a binary relation $Z \subseteq$ guarded$(M) \times$ guarded(N) (resp. $Z \subseteq$ guarded$^k(M) \times$ guarded$^k(N)$) such that the following hold for every pair $(\bar{a}, \bar{b}) \in Z$, where $\bar{a} = a_1, \ldots, a_m$ and $\bar{b} = b_1, \ldots, b_n$*

- *$(M, \bar{a})$ and $(N, \bar{b})$ are locally isomorphic (and in particular, $m = n$)*

- *[**Forward clause**] For every finite set $X \subseteq dom(M)$ (resp. $X \subseteq dom(M)$ and $|X| \leq k$) there is a partial homomorphism $h : M \to N$ whose domain is X, such that $h(a_i) = b_i$ for all a_i in X, and such that for every $\bar{a}' \in guarded(M)$ (resp. $\bar{a}' \in guarded^k(M)$) consisting of elements in the domain of h, the pair $(\bar{a}', h(\bar{a}'))$ belongs to Z.*

- *[**Backward clause**] Likewise in the other direction, where $X \subseteq dom(N)$.*

Note that if X above is restricted to guarded sets then we obtain a definition of guarded bisimulation. We write $M \approx_{GN} N$ if there is a non-empty GN-bisimulation between M and N and write $M \approx^k_{GN} N$ if the GN-bisimulation has width k. Notice that $M \approx_{GN} N$ implies that $M \approx^k_{GN} N$ for all k.

It is not difficult to see that the existence of a GN-bisimulation implies indistinguishability by GNFP sentences, and that the existence of a GN-bisimulation of width k implies indistinguishability in GNFP^k.

Proposition 28. *For $k \geq 1$, if $M \approx^k_{GN} N$ then M and N satisfy the same* GNFP^k *sentences. In particular, if $M \approx_{GN} N$ then M and N satisfy the same* GNFP *sentences.*

Proof. The proof is by induction on the nesting of fixpoints and existential quantification in the formula. We assume without loss of generality that all formulas are in GN-normal form. It is convenient to state the induction hypothesis for formulas $\phi(\bar{x})$ with several free second-order variables. The induction hypothesis then becomes: for all formulas $\phi(\bar{x}, \bar{Y})$, and for all GN-bisimulation Z of width k between $(M, \bar{P})$ and $(N, \bar{Q})$, and all pair $(\bar{a}, \bar{b}) \in Z$, we have $(M, \bar{P}, \bar{a}) \models \phi$ iff $(N, \bar{Q}, \bar{b}) \models \phi$. We show only the important cases of the inductive step. Let M, N be two structures, Z be a GN-bisimulation of width k between M and N, $\bar{P}$ and $\bar{Q}$ be valuations of $\bar{Y}$ respectively on M and N, and $(\bar{a}, \bar{b}) \in Z$.

- $\phi(\bar{x}, \bar{Y})$ starts with an existential quantifier. Then, by definition of GN-normal form, ϕ is of the form $q[\varphi_1/U_1, \ldots, \varphi_s/U_s]$ for some UCQ q and each φ_i is an answer-guarded formula also of the form $\varphi_i(\bar{y}, \bar{Y})$. Let $z_1, \ldots, z_n$ be the existentially quantified variables of q and let $m = |\bar{x}|$. In particular $m + n \leq k$.

  First, suppose $(M, \bar{P}, \bar{a}) \models \phi$. Let $\{c_1, \ldots, c_n\}$ be the quantified elements of M witnessing the truth of ϕ and let $X = \{c_1, \ldots, c_n\} \cup \{a_1, \cdots, a_m\}$. By the definition of GN-bisimulation, there is a partial homomorphism $h : M \to N$ of domain X such that $h(\bar{a}) = \bar{b}$ and such that $(\bar{u}, h(\bar{u})) \in Z$ for all k-guarded tuple $\bar{u} \subseteq X$. For each i, let $\bar{u}_i$ be the subset of X making φ_i true on $(M, \bar{P})$. As φ_i is answer-guarded and belongs of GNFO^k, $\bar{u}_i$ is k-guarded. Therefore, as by induction hypothesis, $(M, \bar{P}, \bar{u}_i) \models \varphi_i(\bar{y}, \bar{Y})$ iff $(N, \bar{Q}, h(\bar{u}_i)) \models \varphi_i(\bar{y}, \bar{Y})$, we have $(N, \bar{Q}, h(\bar{u}_i)) \models \varphi_i(\bar{y}, \bar{Y})$. Hence the assignment that sends $z_1, \ldots, z_n$ to $h(c_1), \ldots, h(c_n)$ makes ϕ true on $(N, \bar{Q}, \bar{b})$.

  The opposite direction, from $(N, \bar{Q}, \bar{b}) \models \phi$ to $(M, \bar{P}, \bar{a}) \models \phi$, is symmetric.

- $\phi(\bar{x}, \bar{Y})$ is any Boolean combination of formulas of the form $\psi(\bar{y}, \bar{Y})$, the result is immediate from the induction hypothesis.

- $\phi(\bar{x}, \bar{Y})$ is of the form $\mu_{Z,\bar{z}}[\,\psi(Z, \bar{Y}, \bar{z})\,](\bar{x})$. We proceed by induction on the fixpoint iterations.

  Let $\mathcal{O}_{\psi,(M,\bar{P})}$ and $\mathcal{O}_{\psi,(N,\bar{Q})}$ be the monotone set-operations induced by ψ on subsets of the domain of $(M, \bar{P})$ and $(N, \bar{Q})$, respectively, and let $\kappa = \max\{|M|, |N|\}$. Recall that the least fixpoint of $\mathcal{O}_{\psi,(M,\bar{P})}$ is equal to ${\mathcal{O}_{\psi,(M,\bar{P})}}^{\kappa}(\emptyset)$, and similarly for the least fixpoint of $\mathcal{O}_{\psi,(N,\bar{Q})}$. A straightforward transfinite induction shows that, for all ordinals λ, and for all $(\bar{a}, \bar{b}) \in Z$, $\bar{a} \in {\mathcal{O}_{\psi,(M,\bar{P})}}^{\lambda}(\emptyset)$ if and only if $\bar{b} \in {\mathcal{O}_{\psi,(N,\bar{Q})}}^{\lambda}(\emptyset)$. We conclude that $(M, \bar{P}, \bar{a}) \models \mu_{Z,\bar{z}}[\,\psi(Z, \bar{Y}, \bar{z})\,](\bar{x})$ if and only if $(N, \bar{Q}, \bar{b}) \models \mu_{Z,\bar{z}}[\,\psi(Z, \bar{Y}, \bar{z})\,](\bar{x})$.

□

In fact, over arbitrary structures, GN-bisimulation invariance can be used to *characterize* GNFO.

Theorem 29. GNFO *is the $\approx_{GN}$-invariant fragment of* FO*, and for all $k \geq 1$,* GNFO^k *is the $\approx^k_{GN}$-invariant fragment of* FO *on arbitrary structures.*

The finite variant of Theorem 29, showing that GNFO^k captures the $\approx^k_{GN}$-invariant fragment of FO on finite structures has recently been established in [Ott12].

Proof. We prove the hard direction, which uses the technique of ω-saturated structures from classical model theory (cf. [Hod93]). We give the proof for the case of GNFO^k. The argument for full GNFO is identical. Let ϕ be any sentence of FO invariant under GN-bisimulations of width k. We want to show that ϕ is equivalent to a GNFO^k-sentence. By a well known argument using Compactness, it is enough to show that, whenever two structures agree on all formulas of GNFO^k, they agree on ϕ. Hence, suppose M and N satisfy the same sentences of GNFO^k. Without loss of generality we can assume that M and N are ω-saturated. Define $Z \subseteq \mathit{guarded}^k(M) \times \mathit{guarded}^k(N)$ as the set of all k-guarded pairs $(\bar{a}, \bar{b})$ such that $(M, \bar{a})$ and $(N, \bar{b})$ satisfy the same GNFO^k-formulas. We claim that Z is a non-empty GN-bisimulation of width k. As ϕ is invariant under GN-bisimulations of width k this implies that M and N agree on ϕ and concludes the proof of the lemma.

That Z is a GN-bisimulation of width k follows immediately from the following lemma where we write $(M, \bar{a}) \equiv_{\text{GNFO}^k} (N, \bar{b})$ (resp. $(M, \bar{a}) \equiv_{\text{GNFO}} (N, \bar{b})$) if for all $\phi \in$ GNFO^k (resp. $\phi \in$ GNFO) we have $M \models \phi(\bar{a})$ iff $N \models \phi(\bar{b})$:

Lemma 30. *For all ω-saturated structures M and N the following hold.*

1. *The relation $\{(\bar{a}, \bar{b}) \in \mathit{guarded}(M) \times \mathit{guarded}(N) \mid (M, \bar{a}) \equiv_{\text{GNFO}} (N, \bar{b})\}$ is a GN-bisimulation.*
2. *The relation $\{(\bar{a}, \bar{b}) \in \mathit{guarded}^k(M) \times \mathit{guarded}^k(N) \mid (M, \bar{a}) \equiv_{\text{GNFO}^k} (N, \bar{b})\}$ is a GN-bisimulation of width k.*

Proof. We prove the second claim. The proof of the first claim is similar. Let $Z = \{(\bar{a}, \bar{b}) \in \mathit{guarded}^k(M) \times \mathit{guarded}^k(N) \mid (M, \bar{a}) \equiv_{\text{GNFO}^k} (N, \bar{b})\}$. Clearly, Z consists of locally isomorphic pairs of tuples. We show that Z satisfies the forward clause, the proof of the backward clause is analogous.

Suppose $(\bar{c}, \bar{d}) \in Z$ and let $X \subseteq \mathit{dom}(M)$ with $|X| \leq k$. For simplicity, assume $\bar{c} \subseteq X$ (the general case is similar). Thus, let $X = \{c_1, \cdots c_l, c_{l+1}, \ldots, c_n\}$ with $\bar{c} = (c_1, \cdots, c_l)$ and $n \leq k$. Let $T[x_1, \ldots, x_n]$ be the set of all formulas $\phi(x_1, \ldots, x_n)$ that are positive Boolean combinations of (i) atomic formulas or (ii) formula of the form $\alpha(\bar{y}) \wedge \neg\psi(\bar{y})$ where ψ is in GNFO^k and α is an atomic formula (possibly an equality statement), and that are true in $(M, c_1, \ldots, c_n)$. We view T as an n-type with l parameters. It is not hard to see that every finite subset $T' \subseteq T$ is realized in N under some assignment that sends $(x_1, \cdots, x_l)$ to $\bar{d}$. Indeed, notice that $\exists x_{l+1} \ldots x_n(\bigwedge T')$ is a formula of GNFO^k true at $(M, \bar{c})$ and therefore it is also true at $(N, \bar{d})$ by hypothesis. Since N is ω-saturated (and treating T as an n-type with parameter $\bar{d}$), it follows that the entire set $T[x_1, \ldots, x_n]$ is realized in N under an assignment g that sends $(x_1, \cdots, x_l)$ to $\bar{d}$. Let h be the mapping sending c_i to $g(x_i)$. As T contains all atomic formulas, then h is a homomorphism. Moreover, as T contains all formula of the form $\alpha(\bar{y}) \wedge \neg\psi(\bar{y})$, for all $\bar{c}'$ in $\mathit{guarded}^k(M)$ with $\bar{c}' \subseteq X$ we have $(M, \bar{c}') \equiv_{\text{GNFO}^k} (N, h(\bar{c}'))$. By definition of Z this implies $(\bar{c}', h(\bar{c}')) \in Z$. □ □

That Z is non-empty follows from ω-saturation: consider $\bar{a} \in \mathit{guarded}^k(M)$ and let $\Sigma(\bar{x})$ be the set of all GNFO^k formulas true for $\bar{a}$. Every finite subset $\Sigma' \subseteq \Sigma(\bar{x})$ is realized in N (notice that $\exists \bar{x} \bigwedge \Sigma'$ is a sentence of GNFO^k that is true in M, and hence in N). Therefore, by ω-saturation, the entire set $\Sigma(\bar{x})$ is realized by an element $\bar{b}$ in N, and hence $(\bar{a}, \bar{b}) \in Z$, which implies that Z is non-empty. □ □

Based on the definition of GN-bisimulation (of width k) it is straightforward to define GN-unraveling (of width k) as an operation constructing from any given structure M a (k-acyclic) companion $M^* \equiv_{\text{GNFO}^k} M$. This provides a natural route for demonstrating the tree-like model property of GNFP. We leave this as an exercise, instead, take a short-cut via the reduction to the guarded fragment introduced in Section 4.

Theorem 31. GNFP *has the tree-like model property.*

Proof. Consider a model M of a GNFP^k-sentence φ. Recall Section 4 and the reduction from GNFP to GFP. We can assume without loss of generality that M contains the relation C_k containing all k tuples on M. In this case M is also a model for $\eta(\varphi)$, where $\eta(\varphi)$ is the GFP formula constructed in section 4. Let M^* be the guarded unraveling of M of width k (cf. e.g. [Gr] for the relevant definitions). It is straightforward to verify that M^* is a k-guarded bisimilar cover of M, that M^* has tree width at most k, and that

M^* is acyclic (in particular, weakly l-acyclic as a cover of M for every $l \in \mathbb{N}$). Therefore, by Lemma 23 we have $M^* \models \eta(\varphi)$ iff $M^* \models \varphi$. By guarded bisimulation we also have $M^* \models \eta(\varphi)$ iff $M \models \eta(\varphi)$. Altogether this shows that $M^* \models \varphi$. □ □

7 Further extensions

7.1 Clique-Guarded Negation

We now consider a further generalization of GNFO called CGNFP, taking inspiration from the clique-guarded fragment. CGNFP is defined just like GNFP except that we allow clique-guards in the place of guards. We say that a conjunction of atomic formulas α *clique-guards* $\bar{x}$ if for every pair $x_i, x_j \in \bar{x}$, α includes a conjunct in which both x_i and x_j appear (in other words, the co-occurrence graph for the variables in $\bar{x}$ is a clique).

The formulas of CGNFP are generated by the following grammar:

$$\phi \quad ::= \quad R(\bar{x}) \mid \text{x=y} \mid \alpha(\bar{x},\bar{y}) \wedge X(\bar{x}) \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \exists x\, \phi \mid \alpha(\bar{x},\bar{y}) \wedge \neg\phi(\bar{x}) \mid \mu_{Z,\bar{z}}[\phi(\bar{Y},Z,\bar{z})\,](\bar{x})$$

where $\alpha(\bar{x},\bar{y})$ is a conjunction of atoms that clique-guards $\bar{x}$. The fixpoint-free fragment of CGNFP is called CGNFO, and we use the notation CGNFP$[\tau]$ or CGNFO$[\tau]$ when restricting to formulas in a particular signature τ. As in the case of GNFO and GNFP, in the above inductive definition, we can equivalently replace the clauses for conjunction, disjunction and existential quantification by a single clause for unions of conjunctive queries. As in the case of GNFP this provides a normal form that is used to define formulas *of width* k, denoted CGNFP^k.

In this section, we show that CGNFO and CGNFP behave similarly to GNFO and GNFP, in terms of the complexity of satisfiability, and in terms of the finite model property. To prove this, we will make use the fact that the n-th Rosati cover $M^{(n)}$ of a structure M is n-*conformal* (cf. Theorem 20). Recall that we call a structure M n-*conformal* (where $n \geq 1$) if every n-tuple that is clique-guarded in M is in fact guarded in M. Using this fact, it turns out that our results for GNFO and GNFP can be lifted to CGNFO and CGNFP without much effort.

Theorem 32. *1. The satisfiability problem for* CGNFO *and for* CGNFP *is* 2ExpTime*-complete.*

2. CGNFO *has the finite model property.*

Theorem 32 generalizes prior decidability results for the *loosely guarded fragment* [vB97], the *packed fragment* [Mar99], and the *clique-guarded fragment* [Gr], which are all subsumed in CGNFO (for sentences). The finite model property for these fragments was first established in [Hod02].

In the remainder of this section, we explain how Theorem 32 is proved.

We extend our translation from GNFP into GFP to a translation from CGNFP into GFP. With a slight abuse of notation, we use the same symbol η also to denote this extension of the translation. For every $k \geq 1$, and for every relational signature τ consisting of at most k-ary relations, we define a translation η from CGNFP$^k[\tau]$ formulas in normal form to GFP$^k[\tau \uplus \{C_k\}]$ formulas, where C_k is a new symbol of arity k, by structural recursion, using the following rules.

$$\begin{aligned}
\eta(\, R(\bar{x})\,) &= R(\bar{x}) && (a)\\
\eta(\, \alpha(\bar{x}\bar{y}) \wedge Z(\bar{x})\,) &= \alpha(\bar{x}\bar{y}) \wedge Z(\bar{x}) && (b)\\
\eta(\, \alpha(\bar{x}\bar{y}) \wedge \neg\psi(\bar{x})\,) &= \alpha(\bar{x}\bar{y}) \wedge \neg\, \eta(\psi(\bar{x})) && (c)\\
\eta(\, \mu_{Z,\bar{z}}[\psi(\bar{Y},Z,\bar{z})\,]\,) &= \mu_{Z,\bar{z}}[\eta(\psi(\bar{Y},Z,\bar{z}))\,] && (d)\\
\eta(\, q[\phi_1/U_1,\ldots,\phi_s/U_s]\,) &= \Lambda_q^{\tau\uplus\{U_1,\ldots,U_s,C_k\}}[\eta(\phi_1)/U_1,\ldots,\eta(\phi_s)/U_s] && (e)
\end{aligned}$$

where in (e) q is a UCQ of signature $\{U_1,\ldots,U_s\}$ disjoint from $\tau \uplus \{\bar{Y}, C_k\}$ and $\phi_1,\ldots,\phi_s \in$ GNFP$^k[\tau \uplus \{\bar{Y}\}]$, where $\bar{Y}$ enumerates the free fixpoint variables occurring in any of the ϕ_i's, each ϕ_i being a answer-clique-guarded formula, and such that $q[\phi_1/U_1,\ldots,\phi_s/U_s]$ is a subformula of $\tilde{\varphi}$.

Recall Theorem 24, which states that the translation $\eta(\cdot)$ from GNFP to GFP is satisfiability preserving. The proof involved passing from a structure M to its (wk^w)-th Rosati cover $M^{(wk^w)}$, which is (wk^w)-acyclic, and applying Lemma 23. Lemma 23, in turn, was proved by induction, where the inductive hypothesis was stated in terms of guarded tuples and guarded relations. Using the (wk^w)-conformality of $M^{(wk^w)}$, which gives us that every clique-guarded (wk^w)-tuple is in fact a guarded tuple, the same arguments apply for clique guarded tuples and therefore when the input is in CGNFP. In this way, we get the following analogue of Theorem 24.

Theorem 33. *A* CGNFP^k*-formula $\tilde{\varphi}$ in normal form is satisfiable (in the finite) if, and only if, the* GFP^k*-formula $\eta(\tilde{\varphi})$ is satisfiable (in the finite).*

Moreover, the same complexity analysis applies as in the case of GNFP. In particular, Theorem 33 implies that satisfiability is 2Exp-complete for CGNFP. Furthermore, observe that the translation η maps CGNFO formulas to GFO formulas. Therefore, Theorem 33 also establishes the finite model property for CGNFO.

7.2 Constant Symbols

Although our definition of GNFO, and of GNFP, does not include constant symbols, they can be added without affecting any of our complexity results. This can be shown using the same technique that was used in [Grb] in the context of the guarded fragment. For the sake of completeness, we explain here how this technique can be applied to CGNFP-sentences (the same argument works also for CGNFP-formulas with free variables).

By a CGNFP-*sentence with constants* we mean a CGNFP-formula where, in addition, constant symbols may freely be used in atomic subformulas (and there is no restriction on the use of constant symbols in negated subformulas).

Proposition 34. *Given any* CGNFP*-sentence ϕ with constant symbols, we can construct in polynomial time a* CGNFP*-sentence ϕ' without constant symbols, such that ϕ and ϕ' are equi-satisfiable, both in the finite and over arbitrary structures.*

Proof. Let ϕ be any CGNFP-sentence over a signature $\sigma = \{R_1, \ldots, R_n, c_1, \ldots, c_k\}$. Consider the relational signature $\tau = \{R'_1, \ldots, R'_n\}$, where the arity of each new relation symbol R'_i is the arity of R_i plus k. Fix fresh variables $z_1, \ldots, z_k$ corresponding to the constants $c_1, \ldots, c_k$. Finally, let $\phi^*(z_1, \ldots, z_k)$ be the CGNFP-formula over τ obtained from ϕ by (i) replacing every occurrence of a constant symbol c_i by the corresponding variable z_i, and, subsequently, (ii) replacing every relational atomic formula $R_j(x_1, \ldots, x_m)$ by $R_j(x_1, \ldots, x_m, z_1, \ldots, z_k)$. Note that $\phi^*(z_1, \ldots, z_k)$ is indeed a CGNFP-formula. This follows from the fact that ϕ was a CGNFP-formula, and the fact $z_1, \ldots, z_k$ occur in every atomic subformula of $\phi^*(z_1, \ldots, z_k)$.

We show that ϕ and $\phi^*(z_1, \ldots, z_k)$ are equi-satisfiable, both in the finite and in the infinite.

Indeed, every model M of ϕ gives rise to a strucure M' over the same domain and such that for each relation symbol $R_i \in \sigma$ we have $R_i'^{M'} = R_i^M \times \{c_1^M\} \times \cdots \times \{c_k^M\}$. It is then immediate to check that $M' \models \phi^*(c_1^M, \ldots, c_k^M)$.

Conversely, if $M' \models \phi^*(u_1, \ldots, u_k)$ for some $u_1, \ldots, u_k \in \mathit{dom}(M')$, then $M \models \phi$, where, for each m-ary relation symbol R_i, $R_i^M = \{(a_1, \ldots, a_m) \mid (a_1, \ldots, a_m, u_1, \ldots, u_k) \in R_i'^{M'}\}$ and and where $c_j^M = u_j$.

Finally, we can turn ϕ^* into a sentence by existentially quantifying out $z_1, \ldots, z_k$. □

The same argument also applies to the model checking problem (as the construction of the model M' from M given in the above proof is polynomial).

8 Discussion

We have provided a logical framework generalizing both GFO and UNFO while preserving their nice properties, in particular decidability of satisfiability. Our results on satisfiability carry over to the validity

and entailment problems for GNFO, and likewise for GNFP, as these problems are all reducible to each other. For instance, a GNFO entailment $\phi(\bar{x}\bar{y}) \models \psi(\bar{x}\bar{z})$ holds if, and only if, for a fresh relation R of appropriate arity $\exists\bar{x}\bar{y}\bar{z}(\phi(\bar{x}\bar{y}) \wedge R(\bar{x}\bar{y}\bar{z}) \wedge \neg\psi(\bar{x}\bar{z})))$ is not satisfiable.

Another immediate consequence of our results is that query answering for unions of conjunctive queries with respect to guarded-negation fixpoint theories (i.e., the analogue of Theorem 6 replacing GFO by GNFP) is decidable and 2ExpTime-complete.

It would be tempting to further generalize by including the two variable fragment of FO (FO^2). Unfortunately this would lead to undecidability. Actually a simple combination of FO^2 with UNFO already yields undecidability as FO^2 can express the fact that a relation correspond to inequality (by $\forall x, y\ (R(x, y) \leftrightarrow x \neq y)$) and the extension of UNFO with inequality is undecidable [CS13]. Similarly, unconstrained universal quantification leads to undecidability, since every subformula of the form $\neg\psi(\bar{x})$ can be trivially guarded using a fresh relation $R(\bar{x})$, adding $\forall\bar{x}\ R\bar{x}$ as a conjunct to the main formula.

Since the publication of the conference proceedings version of the present paper, a number of new results and applications of guarded-negation logics have been established. In particular, in [BtO12], it was shown that boundedness is decidable for guarded-negation datalog; in [BBtC13], GNFO was shown to satisfy Craig interpolation as well as various model-theoretic preservation theorems; and in [BBtC13] and [BtCLW13], open-world query answering and query rewritability were studied for database queries and constraints specified in GNFO.

References

[ABN98] Hajnal Andréka, Johan van Benthem, and István Németi. Modal languages and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27:217–274, 1998.

[AN01] André Arnold and Damian Niwinski. *Rudiments of mu-calculus*. Elsevier, 2001.

[BB12] Vince Bárány and Mikołaj Bojańczyk. Finite satisfiability for guarded fixpoint logic. *Information Processing Letters*, 112(10):371–375, 2012.

[BBtC13] Vince Bárány, Michael Benedikt, and Balder ten Cate. Rewriting guarded negation queries. In *Intl. Symp. on Mathematical Foundations of Computer Science (MFCS)*, 2013.

[BG01] Dietmar Berwanger and Erich Grädel. Games and model checking for guarded logics. In *Logic for Programming, Artificial Intelligence, and Reasoning (LPAR)*, 2001.

[BGO14] Vince Bárány, Georg Gottlob, and Martin Otto. Querying the guarded fragment. *Logical Methods in Computer Science (LMCS)*, 10(2), 2014. Preliminary version in Proc. Symp. on Logic In Computer Science (LICS), 2010.

[Boj03] Mikołaj Bojańczyk. The finite graph problem for two-way alternating automata. *Theoretical Computer Science*, 3(298):511–528, 2003.

[BtCLW13] Meghyn Bienvenu, Balder ten Cate, Carsten Lutz, and Frank Wolter. Ontology-based data access: a study through disjunctive datalog, csp, and mmsnp. In *Symp. on Principles of Database Systems (PODS)*, 2013.

[BtCS11] Vince Bárány, Balder ten Cate, and Luc Segoufin. Guarded negation. In *International Conference on Automata, Languages and Programming (ICALP)*, volume 6756 of *Lecture Notes in Computer Science*, pages 356–367. Springer, 2011.

[BtO12] Vince Bárány, Balder ten Cate, and Martin Otto. Queries with guarded negation. In *Intl. Conf. on Very Large Databases (VLDB)*, 2012.

[CS13] Balder ten Cate and Luc Segoufin. Unary negation. *Logical Methods in Computer Science (LMCS)*, 9, 2013.

[DG02] Anuj Dawar and Yuri Gurevich. Fixed point logics. *Bulletin of Symbolic Logic*, 1(8):65–88, 2002.

[FFG02] Jörg Flum, Markus Frick, and Martin Grohe. Query evaluation via tree-decompositions. *J. ACM*, 49(6):716–752, 2002.

[GOR99] Erich Grädel, Martin Otto, and Eric Rosen. Undecidability results on two-variable logics. *Arch. Math. Log.*, 38(4-5):313–354, 1999.

[GP03] Georg Gottlob and Christos H. Papadimitriou. On the complexity of single-rule datalog queries. *Information and Computation*, 183(1):104–122, 2003.

[Gr] Erich Grädel. Decision procedures for guarded logics. In *Intl. Conf. on Automated Deduction (CADE)*, 1999.

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

[Grä01] Erich Grädel. Why are modal logics so robustly decidable? In *Current Trends in Theoretical Computer Science*, pages 393–408. 2001.

[GW99] Erich Grädel and Igor Walukiewicz. Guarded fixed point logic. In *Symposium on Logic In Computer Science (LICS)*, 1999.

[Hod93] W. Hodges. *Model Theory*. Cambridge University Press, 1993.

[Hod02] Ian Hodkinson. Loosely guarded fragment of first-order logic has the finite model property. *Studia Logica*, 70(2):205–240, 2002.

[Mar99] Maarten Marx. Tolerance logic. *Journal of Logic, Language and Information*, 10:2001, 1999.

[Mor75] Michael Mortimer. On languages with two variables. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 21(8):135–140, 1975.

[Ott04] Martin Otto. Modal and guarded characterisation theorems over finite transition systems. *Ann. Pure Appl. Logic*, 130(1-3):173–205, 2004.

[Ott12] Martin Otto. Expressive completeness through logically tractable models. *Annals of Pure and Applied Logic*, 12:1418–1453, 2012.

[Var96] Moshe Y. Vardi. Why is modal logic so robustly decidable? In *Descriptive Complexity and Finite Models*, pages 149–184, 1996.

[vB97] Johan van Benthem. Dynamic bits and pieces. Technical Report LP-97-01, Institute for Logic, Language and Computation (ILLC), 1997.

[Yan81] Mihalis Yannakakis. Algorithms for acyclic database schemes. In *Proceedings of the seventh international conference on Very Large Data Bases - Volume 7*, VLDB '81, pages 82–94. VLDB Endowment, 1981.