CONSERVATIVE EXTENSIONS IN DECIDABLE FRAGMENTS OF FIRST-ORDER LOGIC

Based on joint work with Carsten Lutz, Jerzy Marcinkowski, Mauricio Martel, Hadrian Pulcini, Thomas Schneider, Frank Wolter
MOTIVATION

- **Conservative extensions** foundational notion in **mathematical logic**

- Theory $T_2$ is conservative extension of theory $T_1$ if $T_1 \subseteq T_2$ and every theorem in $T_2$ formulated in the language of $T_1$ is also in $T_1$

- Extension of a theory with a definition is always conservative, e.g.,

  extension of Presburger Arithmetic with new constant 5

- But useful in many other branches of computer science
  - software specification (e.g. formalization of refined systems)
  - theorem proving
  - **ontologies**
ONTOLOGIES

Logic-Based Representation of domain knowledge (concepts + relations)

- medicine, e.g., SNOMED Clinical Terms
- biology/genetics, e.g., Gene Ontology
- meteorology, geography, ...

Provide a common understanding of domain across different parties

Computer processable

Ontology ≡ set of logical formulas

100s of thousands of concepts (genes, alleles, diseases) and their relations (causes, part of, is_a, interacts_with, ...)
Ontologies are **dynamic objects**

- change of axioms
- addition of axioms
- merging/import of ontologies

**Module** of ontology $\mathcal{O}$ is a subset $\mathcal{M} \subseteq \mathcal{O}$

“suitable” to answer queries over restricted signature

⇒ **Need to be able to tell the “difference” between ontologies**


Notions of “difference”: **model-theoretic, deductive, w.r.t. queries+data**
- all depend on the ontology language
- give rise to decision problem
THE UNAVOIDABLE

decidable

undecidable
ONTOSOLOGY LANGUAGES

**Description Logics (DLs)** = family of knowledge representation languages
- basic member $\mathcal{ALC} = \text{notational variant of multi-modal logic } \mathcal{K}_n$
- extensions: inverses ($\mathcal{I}$), constants ($\mathcal{O}$), counting ($\mathcal{F}$, $\mathcal{Q}$), ..., e.g. $\mathcal{ALCQI}$
- Horn-DLs: $\mathcal{EL}$, $\mathcal{ELI}$, DL-Lite
  
  $\text{Mother} \equiv \text{Person} \sqcap \text{Female} \sqcap \exists \text{hasChild}. \top$

**Guarded Fragment (GF) and Two-Variable Fragment (FO$^2$)** generalize many DLs

\[
\forall x. (\text{Mother}(x) \leftrightarrow \text{Person}(x) \land \text{Female}(x) \land \exists y. \text{hasChild}(x, y))
\]

**Tuple-Generating Dependencies (TGD) or Existential Rules**

stem from database and dependency theory, also generalize many Horn DLs

\[
\forall x \forall y \left( \psi(x, y) \rightarrow \exists z \phi(x, z) \right) \text{ for conjunctive queries } \psi, \phi \quad \text{(CQ=FO( } \land, \exists))
\]
MODEL-THEORETIC CONSERVATIVE EXTENSIONS

\( \varphi, \psi \) are model-theoretically \( \Sigma \)-inseparable if
\[
\{ \mathcal{A} \mid \mathcal{A} \models \varphi \} = \{ \mathcal{A} \mid \mathcal{A} \models \psi \}
\]
\( \varphi \land \varphi' \) is model-theoretic CE of \( \varphi \) if \( \varphi, \varphi \land \varphi' \) are \( \text{sig}(\varphi) \)-inseparable.

\( \Sigma \)-inseparability only interesting if \( \Sigma \) is a restriction of \( \text{sig}(\varphi, \psi) \)
(otherwise, coincides with logical equivalence \( \equiv \))

relevant notion for normal forms, e.g., Scott Normal form for \( \text{FO}^2 \)

very strong (typically entails other notions) and

decision problem often undecidable, but exceptions:

some very lightweight Description Logics (e.g., some DL-Lite dialects)

if non-\( \Sigma \)-symbols are only unary, then also expressive DLs work
e.g. \text{ALC}; based on FMP, so might extend to \( \text{FO}^2 \) and GF
In applications, it often suffices to **preserve the consequences**

\[ \varphi \text{ (deductively) } \Sigma\text{-entails } \psi \text{ if for every } \mathcal{L}(\Sigma)\text{-sentence } \vartheta:\]

\[ \psi \models \vartheta \text{ implies } \varphi \models \vartheta \]

We write \( \varphi \models_{\Sigma} \psi \). \( \varphi \land \varphi' \) is deductive CE of \( \varphi \) if \( \varphi \models_{\Sigma} \varphi \land \varphi' \).

Let \( \mathcal{L} = GF \) and \( \Sigma = \{ R \} \)

\[ \varphi = \forall x \exists y Rxy \text{ and } \psi = \forall x ( \exists y (Rxy \land Ay) \land \exists y (Rxy \land \lnot Ay) ) \]

\[ \varphi \not\models_{\Sigma} \psi: \text{ consider } \vartheta = \exists xy (Rxy \land x \neq y) \]

\[ \varphi' = \forall x \exists y (Rxy \land x \neq y) \]

\[ \varphi' \models_{\Sigma} \psi: \text{ GF cannot count the number of } R\text{-successors} \]

\[ \triangleright \text{ non-}\Sigma\text{-predicates add "expressive power"} \]
Model-theoretic CE implies deductive CE (but not vice versa)

entailment implies $\Sigma$-entailment and
$\Sigma$-entailment coincides with entailment if $\Sigma \supseteq \text{sig}(\varphi, \psi)$

Deciding $\Sigma$-entailment at least as hard as satisifiability $\mathcal{L}$:

$\varphi$ is unsatisfiable  iff  $\varphi \vdash_{\text{sig}(\varphi)} \exists x. A(x) \land \neg A(x)$

$\Sigma$-entailment in $\mathcal{L}$ does not imply $\Sigma$-entailment in $\mathcal{L}' \not\supset \mathcal{L}$
A logic $\mathcal{L}$ ...

... enjoys Craig interpolation Property (CIP) if for any two formulas $\varphi, \psi$ with $\varphi \vdash \psi$, there is $\theta$ with $\varphi \vdash \theta \vdash \psi$ and $\text{sig}(\theta) \subseteq \text{sig}(\varphi) \cap \text{sig}(\psi)$.

... enjoys uniform interpolation Property (UIP) if the interpolant depends only on $\varphi$ and $\text{sig}(\varphi) \cap \text{sig}(\psi)$

- Effective UIP + decidable entailment $\Rightarrow$ decidable $\Sigma$-entailment:
  
  $$\varphi \models_\Sigma \psi \quad \text{iff} \quad \varphi \models UI_\Sigma(\psi)$$

- $\Sigma$-entailment decidable for modal logics $\mathbf{K}$, $\mathbf{GL}$, intuitionistic logic, $\mu$-calculus
  complexity not necessarily optimal: $UI_\Sigma(\psi)$ might be huge

- Conversely, if $\mathcal{L}$ enjoys CIP, then uniform interpolant recognition reduces to $\Sigma$-entailment
CASE STUDY: MODAL LOGIC $\mathcal{K}$

[Ghilardi, Lutz, Wolter, Zakharyaschev 2006]

- formulas built according to $\varphi ::= p \mid \varphi \land \varphi \mid \neg \varphi \mid \Diamond \varphi$
- interpreted in pointed structures $\mathcal{A}, a$
- relevant model-theoretic notion: bisimulations between $\mathcal{A}$ and $\mathcal{B}$

$\Sigma$-bisimulation is relation $S \subseteq \text{dom}(\mathcal{A}) \times \text{dom}(\mathcal{B})$ s.t. for every $aSb$

- (atom) $\mathcal{A}, a \models p \iff \mathcal{B}, b \models p$ for all $p \in \Sigma$
- (forth) if $aR^\mathcal{A}a'$ then there is $b'$ with $bR^\mathcal{B}b'$ and $a'Sb'$
- (back) if $bR^\mathcal{B}b'$ then there is $a'$ with $aR^\mathcal{A}a'$ and $a'Sb'$

$a \sim_\Sigma b$ if there is a $\Sigma$-bisimulation $S$ with $aSb$

- model-theoretic characterization

$\varphi \models_\Sigma \psi$ iff for every $\mathcal{A}, a \models \varphi$ there is $\mathcal{B}, b \models \psi$ with $a \sim_\Sigma b$

- if $\varphi \not\models_\Sigma \psi$ there is a small (exp-size) witness $\mathcal{A}, a \Rightarrow \text{NExpTime}$
DESCRIPTION LOGICS


- semantically like modal logic, except **global consequence**
- very similar **model-theoretic characterizations**, e.g.,

\[ \mathcal{O}_1 \models_\Sigma \mathcal{O}_2 \iff \text{for every } \mathcal{A} \models \mathcal{O}_1 \text{ and every } a \in \text{dom}(\mathcal{A}) \text{ there is } \mathcal{B} \models \mathcal{O}_2 \text{ and } b \in \text{dom}(\mathcal{B}) \text{ with } a \sim_{\mathcal{L}C(\Sigma)} b \]

- holds for **tree-shaped** \( \mathcal{A}, \mathcal{B} \Rightarrow \text{use (two-way, alternating) tree-automata:} \)

  - \( \mathcal{A}_1 \) accepts all tree-models of \( \mathcal{O}_1 \)
  - \( \mathcal{A}_2 \) accepts a tree-model if every element is bisimilar to an element in a model of \( \mathcal{O}_2 \)

\[ \mathcal{O}_1 \#_\Sigma \mathcal{O}_2 \iff L(\mathcal{A}_1) \cap \overline{L(\mathcal{A}_2)} \neq \emptyset \]

- works for many DLs: \( \mathcal{L}C, \mathcal{L}C\mathcal{I}, \mathcal{L}C\mathcal{Q}, \mathcal{L}C\mathcal{O} \)
- mostly with **tight complexity:** \( 2/3\text{ExpTime} \)
- exception \( \mathcal{L}C\mathcal{Q}\mathcal{I}\mathcal{O} \Rightarrow \text{undecidable} \)
WHY ARE …

… \( \Sigma \)-entailment and conservative extensions in modal and description logics so robustly decidable?

same question asked \textbf{for satisfiability} by [Vardi, 1996] and [Grädel 1999]

Answers by Grädel and Vardi:
- \textbf{tree-model property} enables use of tree-automata techniques even when extended with fixed points
  \( \Rightarrow \) guardedness serves as a better explanation than \( \text{FO}^2 \):
  - loss of nice model-theoretic properties and
  - less robust

What about \( \Sigma \)-entailment/conservative extensions in GF and \( \text{FO}^2 \)?
RESULTS IN A NUTSHELL

Conservative Extensions and $\Sigma$-entailment are
- **undecidable** in GF
- **undecidable** in $\mathbf{FO}^2$
- **undecidable** in any logic $\mathcal{L} \subseteq \mathbf{FO}$ that contains GF or $\mathbf{FO}^2$
- **decidable** and 2ExpTime-complete in $\mathbf{GF}^2 = \mathbf{GF} \cap \mathbf{FO}^2$

**Conclusion**  Decidability even more closely linked to real tree-model property (treewidth 1) not tree decompositions!

- upper bound for $\mathbf{GF}^2$ is far from trivial
  (in contrast to relatively easy decidability of satisfiability)
- consequences for uniform interpolant recognition:
  - undecidable in GNF
  - decidable in 2ExpTime for $\mathbf{GF}^2$
UNDECIDABILITY IN GF

Reduction from halting problem of two register machines $M$

Construct $\varphi_1, \varphi_2$ such that

1. if $M$ halts, then there is $\psi$ sucht that $\varphi_1 \land \varphi_2 \models \psi$, but $\varphi_1 \not\models \psi$

2. if there is a $\Sigma$-structure that satisfies $\varphi_1$, which cannot be extended to a model of $\varphi_2$, then $M$ halts

**Intended model**

In 1, $\psi$ is a description of the (finite!) computation of $M$

$\varphi_1$ enforces the very basic structure of the model above

$\varphi_2$ enforces finiteness and proper updating using non-$\Sigma$-symbols

In 2, the premise is from model-theoretic $\text{CE} \Rightarrow$ holds for $\text{GF} \subseteq \mathcal{L} \subseteq \text{FO}$
INITIAL OBSERVATION FOR GF\(^2\)

- GF\(^2\) has tree-model property and bisimulation \(\sim_{GF^2(\Sigma)}\),
  but natural characterization **fails**:

\[
\varphi_1 \models_{\Sigma} \varphi_2 \iff \text{for every } \mathcal{A} \models \varphi_1 \text{ there is } \mathcal{B} \models \varphi_2 \text{ such that}

\text{there is a } \text{global } GF^2(\Sigma)\text{-bisimulation between } \mathcal{A}, \mathcal{B}.
\]

- \(\Sigma = \{R\}\)
  \(\varphi_1 = \forall x \exists y Rxy\)
  \(\varphi_2 = \varphi_1 \land \exists x Bx \land \forall x Bx \rightarrow (\exists y Ryx \land By)\)

- model of \(\varphi_1\):

\[
\begin{array}{c}
\bullet & \bullet & \bullet & \cdots \\
R & R & R & \cdots
\end{array}
\]

- any model of \(\varphi_2\) contains:

\[
\begin{array}{c}
\ldots & \bullet & \bullet & \bullet & \bullet & \cdots \\
R & R & R & R & R & \cdots
\end{array}
\]

- \(\varphi_1 \models_{\Sigma} \varphi_2\), but **no global bisimulation**:
  - GF\(^2\)-bisimulations also travel “backwards”
**BOUNDED BISIMULATIONS TO THE RESCUE**

- *k*-bounded bisimulation is the *k*-round version of bisimulations: $\sim^k_{\text{GF}^2(\Sigma)}$

  ⇒ characterize expressive power of $\text{GF}^2$

\[
\varphi_1 \models_\Sigma \varphi_2 \text{ iff for every finite-outdegree forest-model } \mathcal{A} \models \varphi_1 \text{ and every } k \geq 0 \text{ there is } \mathcal{B} \models \varphi_2 \text{ such that } \\
1. \text{ for every } a \in \text{dom}(\mathcal{A}) \text{ there is } b \in \text{dom}(\mathcal{B}) \text{ with } a \sim_{\text{GF}^2(\Sigma)}^k b \\
2. \text{ for every } b \in \text{dom}(\mathcal{B}) \text{ there is } a \in \text{dom}(\mathcal{A}) \text{ with } a \sim_{\text{GF}^2(\Sigma)}^k b
\]

- $\Sigma = \{R\}$  
  $\varphi_1 = \forall x \exists y Rxy$  
  $\varphi_2 = \varphi_1 \land \exists x Bx \land \forall x Bx \rightarrow (\exists y Rxy \land By)$

  model of $\varphi_1$: 
  ![Model of $\varphi_1$]

  model of $\varphi_2$: 
  ![Model of $\varphi_2$]

- nice characterization, but **impossible to implement** in tree automaton
Here: $a \sim_{\text{GF}^2(\Sigma)}^X b$ is variant of bisimulation that
- behaves normally when “going down” the tree
- stops when finding the second $X$ when “going up”

$\Sigma$-entailment can be decided in time $\exp(\cdot)$ in $|\varphi_1|$, double exp. in $|\varphi_2|$

matching lower bound by reduction from exp space bounded ATMs
Another relevant application of ontologies is **ontology-mediated querying**: For database $\mathcal{D}$, ontology $\mathcal{O}$, conjunctive query $q(\mathbf{x})$, decide:

$$\mathcal{D} \cup \mathcal{O} \models q(\mathbf{a})$$

Ontology provides semantic layer to access the data in a more informative and user-friendly way. An appropriate notion of conservative extensions **preserves query answers**:

$\mathcal{O}_1 \Sigma_D, \Sigma_Q$-query entails $\mathcal{O}_2$, if for every $\Sigma_D$-database $\mathcal{D}$ and every $\Sigma_Q$-CQ $q$ and every $\mathbf{a}$:

$$\mathcal{D} \cup \mathcal{O}_2 \not\models q(\mathbf{a}) \text{ implies } \mathcal{D} \cup \mathcal{O}_1 \not\models q(\mathbf{a})$$

We write $\mathcal{O}_1 \not\models^{CQ}_{\Sigma_D, \Sigma_Q} \mathcal{O}_2$.

$\mathcal{O}_2 \cup \mathcal{O}_1$ is $\Sigma_D, \Sigma_Q$-query-CE of $\mathcal{O}_1$ iff $\mathcal{O}_1 \not\models^{CQ}_{\Sigma_D, \Sigma_Q} \mathcal{O}_1 \cup \mathcal{O}_2$.
Query-conservative extensions often more difficult than deductive conservative extensions

- **Undecidable** in $\mathcal{ALC}$ [Botoeva, Lutz, Ryzhikov, Wolter, Zakharyaschev 2019]
- Not monotone in the query language
- **Decidable** in $\mathcal{ALC}$ when considering UCQs [Botoeva et al 2019]
- Decidable for several Horn DLs such as DL-Lite and $\mathcal{ELI}$ [Konev, Kontchakov, Ludwig, Schneider, Wolter, Zakharyaschev 2011] [J, Lutz, Martel, Schneider 2020]

What about generalizations of DL-Lite and $\mathcal{ELI} \Rightarrow$ TGDs?
TUPLE-GENERATING DEPENDENCIES

Take the form:

\[ \forall x \forall y (\psi(x, y) \rightarrow \exists z \varphi(x, z)) \]  for CQs \( \psi, \varphi \)

Prominent classes

- **linear**
  
  \[ R(x, y) \rightarrow \exists z \varphi(x, y, z) \]

- **guarded**
  
  \[ R(x, y, z) \land S(y, z) \land P(x) \rightarrow \exists z \varphi(x, y, z) \]

- **frontier-one**
  
  \[ \psi(x, y) \rightarrow \exists z \varphi(z, x) \]

Examples

- \( \text{Movie}(x) \rightarrow \exists y \exists z \text{directedBy}(x, y) \land \text{Director}(y) \land \text{hasLocation}(x, z) \land \text{GeoLocation}(z) \)

- \( \text{Movie}(x) \land \text{hasScene}(x, y) \land \text{Violent}(y) \rightarrow \exists z \text{hasRating}(x, z) \land \text{AgeRestriction}(z) \)

(both actually \( \mathcal{ELI} \))
RESULTS IN A NUTSHELL

Query-conservative extensions and $\Sigma$-query entailment are

- **undecidable** for linear TGDs
- **undecidable** for guarded TGDs, even if first ontology is $\emptyset$
- **undecidable** for linear/guarded TGDs also when infinite CQs admitted
- **decidable** and PSpace-c. for linear TGDs when the first ontology is $\emptyset$
- **decidable** and 2ExpTime ... 3ExpTime for frontier-one TGDs

2ExpTime lower bound known from $\mathcal{L}\mathcal{I}$

CQ answering $\Rightarrow$ homomorphisms instead of bisimulations

We will be using the **chase** = unique hom-universal model of $\mathcal{O}, \mathcal{D}$

$\mathcal{O} \cup \mathcal{D} \models q(a)$ iff $\text{ch}_{\mathcal{O}}(\mathcal{D}) \models q(a)$
Pyramus and Thisbe are lovers whose parents are not on good terms

After one encounter, they are not allowed to meet each other

Instead they are doomed to communicate forever through a crack in the wall

(not important for the reduction: when their plans to flee Babylon fail they kill themselves)
Conway functions $F$ given by

$$S = \gamma, \alpha_0, \beta_0, \ldots, \alpha_{\gamma-1}, \beta_{\gamma-1}$$

$$F(n) = \frac{n\beta_i}{\alpha_i} \quad \text{for } i = n \mod \gamma$$

The following problem is **undecidable**:

Given $S$, is there $m \geq 0$ such that $F^m(2) = 1$?

Similar to two-register machines
creates (in the chase) the “union” of all rivers of this shape

such that $t_i = F(p_i)$ for all $i$

and $p_0 = 1$

(can be done with linear TGD)
Eternity

Bridges

homomorphism?
We start with possibly infinite CQs

The following are equivalent:

1. $\emptyset_1 \models_\Sigma^\omega_{\Sigma_D, \Sigma_Q} \emptyset_2$

2. $\text{ch}_{\emptyset_2}(\mathcal{D}) \rightarrow_{\Sigma_Q} \text{ch}_{\emptyset_1}(\mathcal{D})$ for all tree-like $\Sigma_D$-databases of width $|\emptyset_1|$

This can be “easily” decided via tree-automata

$\mathcal{D}$ and $\text{ch}_{\emptyset_1}(\mathcal{D})$ in the input

$\text{ch}_{\emptyset_2}(\mathcal{D})$ computed on the fly in the states of the automaton
CQ: DECIDABILITY EXAMPLE

\[ O_1 = \{ A(x) \rightarrow \exists y \, S(x, y), B(y), \quad B(x) \rightarrow \exists y \, R(x, y), B(y) \} \]
\[ O_2 = \{ A(x) \rightarrow \exists y \, S(x, y), B(y), \quad B(x) \rightarrow \exists y \, R(y, x), B(y) \} \]
\[ \Sigma_D = \{ A \} \quad \Sigma_Q = \{ R \} \]

\[ \text{ch}_{O_1}(D) \quad \text{ch}_{O_2}(D) \]

⇒ characterization based on homomorphisms does not work for CQs: bounded homomorphisms
The following are equivalent:

1. \( \mathcal{O}_1 \models^\text{CQ} \Sigma_D, \Sigma_Q \mathcal{O}_2 \)
2. \( \text{ch}_{\mathcal{O}_2}(\mathcal{D}) \to^\lim \Sigma_Q \text{ch}_{\mathcal{O}_1}(\mathcal{D}) \) for all tree-like \( \Sigma_D \)-databases of width \( |\mathcal{O}_1| \)

\( \Rightarrow \) Here \( I \to^\lim \Sigma \mathcal{J} \) means that \( I' \to \Sigma \mathcal{J} \) for every finite subinstance \( I' \) of \( I \)

\( \Rightarrow \) limit homomorphisms cannot be implemented directly in automaton

\( \Rightarrow \) We establish **automata-friendly intermediate characterization** carefully distinguishing where subinstances of \( \text{ch}_{\mathcal{O}_2}(\mathcal{D}) \) are mapped

\( \Rightarrow \) **Important:** case of \( \Sigma_Q \)-disconnected subinstances that map further and further away from \( \mathcal{D} \) can (and has to) be precomputed
(deductive) conservative extensions and $\Sigma$-entailment

many DLs

$\mathbf{GF^2}$

$\mathbf{GF, GNFO}$

$\mathbf{FO^2}$

$\mathbf{ELI}$, frontier-one TGDs

linear TGDs

guarded TGDs, $\mathbf{ALC}$

query conservative extensions and $\Sigma$-query entailment
CONCLUSION

Picture for model-theoretic, deductive, query-conservative extension quite well-understood

Interesting open questions

- UNFO
- GC²
- Horn-\(\mathcal{ALC}\) / Horn-Modal Logic
- Transitivity and temporal logics
- Other classes of TGDs

Practical algorithms missing
Decision procedures: (double-triple) exponentially large automata
THANK YOU VERY MUCH!

QUESTIONS?