Complexity and Expressive Power of Ontology-Mediated Queries

Carsten Lutz, University of Bremen
A Teaser

NP is an **interesting** class of problems
...and a **difficult** one! [Ladner73]

**Subclasses** might be more approachable:

FederVardi [93] conjecture: no 😱

Concrete conjecture on frontier from univ. algebra [BulatovJeavonsKrokhin05]

Many **subclasses** arise in **database theory**: consistent query answering, ontology-mediated queries, view-based deletion propagation, etc.

**Where can we classify complexities, where decide associated meta-problems?**
Today data is often highly incomplete and very heterogeneous.

Examples include web data and large-scale data integration.

Querying such data can be a challenging problem.

Ontology is a logical theory that:

- adds domain knowledge (@ incompleteness)
- interrelates diverging vocabularies (@ heterogeneity)
- provides unified view of the data
Ontology-Mediated Query: Example

Ontology $\mathcal{O}$:

$$\forall x (\text{Director}(x) \rightarrow (\text{Person}(x) \land \exists y (\text{directed}(x, y) \land \text{Movie}(y))))$$

Data $D$:

- Person: jj
- Movie: dbl
- Director: ww

Query $q(x)$:

$$\exists y (\text{Person}(x) \land \text{directed}(x, y) \land \text{Movie}(y))$$

Answers:

- jj
- ww
Ontology-Mediated Query: Example

Ontology $\mathcal{O}$:

$$\forall x (\text{Director}(x) \rightarrow (\text{Person}(x) \land \exists y (\text{directed}(x, y) \land \text{Movie}(y)))) \lor \text{Movie}(y) \lor \text{TVseries}(y))$$

Data $D$:

- Person: jj directed dbl
- Movie: directed
- Director: ww directed
- TVseries: directed

Query $q(x)$:

$$\exists y (\text{Person}(x) \land \text{directed}(x, y) \land \text{Movie}(y))$$

Answers: jj ww

Semantics: consider all extensions of the data that satisfy ontology and return answers valid in all of them (certain answers)
Description Logic

The World Wide Web Committee (W3C) has standardized a family of ontology languages for the web called OWL2.

OWL2 is based on a family of logics from KR/AI: description logics.

I will consider:

- ontologies formulated in description logics
- queries that are unions of conjunctive queries (UCQs)
  in other words: existential positive FO sentences

Description logics are...

- ...decidable fragments of FO...
- ...related to modal logics...
- ...often contained in the guarded fragment and in FO2
Only unary and binary predicates (concept names and role names)

Operators available in $\mathcal{ALC}$: (attribute concept language with complement [Schmidt-SchaußSmolka91])

- $A(x)$
- $\neg C(x)$, $C(x) \land D(x)$, $C(x) \lor D(x)$
- $\exists r. C$
- $\forall r. C$
- $\forall x C(x) \rightarrow D(x)$
- $\forall y r(x, y) \land C(y)$
- $\forall y r(x, y) \rightarrow C(y)$

Ontology: finite set of $C \subseteq D$

For example: Director $\sqsubseteq$ Person $\sqcap \exists$directed.(Movie $\sqcap$ TVseries)

Theorem. An FO-sentence is equivalent to an $\mathcal{ALC}$-ontology iff it is preserved under global bisimulation and disjoint union.
Ontology-Mediated Queries

Ontology-mediated query (OMQ): triple \((\mathcal{O}, \Sigma, q)\) where

- \(\mathcal{O}\) is ontology
- \(q\) is query
- \(\Sigma\) is signature of the data (set of allowed symbols)

Example: \(\Sigma = \{e\}, e\) binary (we speak about digraphs)

Ontology \(\mathcal{O}\):

\[
\top \subseteq R \cup G \cup B \quad \forall x (R(x) \lor G(x) \lor B(x))
\]

\[
C \cap \exists e.C \subseteq D \quad \text{for } C \in \{R, G, B\} \quad \forall x \forall y (C(x) \land e(x, y) \land C(y) \rightarrow D(x))
\]

Query: \(q() = \exists x D(x)\)

Expresses non-3-colorability, thus coNP-complete
Ontology-Mediated Queries

Ontology-mediated query (OMQ): triple \((\mathcal{O}, \Sigma, q)\) where

Central questions in the field:

- how to distinguish tractable from intractable OMQs?
- when is an OMQ rewritable into FO / into Datalog?
  - SQL
- how to decide whether OMQ belongs to these classes?
- how to compute rewritings when they exist?

Query: \(q(x) \equiv \exists y \ (\Sigma (x, y) \land \exists z \ (\Sigma (y, z) \land \Sigma (z, x)))\)

Expresses non-3-colorability, thus coNP-complete
Ontology-Mediated Queries

OMQ language:

pair \((L, Q)\) with \(L\) ontology language and \(Q\) query language

for example \((\text{ALC}, \text{UCQ})\), but many other choices:

- extensions of \(\text{ALC}\) such as \(\text{ALCI}, \text{ALCF}\), the guarded fragment of FO
- other queries such as tree-shaped UCQs (tUCQs)

All of these are subclasses of coNP
CSP Equivalence

We start with tree-shaped queries: 

\[(\mathcal{ALC}, \text{tUCQ}) = \text{coCSP}\]

CSP is homomorphism problem:

for finite relational structure \(T\) (template), \(\text{CSP}(T) = \{S \mid S \rightarrow T\}\)

**Theorem [BienvenuTenCateL_Wolter13]**

Every OMQ from \(\mathcal{ALC}, \text{tUCQ}\) is equivalent to the complement of a CSP and vice versa.

From CSP to OMQ:

\[T \sqsubseteq R \sqcup G \sqcup B\]

\[C \sqcap \exists e.C \sqsubseteq D \quad \text{for } C \in \{R, G, B\}\]

\[q() = \exists x \ D(x)\]
Take disjoint union of all finite models of $\mathcal{O}$ that make $q$ false

then make finite by applying filtration:

- 1-type of an element is set of all subformulas of $\mathcal{O}$ and $q$
  that are true at that element

- identify all elements with same 1-type

Ontology:

\[
A \subseteq \forall r. A \\
B \subseteq \exists r. B
\]

Query:

\[
q(x) = \exists x \ A(x) \land B(x)
\]
Meta-Problems

Many relevant results transfer from CSP to OMQ, e.g.:

- (co)NP/PTime dichotomy iff the FV conjecture is true
- Tractability via Datalog + group theory / certain polymorphism [FederVardi93, CohenJeavonsGyssens97, many others]
- No complexities between FO=AC$_0$ and LogSpace [LaroseTesson07]
- FO- and Datalog-rewritability decidable [LaroseLotenTardiff07, BartoKozik09]

**Theorem [BienvenuTenCateL_Wolter13]**

FO-rewritability and Datalog-rewritability in (\(\mathcal{ALC}\), tUCQ) is decidable and NEXPTIME-complete.

Upper bound from CSP translation, lower bound from tiling problem
Obstructions

Structure $O$ is obstruction for $T$ if $O \rightarrow S$ implies $S \not
\rightarrow T$ for all $S$

Obstruction set $\mathcal{O}$ complete if $S \not
\rightarrow T$ implies $O \rightarrow S$ for some $O \in \mathcal{O}$

For example $T = T_2$ and $\mathcal{O} = \{P_3\}$.

$\mathcal{O}$ complete obstruction set for $T_Q \Rightarrow \bigvee \mathcal{O}$ existential positive rewriting of $Q$
Computing Rewritings

[Atserias05, NesetrilTardiff00]:

\[
\text{FO rewritability} = \text{finite obstruction set of finite trees}
\]

Can guide the computation of FO-rewritings (tUCQ suffices!)

Datalog-rewritability connected to obstructions of bounded tree width

But there is something better:

(2,3)-canonical Datalog program is most complete Datalog-approximation of coCSP, complete if Datalog-rewritable

[FederVardi93, BartoKozik09]

None of this immediately practical (large blowups involved)
Beyond Trees

Consider \((\mathcal{ALC}, \text{UCQ})\) where queries are no longer trees.

For special case \((\emptyset, \Sigma, q)\), we need template \(T_q\) such that

\[ q \rightarrow D \quad \text{iff} \quad D \not\models T_q \quad \text{for all data sets } D \]

Such singleton duality exists iff \(q\) is tree-shaped \([\text{NesetrilTardiff00}]\).

We thus need generalization of CSP: MMSNP to the rescue!

\[
\begin{align*}
\text{NP} & = \exists \text{SO} \\
\cup \cup & \approx_{\text{PTime}} \text{MMSNP} \\
\exists S_1 \cdots \exists S_n \forall x_1 \cdots \forall x_m \, \varphi \text{ with } \varphi \text{ conjunction of } \bigwedge_i P_i(\overline{x}_i) \rightarrow \bigvee_i S_i(x_i)
\end{align*}
\]

**Theorem [BienvenuTenCateL_Wolter13]**

Every OMQ from \((\mathcal{ALC}, \text{UCQ})\) is equivalent to the complement of an MMSNP sentence and vice versa.
Meta-Problems

Thus ($\mathcal{ALC}$,UCQ) has coNP/PTime dichotomy iff FV conjecture holds.

Meta-problems can be attacked via FV-translation of MMSNP to CSP:

- FO-rewriting of CSP yields FO-rewriting of MMSNP sentence
- Conversely, we get a rewriting of the CSP that is complete only on inputs whose girth exceeds rule diameter of MMSNP sentence

Fill gap by analysing obstructions for FO-rewritable MMSNP sentences: finite sets of structures of treewidth $(1,k)$, $k$ diameter of sentence

**Theorem [BourhisL_16,unpublished]**

FO- and monadic Datalog-rewritability are decidable and 2$\text{NExpTime}$-complete in MMSNP and in ($\mathcal{ALC}$, UCQ).

Currently working on Datalog-rewritability (2NExpTime-hardness clear)
Guarded Fragment

Consider (GF, UCQ) where GF is guard fragment of FO

**Theorem [BienvenuTenCateL_Wolter13]**
(GF,UCQ) is strictly more expressive than the complement of MMSNP

Non-expressible property:
“cartwheel” reachability

(Proof via coloured forbidden patterns)

GMSNP: non-monadic MMSNP, but implications must be guarded

such as \( \exists S \forall x \forall y R(x, y) \rightarrow S(x, y) \lor S(x, x) \)

**Theorem [BienvenuTenCateL_Wolter13]**
Every OMQ from (GF,UCQ) is equivalent to the complement of a GMSNP sentence and vice versa.

PTime / NP dichotomy status of GMSNP is interesting open problem
Overview

- coNP
  - coGMSNP
    - GF, UCQ
  - (GNFO, UCQ)
  - coMMSNP
    - (ALC, UCQ)
    - (ALCI, UCQ)
  - coCSP
    - (ALC, tUCQ)
    - (ALCI, tUCQ)
    - (SHI, tUCQ)
    - (SHI, UCQ)
    - (UNFO, UCQ)
\(\mathcal{ALCF}\) extends \(\mathcal{ALC}\) with functional role declarations:

\[
\text{func(hasMother)} \quad \text{func(hasFather)}
\]

\(\mathcal{ALCF}\) does not admit filtration, so template construction fails!

Theorem [L_Wolter12]

\[
(\mathcal{ALCF}, \text{tCQ}) \approx_{\text{PTime}} \text{coNP}
\]

trees, no disjunctions!

Thus \((\mathcal{ALCF}, \text{tCQ})\) has no PTime/coNP dichotomy (unless PTime = NP)

The proof requires the ontology to `verify' a grid-structure in the data which relies on some relations being partial functions
Closed Predicates

Back to \((\mathcal{ALC}, \text{tUCQ})\), but now some predicates can be closed in data

\[ V \xleftrightarrow{e} R \] \[ V \xleftrightarrow{e} e \] \[ V \xrightarrow{e} V \]

Ontology: \( S \subseteq \exists r. (V \cap R) \)

monadic predicate \( V \) closed

Natural setup: corresponds to partially complete data

Theorem [SeylanL_Wolter15]

\((\mathcal{ALC}, \text{tUCQ})\) w. closed preds “equivalent” to surjective CSPs, but

- not same expressivity, only same complexities up to FO-reductions
- small gap: OMQs translate into surjective multi-template CSPs

Surjective CSPs are difficult and little is known about them:

simple templates of unknown complexity (6-cycle), no dichotomy conjecture
Overview

- $\text{coCSP} (\mathcal{ALCF}, \text{tCQ})$
- $\text{coMMSNP} (\mathcal{ALC}, \text{UCQ})$
- $\text{coNP} (\mathcal{ALC}, \text{tUCQ})$
- $\text{surjective CSP} (\mathcal{ALC}, \text{tUCQ})$
- $\text{+closed}$
- $\text{coGMSNP} (\text{GF, UCQ})$

Universität Bremen
Beyond Ontology-Mediated Queries

Other areas of database theory provides more subclasses of (co)NP:

- consistent query answering,
- deletion propagation,
- peer data exchange,
- causality,
- resilience,
- disjunctive query languages, etc.

All based on positive existential queries, thus homomorphisms

There should be some kind of connection to CSP!?

Quick look at Consistent Query Answering (CQA)

Assume a set of constraints $C$ and data $D$ that violates $C$

- **Repair** is data $D'$ that satisfies $C$ and with $D' \Delta D$ minimal

- Given $C$, $D$, and query $q$, we want to compute the answers to $q$ on which all repairs agree

A CQA problem is a pair $(C, q)$. 
Consistent Query Answering (CQA)

First connection to CSP established in [Fontaine13]

**Theorem [Fontaine13]**

For every CSP, there is a CQA problem that has the same complexity up to PTime reductions.
Constraints take form $\forall \bar{x} \neg \varphi(\bar{x})$ with $\varphi$ a UCQ, queries are UCQs too

How close is the connection between CQA and CSP?
No equivalence-preserving translations can be expected!

Important reason for **clean connection between OMQ and CSP:**
Relevant ontology languages are **essentially unary** in the sense that 1-types “largely suffice to describe modes” (recall filtration!)
First Crack at CQA

**Theorem [L_Wolter15]**

- coCSP \( \approx_{FO} \) CQA: constraints \( \forall x \neg (A_1(x) \land \cdots \land A_n(x)) + tUCQs \)
- coMMSNP \( \approx_{FO} \) CQA: same constraints + UCQs
- coGMSNP \( \approx_{FO} \) CQA: constraints \( \forall x \neg (R_1(x) \land \cdots \land R_n(x)) + UCQs \)

where all \( R_i(x) \) have same arity and use same vars in same order

More important are **key constraints (functional relations)**

Known:

- PTime / coNP dichotomy for “self-join free queries” [KoutrisWijsen15]
- at least as hard to classify as conservative CSPs [Fontaine13]

No full understanding of relation to CSP yet…
Thank you!

- Close connection between ontology-mediated queries and CSP
- New interest in MMSNP, definability questions, dualities
- Analyzing subclasses of NP is fun, and there’s a lot left to be done!