Boundedness of Conjunctive Regular Path Queries

Pablo Barceló (Univ. of Chile & IMFD)   Diego Figueira (CNRS & LaBRI)

Miguel Romero (Univ. of Oxford)

ICALP 2019, July 11, Patras, Greece
The Boundedness problem

• Basic optimization task for recursive queries
• **Question:** Can we remove recursion from a recursive query?
• **Motivation:** Non-recursive queries behave better!

This talk: Datalog and fragments
(Unions of conjunctive queries (UCQs) + recursion)

**Definition:**
A Datalog program is **bounded** if it is *equivalent* to a UCQ

**Boundedness problem:**
Given a Datalog program, is it bounded?
Previous work

- **Undecidable** for Datalog (even linear)  
  (Gaifman, Mairson, Sagiv, Vardi LICS’87)

- Several decidability/undecidability result since then…
  - Arity of intentional predicates, number of rules, connectivity, …

- **Decidable** for *monadic* Datalog  
  (Cosmadakis, Gaifman, Kanellakis, Vardi STOC’88)
  - 2EXPTIME-complete  
    (Benedikt, ten Cate, Colcombet, Vanden Boom LICS’15)

- **Decidable** for *guarded* Datalog  
  (Blumensath, Otto, Weyer LMCS’14)
  - 2EXPTIME-complete  
    (Benedikt, ten Cate, Colcombet, Vanden Boom LICS’15)

- **Decidable** for *guarded* Datalog + parameters
  - Non-elementary upper bound  
    (Benedikt, Bourhis, Vanden Boom LICS’16)
Contributions

We consider unions of conjunctive two-way regular path queries (UC2RPQs)

- Basic navigational language for graph databases

UC2RPQs are subsumed by guarded Datalog + parameters

- Decidability of boundedness and non-elementary upper bound from Benedikt, Bourhis, Vanden Boom LICS’16

Main Question:
What is the precise complexity of boundedness for UC2RPQs?

- Is it elementary?
Contributions

Boundedness for UC2RPQs is \textbf{EXPSPACE-complete}

- Same as containment (Calvanese, Giacomo, Lenzerini, Vardi KR’00)

Tight size bounds of equivalent UCQs (triple exponential)

Better-behaved restrictions of UC2RPQs

- \textit{Acyclic} UC2RPQs of \textit{bounded thickness}
- Boundedness is \textbf{PSPACE-complete}
General picture

Datalog

Guarded Datalog + parameters
Non-elementary (Benedikt et al. '16)

Guarded Datalog
2EXPTIME-complete
(Blumensath et al.'88; Benedikt et al.'15)

Monadic Datalog
2EXPTIME-complete
(Cosmadakis et al.'88; Benedikt et al.'15)

Linear Datalog
Undecidable (Gaifman et al. '87)

UC2RPQ
EXPSPACE-complete (this paper)

UCQ

Undecidable (Gaifman et al. '87)
Graph databases and 2RPQs

Graph databases:
- Binary relational schema $S$
- Edge-labeled directed graphs

**Definition:**
A regular path query (RPQ) $L$ is a regular language over $S$

**Semantics:**
$$L(G) := \{(u,v): \text{there is directed path from } u \text{ to } v \text{ in } G \text{ whose label satisfies } L\}$$

**Examples:**
$S = \{\text{knows, friends}\}$
$L = (\text{knows + friends})^*$
Graph databases and 2RPQs

Definition:

A two-way RPQ (2RPQ) $L$ is a regular language over $S \cup S^{-1}$

$S^{-1} := \{a^{-1} : a \text{ in } S\}$ is the set of inverse symbols

Oriented path = forward and backward edges

Semantics:

$L(G) := \{(u, v) : \text{there is oriented path from } u \text{ to } v \text{ in } G$ whose label satisfies $L\}$

Examples: $S = \{\text{knows, friends}\}$

$L = (\text{knows.knows}^{-1})^*$

$\text{Label} = \text{a b a}^{-1} \text{ a b}^{-1}$
Unions of Conjunctive 2RPQs (UC2RPQs)

**Definition:**
A conjunctive 2RPQ (C2RPQ) $Q(x)$ is an expression:

$$Q(x) = \exists z \left( L_1(w_1, y_1) \land \cdots \land L_m(w_m, y_m) \right)$$

where

- Each $L_i$ is a 2RPQ
- Each $w_i, y_i$ is in $z$
- $x$ are the *free variables*

A mapping $h$ from the variables of C2RPQ $Q(x)$ to database $G$ is a **homomorphism** if for each $i$, $(h(w_i), h(y_i))$ is in $L_i(G)$

**Semantics:**

$$Q(G) := \{ h(x) : h \text{ is a homomorphism from } Q \text{ to } G \}$$
Unions of Conjunctive 2RPQs (UC2RPQs)

Definition:
A union of C2RPQs (UC2RPQ) $Q(x)$ is an expression:

$$Q(x) = Q_1(x) \lor \cdots \lor Q_n(x)$$

Semantics: $Q(G) := \bigcup_{1 \leq i \leq n} Q_i(G)$

UC2RPQs = core of most navigational graph query languages

Remark:
A UCQ is a UC2RPQ where each 2RPQ L is a single symbol
Main result

Main Theorem:
Boundedness for UC2RPQs is **EXPSPACE-complete**

- Same as for containment (and equivalence) (Calvanese, Giacomo, Lenzerini, Vardi KR’00)
- Lower bound from containment (EXPSPACE-hard even for Boolean CRPQs)
- Bounds for the size of equivalent UCQ

Theorem:
Every bounded UC2RPQ is equivalent to a UCQ with
- at most triply-exponentially many disjuncts
- each of them of size at most double exponential
and hence of at most triple exponential size. This is tight in general.
EXPSPACE upper bound

- Classical automata techniques used for containment + cost automata
- Well-known approach (Blumensath et al.’14; Benedikt et al.’15,’16): Reduce boundedness to limitedness of cost automata
- Non-elementary bound Benedikt et al.’16: sophisticated cost automata on trees

Observation:
For UC2RPQs, we can use distance automata over finite words
EXPSPACE upper bound

- A UC2RPQ $Q$ is bounded iff it is bounded over its canonical models (expansions)
EXPSPACE upper bound

- A UC2RPQ $Q$ is bounded iff it is bounded over its canonical models (expansions). Replace each 2RPQ $L(x,y)$ by a “fresh oriented path” from $x$ to $y$ with label in $L$. 


EXPSPACE upper bound

- A UC2RPQ $Q$ is bounded iff it is bounded over its canonical models (expansions)
  - There is $k$ such that for every canonical model $C$ of $Q$ the “cost of mapping” $Q$ to $C$ is at most $k$
EXPSPACE upper bound

- A UC2RPQ $Q$ is bounded iff it is bounded over its canonical models (expansions).
- There is $k$ such that for every canonical model $C$ of $Q$ the “cost of mapping” $Q$ to $C$ is at most $k$.

**Minimal size** of an expansion of $Q$ that maps homomorphically to $C$. 
EXPSPACE upper bound

- A UC2RPQ Q is bounded iff it is bounded over its canonical models (expansions)
  - There is $k$ such that for every canonical model $C$ of Q the “cost of mapping” Q to C is at most $k$
- We construct for Q a distance automata $A_Q$ of exponential size that given an (encoding) of a canonical model C computes “cost of mapping” Q to C
  - Q is bounded iff $A_Q$ is limited
- Upper bound follows from the following result:

Theorem (Leung’91; Leung, Podolskiy’04): The limitedness problem for distance automata is $\text{PSPACE-complete}$
Better-behaved UC2RPQs: acyclicity + bdd thickness

Theorem:
Fix positive integer $k$.
Boundedness for acyclic UC2RPQs of thickness at most $k$ is \textbf{PSPACE-complete}
Better-behaved UC2RPQs: acyclicity + bdd thickness

Theorem:
Fix positive integer $k$.
Boundedness for acyclic UC2RPQs of thickness at most $k$ is PSPACE-complete

Underlying graphs of C2RPQs are acyclic

Maximum number of 2RPQs between two distinct variables
Theorem:
Fix positive integer $k$.
Boundedness for acyclic UC2RPQs of thickness at most $k$ is **PSPACE-complete**

- Same as for containment (and equivalence) (implicit in Barceló, R., Vardi SICOMP’16)
- Both conditions are necessary:
  - EXPSPACE-hard for acyclic UC2RPQs
  - EXPSPACE-hard for thickness-1 UC2RPQs of treewidth 2
- Reduction to alternating two-way distance automata

Theorem:
The limitedness problem for alternating two-way distance automata is **PSPACE-complete**
Concluding remarks

- Elementary tight bounds for boundedness of UC2RPQs

Open questions:

- Can we use only classical automata techniques?
- More fragments of Datalog with elementary boundedness?
General picture

- **Datalog**
  - Undecidable (Gaifman et al. ’87)

- **Linear Datalog**
  - Undecidable (Gaifman et al. ’87)

- **Guarded Datalog + parameters**
  - Non-elementary (Benedikt et al.’16)

- **Guarded Datalog**
  - 2EXPTIME-complete
    - (Blumensath et al.’88; Benedikt et al.’15)

- **Monadic Datalog**
  - 2EXPTIME-complete
    - (Cosmadakis et al.’88; Benedikt et al.’15)

- **UCQ**

- **UC2RPQ**
  - EXPSPACE-complete (this paper)
General picture

Datalog

Undecidable (Gaifman et al. ’87)

Linear Datalog

Undecidable (Gaifman et al. ’87)

Regular Datalog?

Containment is 2EXPSPACE-complete
(Reutter, R., Vardi ICDT’15)

Guarded Datalog + parameters

Non-elementary (Benedikt et al.’16)

Guarded Datalog

2EXPTIME-complete
(Blumensath et al.’88; Benedikt et al.’15)

Monadic Datalog

2EXPTIME-complete
(Cosmadakis et al.’88; Benedikt et al.’15)

UCQ

UC2RPQ

EXPSPACE-complete (this paper)
Concluding remarks

• Elementary tight bounds for boundedness of UC2RPQs

Open questions:

• Can we use only classical automata techniques?
• More fragments of Datalog with elementary boundedness?
  • Natural candidate: Regular Datalog

Thank you!