Semantic Acyclicity on Graph Databases

Pablo Barceló
Miguel Romero
Moshe Y. Vardi

Univ. of Chile
Univ. of Chile
Rice University

June 26, PODS 2013, New York, USA
Motivation

**Acyclicity** is a restriction that ensures efficient evaluation for:
1. conjunctive queries (relational databases)
2. conjunctive regular path queries (graph databases)

**Goal:**
Understand the **semantic space** defined by acyclicity.

- The class of queries that are equivalent to an acyclic one, i.e.,
  the class of **semantically acyclic** queries.

1. Is this notion decidable?
2. What is the complexity of evaluating queries in this class?
Motivation

Semantic acyclicity for conjunctive queries is well understood thanks to connections with CSP.

In this work:

We study semantic acyclicity in the graph database context

- We study semantic acyclicity for **unions of conjunctive regular path queries with inverse (UC2RPQs)**.
Contributions

1. Checking semantic acyclicity for UC2RPQs is decidable.
2. Evaluating semantically acyclic UC2RPQs is fixed-parameter tractable.

Our approach:
- Study the notion of maximally contained acyclic UC2RPQs.
- Apply this to semantic acyclicity recognition and evaluation.
Outline

Motivation

Graph databases

Interlude on conjunctive queries

Semantic acyclicity of UC2RPQs

Open questions
Outline

Motivation

Graph databases

Interlude on conjunctive queries

Semantic acyclicity of UC2RPQs

Open questions
Graph databases

**Graph database:** Directed graph whose edges are labeled over a finite alphabet $\Sigma$.

**Example:** A fragment of the RDF representation of DBLP.
Regular path queries

Basic querying mechanism for Graph DBs: Regular path queries with inverse (2RPQs).

- 2RPQ $r = \text{Regular expression over alphabet } \Sigma \cup \{a^- | a \in \Sigma\}$.

Semantics: Evaluation $[r]_G$ of 2RPQ $r$ over graph DB $G$:

- Pairs $(u, v)$ of nodes in $G$ that are linked by a path whose label satisfies $r$, where:
  1. Forward traversals of edges with label $a$ are interpreted as $a$.
  2. Backward traversals of edges with label $a$ are interpreted as $a^-$. 
Example of 2RPQ

**Example:** The 2RPQ

\[
(creator \cdot ((partOf \cdot series) \cup journal))
\]

computes pairs \((a, v)\) such that \(a\) is an author and \(v\) is either a conference or a journal in which \(a\) published a paper.
Example of 2RPQ

**Example:** The 2RPQ

\[
(\text{creator}^- \cdot ((\text{partOf} \cdot \text{series}) \cup \text{journal}))
\]

computes pairs \((a, v)\) such that \(a\) is an author and \(v\) is either a conference or a journal in which \(a\) published a paper.
Example of 2RPQ

**Example:** The 2RPQ

\[
(\text{creator}^{-}\cdot ((\text{partOf} \cdot \text{series}) \cup \text{journal}))
\]

computes pairs \((a, v)\) such that \(a\) is an author and \(v\) is either a conference or a journal in which \(a\) published a paper.
Conjunctive 2RPQs

**Conjunctive 2RPQs (C2RPQs):** 2RPQs extended with conjunction and existential quantification.

Rules of the form:

\[
\text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m),
\]

such that

- the \(x_i, y_i\) are variables,
- each \(r_i\) is a 2RPQ,
- \(\bar{z}\) has some variables among the \(x_i, y_i\).

**CRPQ:** C2RPQ without inverse. **UC2RPQ:** Union of C2RPQs.
Conjunctive 2RPQs: Semantics

Semantics: Given graph DB $G$ with nodes $V$ and $Q(\bar{z})$ of the form

$$\text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m).$$
Conjunctive 2RPQs: Semantics

**Semantics:** Given graph DB $G$ with nodes $V$ and $Q(\vec{z})$ of the form

$$\text{Ans}(\vec{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m).$$

- Valuation $\sigma : \{x_1, \ldots, x_m, y_1, \ldots, y_m\} \rightarrow V.$
Semantics: Given graph DB $G$ with nodes $V$ and $Q(\bar{z})$ of the form

$$\text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m).$$

- Valuation $\sigma : \{x_1, \ldots, x_m, y_1, \ldots, y_m\} \rightarrow V$.
- It satisfies $Q$ if $(\sigma(x_i), \sigma(y_i)) \in [r_i]_G$, for each $i$. 
Conjunctive 2RPQs: Semantics

Semantics: Given graph DB $G$ with nodes $V$ and $Q(\vec{z})$ of the form

$$Ans(\vec{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m).$$

- Valuation $\sigma : \{x_1, \ldots, x_m, y_1, \ldots, y_m\} \rightarrow V$.
- It satisfies $Q$ if $(\sigma(x_i), \sigma(y_i)) \in [r_i]_G$, for each $i$.
- $[Q]_G$ has the tuples $\sigma(\vec{z})$ for all $\sigma$'s that satisfy $Q$. 
Semantics: Given graph DB $G$ with nodes $V$ and $Q(\bar{z})$ of the form

$$\text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m).$$

- Valuation $\sigma : \{x_1, \ldots, x_m, y_1, \ldots, y_m\} \rightarrow V$.
- It satisfies $Q$ if $(\sigma(x_i), \sigma(y_i)) \in \llbracket r_i \rrbracket_G$, for each $i$.
- $\llbracket Q \rrbracket_G$ has the tuples $\sigma(\bar{z})$ for all $\sigma$'s that satisfy $Q$.

If $Q$ is UC2RPQ $\bigcup_i Q_i$, then $\llbracket Q \rrbracket_G = \bigcup_i \llbracket Q_i \rrbracket_G$. 
Example of C2RPQ

Example: The C2RPQ

\[\text{Ans}(x, u) \leftarrow (x, \text{creator}^-, y), (y, \text{partOf} \cdot \text{series}, z), (y, \text{creator}, u)\]

computes pairs \((a_1, a_2)\) such that \(a_1\) and \(a_2\) are coauthors of a conference paper.
Example of C2RPQ

Example: The C2RPQ

\[ \text{Ans}(x, u) \leftarrow (x, \text{creator}^-, y),(y, \text{partOf} \cdot \text{series}, z),(y, \text{creator}, u) \]

computes pairs \((a_1, a_2)\) such that \(a_1\) and \(a_2\) are coauthors of a conference paper.
Complexity of evaluation of UC2RPQs

Complexity of evaluation of UC2RPQs is **NP-complete**. Problem remains hard even for CRPQs.

Problem is also hard in **parameterized complexity**:

- It cannot be solved in time $O(|G|^c \cdot f(|Q|))$, for $c \geq 1$ and computable function $f : \mathbb{N} \rightarrow \mathbb{N}$.

**Idea**: Find restrictions that lead to (fixed-parameter) tractability.

- We can transfer restrictions identified for evaluation of conjunctive queries over relational databases.
- The oldest and most common such restriction is **acyclicity**.
Outline

Motivation

Graph databases

Interlude on conjunctive queries

Semantic acyclicity of UC2RPQs

Open questions
Interlude on conjunctive queries

Conjunctive queries (CQs) are the most basic queries for relational databases.

A CQ is a rule of the form:

$$\text{Ans}(\bar{z}) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m),$$

such that

- the $\bar{x}_i$’s are tuples of variables,
- each $R_i$ is a relation,
- $\bar{z}$ has some variables among the $\bar{x}_i$’s.

Evaluation: Over database $D$, this CQ retrieves tuples $\sigma(\bar{z})$ such that $\sigma : \bigcup_i \bar{x}_i \rightarrow D$ and $R_i(\sigma(\bar{x}_i)) \in D$, for each $i$. 
Acyclic CQs

A CQ is **acyclic** if it can be decomposed into a tree that “preserves” its structure.

Formally:

A CQ $\text{Ans}(\bar{z}) \leftarrow R_1(\bar{x}_1), \ldots, R_m(\bar{x}_m)$ is acyclic if it has a join tree, i.e., a tree $T$ such that:

1. The nodes of $T$ are the **atoms** of the CQ.
2. For each variable $x$ the set $\{t \in T \mid x \in t\}$ is connected.
Evaluation of acyclic CQs

Acyclic CQs can be evaluated in linear time in the size of the data and the query:

**Theorem (Yannakakis, 1981)**

Acyclic CQs can be evaluated in time $O(|D| \cdot |Q|)$.

This is also true for unions of acyclic CQs.
Semantically acyclic UCQs

A union of CQs (UCQs) is semantically acyclic if it is equivalent to a union of acyclic CQs.

1. Is this notion decidable?
2. What is the complexity of evaluating queries in this class?

Using known CSP techniques we can answer these questions.
Checking semantic acyclicity

If a UCQ is semantically acyclic, then it is equivalent to a union of acyclic UCQs of at most polynomial size.

- Checking whether a UCQ is semantically acyclic is in NP.
- Checking semantic acyclicity is NP-hard, even for CQs over graphs (Dalmau, Kolaitis, Vardi, 2004).
Evaluation of semantically acyclic UCQs

Corollary

Evaluation of semantically acyclic UCQs is **fixed-parameter tractable**:

\[
|Q'| \leq \text{poly}(|Q|)
\]

\[
O(D^{O(|Q|)}) + D \cdot \text{poly}(|Q|)
\]
Evaluation of semantically acyclic UCQs (Revisited)

Surprisingly, the previous result can be improved:

**Theorem (Chen, Dalmau, 2005)**

Semantically acyclic UCQs can be evaluated in polynomial time.

**Remark:** Evaluation is polynomial when seen as a promise problem.

**Conclusion**

- Semantic acyclicity extends tractability of acyclicity for UCQs
- but recognition of queries in the class has an important cost in complexity.
Outline

Motivation

Graph databases

Interlude on conjunctive queries

Semantic acyclicity of UC2RPQs

Open questions
C2RPQs vs conjunctive queries

C2RPQs are conjunctive queries over graph databases, in which atomic formulas are defined by 2RPQs.

We associate with each C2RPQ

$$Ans(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m)$$

an underlying CQ

$$Ans(\bar{z}) \leftarrow r_1(x_1, y_1), \ldots, r_m(x_m, y_m).$$
Acyclic C2RPQs

The C2RPQ

\[
\text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m)
\]

is acyclic, if its underlying CQ

\[
\text{Ans}(\bar{z}) \leftarrow r_1(x_1, y_1), \ldots, r_m(x_m, y_m)
\]

is acyclic.
Evaluation of acyclic C2RPQs

Acyclic C2RPQs are tractable:

**Proposition**

Evaluation of an acyclic C2RPQ

\[ Q = \text{Ans}(\bar{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m) \]

over a graph DB \( G \) takes time \( O(|G|^2 \cdot |Q|^2) \).

This extends to unions of acyclic C2RPQs.
Acyclic C2RPQs (Revisited)

Since graph DBs consist only of binary relations, acyclicity of a C2RPQ \( \text{Ans}(\vec{z}) \leftarrow (x_1, r_1, y_1), \ldots, (x_m, r_m, y_m) \) can also be defined in a more intuitive way.

**Proposition**

A C2RPQ \( Q \) is acyclic iff its underlying simple and undirected graph \( \mathcal{U}(Q) \) is acyclic, where \( \mathcal{U}(Q) = (V, E) \) for:

- \( V = \{x_1, y_1, \ldots, x_m, y_m\} \);
- \( E = \{\{x_i, y_i\} \mid 1 \leq i \leq m \text{ and } x_i \neq y_i\} \).
Acyclic C2RPQs: Examples

\[ \text{Ans}(x, u) \leftarrow (x, \text{creator}^-, y), (y, \text{partOf} \cdot \text{series}, z), (y, \text{creator}, u). \]
Acyclic C2RPQs: Examples

\[ \text{Ans}(x, u) \leftarrow (x, \text{creator}^-, y), (y, \text{partOf} \cdot \text{series}, z), (y, \text{creator}, u). \]

\[ Q \]

\[ U(Q) \]

\( Q \) is acyclic.
Acyclic C2RPQs: Examples

Remark: Acyclicity allows loops and multi-edges.
Acyclic C2RPQs: Examples

$Q$ is acyclic.

Remark: Acyclicity allows loops and multi-edges.
Acyclic C2RPQs: Examples
Acyclic C2RPQs: Examples

\[ Q \] is not acyclic.

\[ U(Q) \]
Semantic acyclicity for UC2RPQs

A UC2RPQ is semantically acyclic if it is equivalent to a union of acyclic C2RPQs.

Question:
Is semantic acyclicity of UC2RPQs the same as semantic acyclicity of the underlying CQs?.

Semantic acyclicity of UC2RPQs is more general than semantic acyclicity of the underlying CQs.
Example of semantically acyclic CRPQ

The CRPQ

\[ L_1 \rightarrow L_2 \rightarrow L_3 \]

is equivalent to the acyclic CRPQ

\[ L_1 \cdot L_2 \cdot L_3 \]
Example of semantically acyclic CRPQ

The CRPQ

is equivalent to the acyclic CRPQ
Non semantically acyclic CRPQ

The following CRPQ is not semantically acyclic:

\[
\begin{array}{c}
\text{\$1} \\
\text{\$2} \\
\text{\$3}
\end{array}
\]

$\begin{array}{c}
\text{\$1} \\
\text{\$2} \\
\text{\$3}
\end{array}$
Semantic acyclicity of UC2RPQs is decidable

Our main result:

Theorem
There is a \textbf{2EXPSPACE} algorithm that on input a UC2RPQ \( Q \):

1. It checks whether \( Q \) is semantically acyclic;
2. if so, it outputs an equivalent union of acyclic C2RPQs \( Q' \) of exponential size.
Proof idea

Given UC2RPQ $Q$:

1. Compute a union of acyclic C2RPQs $Q'$ that is maximally contained in $Q$.
   - if $Q$ is semantically acyclic, it must be equivalent to $Q'$.
2. Check whether $Q$ is equivalent to $Q'$.
3. if so, output $Q'$. 
Approximations of UC2RPQs

The query $Q'$ is acyclic and maximally contained in $Q$.

In terms of [Barceló, Libkin, R., PODS 2012], $Q'$ is an approximation of $Q$ among the class of unions of acyclic UC2RPQs.

**Proposition**

There is an EXPSPACE algorithm that on input a UC2RPQ $Q$:
- it outputs an approximation of $Q$ – of exponential size – in the class of unions of acyclic C2RPQs.
Theorem

There is a $2\text{EXPSPACE}$ algorithm that on input a UC2RPQ $Q$:

1. It checks whether $Q$ is semantically acyclic;
2. if so, it outputs an equivalent union of acyclic C2RPQs $Q'$ of exponential size.

Given UC2RPQ $Q$:

1. Compute in $\text{EXPSPACE}$ a union of acyclic C2RPQs $Q'$ – of exponential size – that is maximally contained in $Q$.
2. Check whether $Q \subseteq Q'$ in $2\text{EXPSPACE}$.
3. If so, output $Q'$. 
Our method might not be optimal, but it cannot be improved dramatically:

Proposition

The problem of checking whether a UC2RPQ is semantically acyclic is EXPSPACE-hard, even for CRPQs.
Corollary

Evaluation of semantically acyclic UC2RPQs is fixed-parameter tractable:

\[
O(f(|Q|) + |G|^2 \cdot 2^{\text{poly}(|Q|)})
\]

where \(Q\) and \(Q'\) denote UC2RPQs and \(G\) represents the graph.
Evaluation of semantically acyclic UC2RPQs

Semantic acyclicity extends fixed-parameter tractability of acyclicity for UC2RPQs.

**Problem:**
Can semantically acyclic UC2RPQs be evaluated in polynomial time?
Outline

Motivation

Graph databases

Interlude on conjunctive queries

Semantic acyclicity of UC2RPQs

Open questions
Open questions

- Complexity of evaluation of semantically acyclic UC2RPQs.
- Extensions to wider tractable classes, e.g., bounded (hyper)treewidth.
- CSP for C2RPQs.