The Tractability Frontier of Well-designed SPARQL Queries

Miguel Romero (University of Oxford)

ACM PODS 2018, 12 June, Houston-USA
Well-designed SPARQL

**SPARQL**: standard query language for RDF graphs

**Well-designed SPARQL** (Perez, Arenas, Gutierrez 2006)

- Evaluation is coNP-complete (PSPACE-complete for SPARQL)

**This work:**

- Well-designed SPARQL restricted to **AND, OPTIONAL, UNION**
Tractable evaluation

Evaluating well-designed SPARQL becomes tractable for some classes

• Most general condition: **local tractability**
  (Letelier, Perez, Pichler, Skritek 2013; Barceló, Pichler, Skritek 2015)

**Main Question:**
Which classes of well-designed SPARQL queries can be evaluated in polynomial time?

**Our Contribution:**
The tractable classes are precisely those of **bounded domination width**
Well-designed Pattern Trees/Forests
(Letelier, Perez, Pichler, Skritek 2013)

Well-designed SPARQL queries with AND, OPTIONAL
= Well-designed Pattern Trees

Well-designed SPARQL queries with AND, OPTIONAL, UNION
= Well-designed Pattern Forests

In this talk:
We focus on (well-designed) pattern forests
Basics of RDF graphs and pattern trees/forests
RDF Graphs

Fix: set of *identifiers* $I$, set of *variables* $V$

**RDF Graph** = finite set of triples from $I \times I \times I$

$$(s, p, o) \quad s \xrightarrow{p} o$$
Conjunctive Queries (CQs) over RDF graphs

Fix: set of *identifiers* $I$, set of *variables* $V$

**Conjunctive query (CQ)** =
AND of triples from $(I \cup V) \times (I \cup V) \times (I \cup V)$ + free variables

$q(\texttt{?y, ?z}) = (\texttt{?x, p, o}) \text{ AND } (\texttt{?y, ?x, a}) \text{ AND } (\texttt{o, ?z, ?y}) \text{ AND } (\texttt{p, ?w, ?w})$

Answer of a CQ $q(X)$ over an RDF graph $G$:

$q(G) = \{h|x : h \text{ is a homomorphism from q to G}\}$

- **Full CQ** = All variables are free (no projection)
Well-designed Pattern Trees

Well-designed Pattern Tree =
(T, pat), where T is rooted tree and pat is a function mapping each node of T to a full CQ such that

- For each variable ?x, the set \{t in T | ?x in pat(t)\} is connected in T
Subtree $T'$ of $P$ = subtree of $T$ containing the root
$\text{pat}(T') = \text{AND of all the CQs in } \{\text{pat}(t) \mid t \in T'\}$
Well-designed Pattern Trees: semantics

Subtree $T'$ of $P$ = subtree of $T$ containing the root
$\text{pat}(T') = \text{AND}$ of all the CQs in $\{\text{pat}(t) \mid t \text{ in } T'\}$
Child of $T'$ = node not in $T'$ whose parent is in $T'$
Well-designed Pattern Trees: semantics

\[ P = (T, \text{pat}) \]

\[ h \text{ is in } P(G) \text{ iff there is a subtree } T' \text{ such that} \]

- \( h \) is a homomorphism from \( \text{pat}(T') \) to \( G \)
- for each child \( t \) of \( T' \), \( h \) cannot be extended to \( \text{pat}(T') \) AND \( \text{pat}(t) \)
Well-designed Pattern Forests

Well-designed Pattern Forest $= \text{Union of well-designed pattern trees}$

Answer of $F=\{P_1, \ldots, P_m\}$ over RDF graph $G$:

$$F(G) = P_1(G) \cup \ldots \cup P_m(G)$$
The Evaluation Problem

Let $C$ be a class of well-designed pattern forests

**EVAL(C)**

**Instance:** well-designed pattern forest $F$ in $C$, RDF graph $G$, mapping $h$

**Question:** does $h$ belong to $F(G)$?
Domination width and main theorem
Main Theorem

Theorem:
Assume \( \text{FPT} \not\subseteq \text{W}[1] \). Let \( C \) be a recursively enumerable class of well-designed pattern forests. Then the following are equivalent:

- \( \text{EVAL}(C) \) can be solved in polynomial time
- \( C \) has **bounded domination width**

Proof based on the corresponding characterisation for conjunctive queries (Dalmau, Kolaitis, Vardi 2002; Grohe 2003)

**Treewidth of a CQ** = measure of tree-likeness

\[
\text{ctw}(q(X)) : = \text{treewidth of the core of } q(X)
\]
The case of Conjunctive Queries

**Theorem** (Dalmau, Kolaitis, Vardi 2002; Grohe 2003)

Assume \( \text{FPT} \not\subseteq \text{W}[1] \). Let \( C \) be a recursively enumerable class of conjunctive queries of **bounded arity**. Then the following are equivalent:

- \( \text{CQ-EVAL}(C) \) can be solved in polynomial time
- \( C \) has **bounded ctw**

Tractability part via the **existential k-pebble game** (Kolaitis, Vardi 1995)

- Relaxation for checking existence of homomorphisms (complete, but not correct)
- Always correct for conjunctive queries \( q \) with \( \text{ctw}(q) < k \)
- Existence of a winning strategy for the Duplicator can be done in \( \text{poly time} \)

Hardness part via a reduction from the **clique problem** (\( \text{W}[1] \)-hardness)
The case of Conjunctive Queries

**Theorem (Dalmau, Kolaitis, Vardi 2002; Grohe 2003)**

Assume $\text{FPT} \not\subseteq W[1]$. Let $C$ be a recursively enumerable class of conjunctive queries of bounded arity. Then the following are equivalent:

- $\text{CQ-EVAL}(C)$ can be solved in polynomial time
- $C$ has **bounded ctw**

Can be extended to **unions** of CQs (UCQs) $Q(X) = \{q_1(X), \ldots, q_m(X)\}$

$$\text{ctw}(Q(X)) = \text{minimum } k \text{ such that for every } q_i(X), \text{ there is } q_j(X) \text{ such that } \text{ctw}(q_j(X)) \text{ is at most } k \text{ and } q_j(X) \text{ can be mapped to } q_i(X) \text{ via a homomorphism}$$
Domination width

\[ P = (T, \text{pat}) \]

Is \( h \) a “potential solution”?

\[ h \text{ in } P(G) \, ? \]

Can be computed in poly time
Domination width

\[ P = (T, \text{pat}) \]

\[ Q_{T'}(G) := \{ q_{t_1}(X), \ldots, q_{t_n}(X) \} \]

\[ h \text{ is not in } P(G) \iff h \text{ is in } Q_{T'}(G) \]
Domination width

$P=(T, \text{pat})$

$h$ in $P(G)$ ?

$h$ is not in $P(G)$ iff $h$ is in $Q^{T'}(G)$

$\text{dw}(P) := \text{maximum } ctw(Q^{T'}(X)), \text{ over all subtree } T'$

$X := \text{vars}(T')$

$UCQ \quad Q^{T'}(X) := \{q_{t_1}(X), \ldots, q_{t_n}(X)\}$

$CQ \quad q_{t_i}(X) := (\text{pat}(T') \text{ AND pat}(t_i))(X)$
Domination width

$P = (T, \text{pat})$

$dw(P) < k$

$CQ \ q_{ti}(X) := (\text{pat}(T') \ \text{AND} \ \text{pat}(t_i))(X)$

$dw(P) := \text{maximum } ctw(QT'(X))$, over all subtree $T'$
Domination width

\[ P = (T, \text{pat}) \]

\[ \text{dw}(P) < k \]

\[ h \text{ in } P(G) ? \]

\[ \text{CQ } q_{ti}(X) := (\text{pat}(T') \text{ AND pat}(ti))(X) \]

\[ \text{dw}(P) := \text{maximum } ctw(Q_{T'}(X)), \text{ over all subtree } T' \]
Domination width

\[ F = \{P_1, \ldots, P_m\} \]

\[ \text{T' AND T''} \]

rename new variables

h in \( P(G) \)?
Domination width

\[ F = \{P_1, \ldots, P_m\} \]

\[ h \text{ in } P(G) ? \]

\[ h \]

\[ G \]

\[ h \text{ is not in } F(G) \text{ iff } h \text{ is in } Q\{T', T''\}(X) \]

\[ Q\{T', T''\}(X) := \{\text{pat}(T') \text{ AND pat}(T'') + \text{choice of children}\} \]

\[ \text{(and renaming)} \]

\[ dw(F) = \text{maximum } ctw(QS(X)), \text{ over all set } S \text{ of subtrees} \]

over the same set of variables \( X \) and satisfying certain closure property
Main Theorem

**Theorem:**
Assume $\text{FPT} \not\subseteq \text{W}[1]$. Let $C$ be a recursively enumerable class of well-designed pattern forests. Then the following are equivalent:

- $\text{EVAL}(C)$ can be solved in polynomial time
- $C$ has **bounded domination width**

**Tractability part:**
Application of the existential $k$-pebble game as for the case of conjunctive queries (Dalmau, Kolaitis, Vardi 2002)

**Hardness part:**
Reduction from clique (Grohe 2003)
+ some basic properties of pattern forests with large $\text{dw}$
The case of UNION-free queries (pattern trees)
Branch Treewidth

$P = (T, \text{pat})$

Branch $B_t$ of $t$
**Proposition:**
For every well-designed pattern tree $P$, we have $\text{dw}(P) = \text{bw}(P)$.  

$$\text{bw}(P) := \text{maximum } \text{ctw}(\text{bt}(X)) \text{ over all node } t \text{ of } T$$

$$\text{CQ } \text{bt}(X) := (\text{pat(Bt)} \text{ AND pat(t))}(X)$$

$$X := \text{vars(Bt)}$$

$P = (T, \text{pat})$

Branch $B_t$ of $t$

Branch Treewidth
Final Remarks

Characterisation of tractable classes of pattern forests
(well-designed SPARQL restricted to AND, OPTIONAL, UNION)

- **Dichotomy**: A class C is either tractable or W[1]-hard

The \{AND, OPTIONAL, UNION\} fragment is maximal with this property:

- **Dichotomy fails** when we add **FILTER** (CQs with inequalities) and **SELECT** (Kroll, Pichler, Skritek 2016)

Open problem:

Characterise fixed-parameter tractable classes of queries with **SELECT**
(Recent characterisation for simple queries, Mengel, Skritek 2018)

Thank you!