Review: Datalog

A rule-based recursive query language

\[
\begin{align*}
\text{father}(alice, bob) \\
\text{mother}(alice, carla) \\
\text{Parent}(x, y) & \leftarrow \text{father}(x, y) \\
\text{Parent}(x, y) & \leftarrow \text{mother}(x, y) \\
\text{SameGeneration}(x, x) \\
\text{SameGeneration}(x, y) & \leftarrow \text{Parent}(x, v) \land \text{Parent}(y, w) \land \text{SameGeneration}(v, w)
\end{align*}
\]

There are three equivalent ways of defining Datalog semantics:

- Proof-theoretic: What can be proven deductively?
- Operational: What can be computed bottom up?
- Model-theoretic: What is true in the least model?

Datalog is more complex than FO query answering:

- ExpTime-complete for query and combined complexity
- P-complete for data complexity

Next question: Is Datalog also more expressive than FO query answering?
Expressivity
Where does Datalog fit in this picture?

- **Tree CQs**
  - $k$-Bounded Hypertree Width
  - everything (sub)polynomial

- **Conjunctive Queries**
  - Data compl.: $\text{AC}^0$; everything else: NP
  - equivalence/containment/emptiness: undec.

- **First-Order Queries**
  - Data compl.: $\text{AC}^0$, Comb./Query compl.: PSpace

- **Arbitrary Query Mappings**
  - everything undecidable

- **Polynomial Time Query Mappings**
Expressivity of Datalog

Datalog is P-complete for data complexity:

- Entailments can be computed in polynomial time with respect to the size of the input database $I$.
- There is a Datalog program $P$, such that all problems that can be solved in polynomial time can be reduced to the question whether $P$ entails some fact over a database $I$ that can be computed in logarithmic space.

$\sim\text{So Datalog can solve all polynomial problems?}$

No, it can’t. Many problems in P that cannot be solved in Datalog:

- **Parity**: Is the number of elements in the database even?
- **Connectivity**: Is the input database a connected graph?
- Is the input database a chain (or linear order)?
- …
Datalog Expressivity and Homomorphisms

How can we know that something is not expressible in Datalog?

A useful property: Datalog is “closed under homomorphisms”

**Theorem 13.1**: Consider a Datalog program $P$, an atom $A$, and databases $I$ and $J$. If $P$ entails $A$ over $I$, and there is a homomorphism $\mu$ from $I$ to $J$, then $\mu(P)$ entails $\mu(A)$ over $J$.

(By $\mu(P)$ and $\mu(A)$ we mean the program/atom obtained by replacing constants in $P$ and $A$, respectively, by their $\mu$-images.)

**Proof (sketch)**:

• Closure under homomorphism holds for conjunctive queries
• Single rule applications are like conjunctive queries
• We can show the claim for all $T_{P,I}^i$ by induction on $i$
Limits of Datalog Expressiveness

Closure under homomorphism shows many limits of Datalog

**Special case:** there is a homomorphism from $\mathcal{I}$ to $\mathcal{J}$ if $\mathcal{I} \subset \mathcal{J}$

$\leadsto$ Datalog entailments always remain true when adding more facts

- **Parity** cannot be expressed
- **Connectivity** cannot be expressed
- It cannot be checked if the input database is a chain
- Many FO queries with negation cannot be expressed (e.g., $\neg p(a)$)
- ... 

However this criterion is not sufficient!

Datalog cannot even express all polynomial time query mappings that are closed under homomorphism
Capturing PTime in Datalog

How could we extend Datalog to capture all query mappings in P?
\[ \sim \text{semipositive Datalog on an ordered domain} \]

**Definition 13.2:** Semipositive Datalog, denoted Datalog⊥, extends Datalog by allowing negated EDB atoms in rule bodies.
Datalog (semipositive or not) with a successor ordering assumes that there are special EDB predicates succ (binary), first and last (unary) that characterise a total order on the active domain.

Semipositive Datalog with a total order corresponds to standard Datalog on an extended version of the given database:

- For each ground fact \( r(c_1, \ldots, c_n) \) with \( \mathcal{I} \not\models r(c_1, \ldots, c_n) \), add a new fact \( \bar{r}(c_1, \ldots, c_n) \) to \( \mathcal{I} \), using a new EDB predicate \( \bar{r} \)
- Replace all uses of \( \neg r(t_1, \ldots, t_n) \) in \( P \) by \( \bar{r}(t_1, \ldots, t_n) \)
- Define extensions for the EDB predicates succ, first and last to characterise some (arbitrary) total order on the active domain.
Theorem 13.3: A Boolean query mapping defines a language in P if and only if it can be described by a query in semipositive Datalog with a successor ordering.

Example 13.4: We can express Connectivity for binary graphs as follows:

```
Reachable(x, x) ←
Reachable(x, y) ← Reachable(y, x)
Reachable(x, z) ← Reachable(x, y) ∧ edge(y, z)
    Connected(x) ← first(x)
Connected(y) ← Connected(x) ∧ succ(x, y) ∧ Reachable(x, y)
Accept() ← last(x) ∧ Connected(x)
```
Datalog Expressivity: Summary

The PTime capturing result is a powerful and exhaustive characterisation for semipositive Datalog with a successor ordering.

Situation much less clear for other variants of Datalog (as of 2018):

- What exactly can we express in Datalog without EDB negation and/or successor ordering?
  - Does a weaker language suffice to capture PTime? $\sim$ No!
  - When omitting negation, do we get query mappings closed under homomorphism? No!¹

- How about query mappings in PTime that are closed under homomorphism?
  - Does plain Datalog capture these? $\sim$ No!²
  - Does Datalog with successor ordering capture these? $\sim$ No!³

¹Counterexample on previous slide
²[A. Dawar, S. Kreutzer, ICALP 2008]
³[S. Rudolph, M. Thomazo, IJCAI 2016]: “We are somewhat baffled by this result: in order to express queries which satisfy the strongest notion of monotonicity, one cannot dispense with negation, the epitome of non-monotonicity.”
The Big Picture

Arbitrary Query Mappings
   everything undecidable

Polynomial Time Query Mappings
   = semipositive Datalog with a successor ordering

Datalog Queries
   Data compl.: PTime, Comb./Query compl.: ExpTime

First-Order Queries
   Data compl.: AC⁰, Comb./Query compl.: PSpace
   equivalence/containment/emptiness: undec.

Conjunctive Queries
   Data compl.: AC⁰; everything else: NP

k-Bounded Hypertree Width
   everything (sub)polynomial

Tree CQs

Note: languages that capture the same query mappings must have the same data complexity, but may differ in combined or in query complexity
Datalog Containment
How can Datalog query answering be implemented?
How can Datalog queries be optimised?

**Recall:** static query optimisation

- Query equivalence
- Query emptiness
- Query containment

$\sim$ all undecidable for FO queries, but decidable for (U)CQs
Learning from CQ Containment?

How did we manage to decide the question $Q_1 \sqsubseteq Q_2$ for conjunctive queries $Q_1$ and $Q_2$?

**Key ideas were:**

- We want to know if all situations where $Q_1$ matches are also matched by $Q_2$.
- We can simply view $Q_1$ as a database $I_{Q_1}$: the most general database that $Q_1$ can match to.
- Containment $Q_1 \sqsubseteq Q_2$ holds if $Q_2$ matches the database $I_{Q_1}$.

$\leadsto$ decidable in NP

A CQ $Q[x_1, \ldots, x_n]$ can be expressed as a Datalog query with a single rule

$\text{Ans}(x_1, \ldots, x_n) \leftarrow Q$

$\leadsto$ Could we apply a similar technique to Datalog?
Checking Rule Entailment

The containment decision procedure for CQs suggests a procedure for single Datalog rules:

- Consider a Datalog program $P$ and a rule $H \leftarrow B_1 \land \ldots \land B_n$.
- Define a database $I_{B_1 \land \ldots \land B_n}$ as for CQs:
  - For every variable $x$ in $H \leftarrow B_1 \land \ldots \land B_n$, we introduce a fresh constant $c_x$, not used anywhere yet.
  - We define $H^c$ to be the same as $H$ but with each variable $x$ replaced by $c_x$; similarly we define $B_i^c$ for each $1 \leq i \leq n$.
  - The database $I_{B_1 \land \ldots \land B_n}$ contains exactly the facts $B_i^c$ ($1 \leq i \leq n$).
- Now check if $H^c \in T_P^\infty (I_{B_1 \land \ldots \land B_n})$:
  - If no, then there is a database on which $H \leftarrow B_1 \land \ldots \land B_n$ produces an entailment that $P$ does not produce.
  - If yes, then $P \models H \leftarrow B_1 \land \ldots \land B_n$.
Example: Rule Entailment

Let $P$ be the program

\[
\text{Ancestor}(x, y) \leftarrow \text{parent}(x, y) \\
\text{Ancestor}(x, z) \leftarrow \text{parent}(x, y) \land \text{Ancestor}(y, z)
\]

and consider the rule $\text{Ancestor}(x, z) \leftarrow \text{parent}(x, y) \land \text{parent}(y, z)$.

Then $I_{\text{parent}(x, y) \land \text{parent}(y, z)} = \{\text{parent}(c_x, c_y), \text{parent}(c_y, c_z)\}$ (abbreviate as $I$)

We can compute $T^\infty_P(I)$:

\[
T^0_P(I) = I \\
T^1_P(I) = \{\text{Ancestor}(c_x, c_y), \text{Ancestor}(c_y, c_z)\} \cup I \\
T^2_P(I) = \{\text{Ancestor}(c_x, c_z) \cup T^1_P(I) \}
\]

Therefore, $\text{Ancestor}(x, z)^c = \text{Ancestor}(c_x, c_z) \in T^\infty_P(I)$, so $P$ entails $\text{Ancestor}(x, z) \leftarrow \text{parent}(x, y) \land \text{parent}(y, z)$. 

Markus Krötzsch, 29th May 2019

Database Theory
Deciding Datalog Containment?

Idea for two Datalog programs $P_1$ and $P_2$:

- If $P_2 \models P_1$, then every entailment of $P_1$ is also entailed by $P_2$.
- In particular, this means that $P_1$ is contained in $P_2$.
- We have $P_2 \models P_1$ if $P_2 \models H \leftarrow B_1 \land \ldots \land B_n$ for every rule $H \leftarrow B_1 \land \ldots \land B_n \in P_1$.
- We can decide $P_2 \models H \leftarrow B_1 \land \ldots \land B_n$.

Can we decide Datalog containment this way?

$\sim$ No! In fact, Datalog containment is undecidable. What’s wrong?
Consider the Datalog queries \( \langle A, P_1 \rangle \) and \( \langle B, P_2 \rangle \):

- Clearly, \( \langle A, P_1 \rangle \) and \( \langle B, P_2 \rangle \) are equivalent (and mutually contained in each other).
- However, \( P_2 \) entails no rule of \( P_1 \) and \( P_1 \) entails no rule of \( P_2 \).

\[\sim \] IDB predicates do not matter in Datalog, but predicate names matter in first-order implications.
Datalog is a fragment of second-order logic: IDB predicates are like variables that can take any set of tuples as value!

**Example 13.5:** The previous query $\langle A, P_1 \rangle$ can be expressed by the formula

$$\forall A. \left( \begin{array}{l}
\forall x, y. A(x, y) \leftarrow \text{parent}(x, y) \\
\forall x, y, z. A(x, z) \leftarrow \text{parent}(x, y) \land A(y, z)
\end{array} \right) \rightarrow A(v, w)$$

- This is a formula with two free variables $v$ and $w$.
  $\leadsto$ query with two result variables
- Intuitive semantics: “$\langle c, d \rangle$ is a query result if $A(c, d)$ holds for all possible values of $A$ that satisfy the rules”
  $\leadsto$ Datalog semantics in other words

We can express any Datalog query like this, with one second-order variable per IDB predicate.
A Datalog program looks like a set of first-order implications, but it has a second-order semantics.

We have already seen that Datalog can express things that are impossible to express in FO queries – that’s why we introduced it!\(^1\)

Consequences for query optimisation:

- Entailment between sets of first-order implications is decidable (shown above)
- Containment between Datalog queries is not decidable (shown next)

---

\(^1\) Possible confusion when comparing FO and Datalog: entailments of first-order implications agree with answers of Datalog queries, so it seems we can break the FO locality restrictions; but query answering is model checking not entailment; FO model checking is much weaker than second-order model checking.
Undecidability of Datalog Query Containment

A classical undecidable problem:

**Post Correspondence Problem:**
- Input: two lists of words $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$
- Output: “yes” if there is a sequence of indices $i_1, i_2, i_3, \ldots, i_m$ such that $\alpha_{i_1}\alpha_{i_2}\alpha_{i_3}\cdots\alpha_{i_m} = \beta_{i_1}\beta_{i_2}\beta_{i_3}\cdots\beta_{i_m}$.

$\Rightarrow$ we will reduce PCP to Datalog containment

We need to define Datalog programs that work on databases that encode words:
- We represent words by chains of binary predicates
- Binary EDB predicates represent letters
- For each letter $\sigma$, we use a binary EDB predicate $\text{letter}[\sigma]$
- We assume that the words $\alpha_i$ have the form $a_i^1 \cdots a_i^{\vert \alpha_i \vert}$, and that the words $\beta_i$ have the form $b_i^1 \cdots b_i^{\vert \beta_i \vert}$
A program $P_1$ to recognise potential PCP solutions.

Rules to recognise words $\alpha_i$ and $\beta_i$ for every $i \in \{1, \ldots, m\}$:

$$A_i(x_0, x_{|\alpha_i|}) \leftarrow \text{letter}[a^i_1](x_0, x_1) \land \ldots \land \text{letter}[a^i_{|\alpha_i|-1}](x_{|\alpha_i|-1}, x_{|\alpha_i|})$$
$$B_i(x_0, x_{|\beta_i|}) \leftarrow \text{letter}[b^i_1](x_0, x_1) \land \ldots \land \text{letter}[b^i_{|\beta_i|-1}](x_{|\beta_i|-1}, x_{|\beta_i|})$$

Rules to check for synchronised chains (for all $i \in \{1, \ldots, m\}$):

$$\text{PCP}(x, y_1, y_2) \leftarrow A_i(x, y_1) \land B_i(x, y_2)$$
$$\text{PCP}(x, z_1, z_2) \leftarrow \text{PCP}(x, y_1, y_2) \land A_i(y_1, z_1) \land B_i(y_2, z_2)$$
$$\text{Accept()} \leftarrow \text{PCP}(x, z, z)$$
Solving PCP with Datalog Containment (2)

**Example:** $\alpha_1 = aa$, $\beta_1 = a$, $\alpha_2 = b$, $\beta_2 = aab$

Example for an intended database and least model (selected parts):

Additional IDB facts that are derived (among others):

- PCP(1, 3, 2)
- PCP(1, 5, 3)
- PCP(1, 6, 6)
- Accept()
**Example:** $\alpha_1 = aaaaa$, $\beta_1 = bbb$

**Problem:** $P_1$ also accepts some unintended cases

Additional IDB facts that are derived:

\[ \text{PCP}(1, 6, 6) \quad \text{Accept()} \]
Solving PCP with Datalog Containment (4)

**Solution:** specify a program $P_2$ that recognises all unwanted cases

$P_2$ consists of the following rules (for all letters $\sigma, \sigma'$):

\[
\begin{align*}
EP(x, x) & \leftarrow \\
EP(y_1, y_2) & \leftarrow EP(x_1, x_2) \land \text{letter}[\sigma](x_1, y_1) \land \text{letter}[\sigma](x_2, y_2) \\
\text{Accept}() & \leftarrow EP(x_1, x_2) \land \text{letter}[\sigma](x_1, y_1) \land \text{letter}[\sigma'](x_2, y_2) \quad \sigma \neq \sigma' \\
NEP(x_1, y_2) & \leftarrow EP(x_1, x_2) \land \text{letter}[\sigma](x_2, y_2) \\
NEP(x_1, y_2) & \leftarrow NEP(x_1, x_2) \land \text{letter}[\sigma](x_2, y_2) \\
\text{Accept}() & \leftarrow NEP(x, x)
\end{align*}
\]

**Intuition:**

- EP defines equal paths (forwards, from one starting point)
- NEP defines paths of different length (from one starting point to the same end point)

$\leadsto P_2$ accepts all databases with distinct parallel paths
What does it mean if \( \langle \text{Accept}, P_1 \rangle \) is contained in \( \langle \text{Accept}, P_2 \rangle \)?

The following are equivalent:

- All databases with potential PCP solutions also have distinct parallel paths.
- Databases without distinct parallel paths have no PCP solutions.
- Linear databases (words) have no PCP solutions.
- The answer to the PCP is “no”.

\[ \sim \text{If we could decide Datalog containment, we could decide PCP} \]

**Theorem 13.6:** Containment and equivalence of Datalog queries are undecidable.

(Note that emptiness of Datalog queries is trivial)
Datalog cannot express all query mappings in P . . .

. . . but semipositive Datalog with a successor ordering can

First-order rule entailment is decidable . . .

. . . but Datalog containment is not.

Next question:

- How can we implement Datalog in practice?