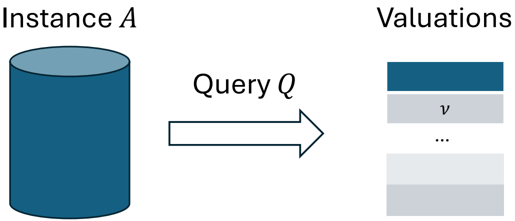
Postulates for Provenance: instance-based provenance for first-order logic

Bart Bogaerts, **Maxime Jakubowski** & Jan Van den Bussche

June 12th, 2024

Instance-based Provenance

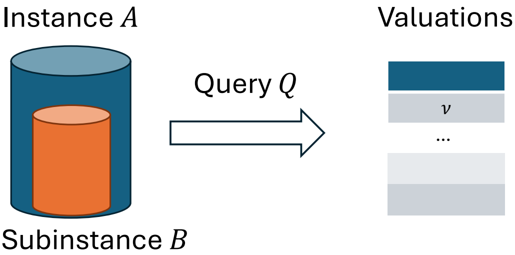
Given a query result setting:



What part of instance A makes ν a valuation of Q?

Instance-based Provenance

The provenance result setting:



We would like to identify an insteresting subinstance $B \subseteq A$.

Provenance Results

A *query result* $\mathbf{r}$ is a tuple $(\mathbf{d}, A, \nu, \varphi)$ with:

- $\mathbf{d}$ the domain;
- A a *three-valued* instance (a set of positive and negative facts);
- ν a valuation; and
- φ a first-order logic query

such that $\nu(\varphi)$ is *true* in A under *certain answer semantics*.

We say $\mathbf{r}$ is *total* if A is two-valued.

Provenance Results

A *query result* $\mathbf{r}$ is a tuple $(\mathbf{d}, A, \nu, \varphi)$ with:

- $\mathbf{d}$ the domain;
- A a *three-valued* instance (a set of positive and negative facts);
- ν a valuation; and
- φ a first-order logic query

such that $\nu(\varphi)$ is *true* in A under *certain answer semantics*.

We say $\mathbf{r}$ is *total* if A is two-valued.

A *provenance result* $\mathbf{p}$ is a tuple $(\mathbf{r}, B)$ with:

- $\mathbf{r} = (\mathbf{d}, A, \nu, \varphi)$ a *total* query result; and
- B a (three-valued) subinstance of A.

such that $(\mathbf{d}, B, \nu, \varphi)$ is a query result.

We say B is *sufficient* for $\mathbf{r}$.

Example of a Provenance Result

To explain that a tuple t belongs to the difference $R - S$ of two relations, the positive and negative information is equally important.

Example of a Provenance Result

To explain that a tuple t belongs to the difference $R - S$ of two relations, the positive and negative information is equally important.

Consider the total query result $\mathbf{r} = (\mathbf{d}, A, \nu, \varphi)$ with:

- $\mathbf{d} = \{a, b\}$;
- $A = \{R(a), R(b), S(a), \neg S(b)\}$;
- φ is $R(x) \wedge \neg S(x)$;
- $\nu = \{x \mapsto b\}$

Example of a Provenance Result

To explain that a tuple t belongs to the difference $R - S$ of two relations, the positive and negative information is equally important.

Consider the total query result $\mathbf{r} = (\mathbf{d}, A, \nu, \varphi)$ with:

- $\mathbf{d} = \{a, b\}$;
- $A = \{R(a), R(b), S(a), \neg S(b)\}$;
- φ is $R(x) \wedge \neg S(x)$;
- $\nu = \{x \mapsto b\}$

A sensible provenance result would be $(\mathbf{r}, \{R(b), \neg S(b)\})$.

Provenance Relations

A *provenance relation* is an infinite set Π of provenance results that is *total* and *generic*.

Examples:

- Π^{tok}: for every query result, B is the set of tokens from the provenance polynomial.
- Π^{cf}: for every query result, B is the set of causal facts.
- $\Pi_{\cap}^{tokcf}$: for every query result, B is the intersection of tokens from the polynomial and the causal facts.

Provenance Relations

A *provenance relation* is an infinite set Π of provenance results that is *total* and *generic*.

Examples:

- Π^{tok}: for every query result, B is the set of tokens from the provenance polynomial.
- Π^{cf}: for every query result, B is the set of causal facts.
- $\Pi^{tokcf}_{\cap}$: for every query result, B is the intersection of tokens from the polynomial and the causal facts.

These were examples of **Deterministic** provenance relations.

Other example:

- Π^{mon}: for every query result B is the set of tokens from one monomial of the provenance polynomial.

Provenance Relations

A *provenance relation* is an infinite set Π of provenance results that is *total* and *generic*.

Examples:

- Π^{tok}: for every query result, B is the set of tokens from the provenance polynomial.
- Π^{cf}: for every query result, B is the set of causal facts.
- $\Pi^{tokcf}_{\cap}$: for every query result, B is the intersection of tokens from the polynomial and the causal facts.

These were examples of **Deterministic** provenance relations.

Other example:

- Π^{mon}: for every query result B is the set of tokens from one monomial of the provenance polynomial.

$\Rightarrow$ Our goal is **not** to propose an ultimate provenance semantics for FO, but to investigate different desiderata.

Provenance Polynomials

We adapt the notion of a provenance polynomial for FO from *Grädel and Tannen (2017)* to the three-valued setting by interpreting unknown facts as 0.

For example: let $\mathbf{d} = \{a, b\}$,

- $A = \{P(a), P(b), \neg Q(a)\}$, and
- φ is $\exists x(P(x) \land \neg Q(x))$

Then $pol(\mathbf{d}, A, \varepsilon, \varphi) = P(a)\overline{Q(a)}$

We also have:

Lemma

Let $\mathbf{r}$ *be a total query result. The set of all tokens of the polynomial for* $\mathbf{r}$ *(denoted* $tokens(\mathbf{r})$*) is sufficient for* $\mathbf{r}$*. In other words,* $(\mathbf{r}, tokens(\mathbf{r}))$ *is a provenance result.*

Causality for Query Results

We adapt the notion of an *actual cause* from *Meliou et al. (2010)*:

Definition

A *supercause* of a query result $\mathbf{r}$ with instance A is a subinstance $C \subseteq A$ such that "flipping" the facts of C in A makes $\nu(\varphi)$ unknown or false. An *actual cause* (or simply *cause*) is a minimal supercause.

We also have:

Lemma

Let $\mathbf{r}$ *be a total query result. The set of all causal facts for* $\mathbf{r}$ *(denoted* $(cf(\mathbf{r}))$ *) is sufficient for* $\mathbf{r}$*. In other words,* $(\mathbf{r}, cf(\mathbf{r}))$ *is a provenance result.*

Polynomials and Causality

The "Hitting-set lemma":

Lemma

Let $\mathbf{r}$ *be a total query result with instance* A *and let* $B \subseteq A$*. Then* B *is sufficient for* $\mathbf{r}$ *if and only if* B *intersects with every cause of* $\mathbf{r}$*.*

We therefore have:

Theorem

For any total query result $\mathbf{r}$*, the intersection* $\mathit{cf}(\mathbf{r}) \cap \mathit{tokens}(\mathbf{r})$ *is sufficient for* $\mathbf{r}$*.*

Properties of Provenance *Results*

Let $\mathbf{p} = (\mathbf{r}, B)$ be a provenance result, with $\mathbf{r}$ a total query result with instance A:

- $\mathbf{p}$ is **proof preserving** (pp) if the provenance polynomial is the same in instances A and B.
- $\mathbf{p}$ is **proof containing** (k) if subinstance B contains at least the tokens of one monomial.
- $\mathbf{p}$ is **proof-relevant** (pr) if B *consists only of* tokens from the polynomial.
- $\mathbf{p}$ is **cause preserving** (cp) if the causes are the same in instances A and B.
- $\mathbf{p}$ is **cause containing** (cc) if B contains a cause from A.
- $\mathbf{p}$ is **cause-relevant** (cr) if B *consists only of* causal facts.

These are the *basic* properties. Let X be a set of basic properties:

- $\mathbf{p}$ is **minimal for** X ($min(X)$) if B is minimal such that $\mathbf{p}$ satisfies all basic properties in X.

Postulates for Provenance *Relations*

We can lift the properties for provenance results to *Postulates* by requiring that every element in the provenance relation satisfies it:

- **Polynomial Preservation** (PP): every $\mathbf{p} \in \Pi$ is proof preserving.
- **Proof Containing** (K): every $\mathbf{p} \in \Pi$ is proof containing.
- **Proof Relevance** (PR): every $\mathbf{p} \in \Pi$ is proof-relevant.
- **Cause Preservation** (CP): every $\mathbf{p} \in \Pi$ is cause preserving.
- **Cause Containing** (CC): every $\mathbf{p} \in \Pi$ is cause containing.
- **Causal Relevance** (CR): every $\mathbf{p} \in \Pi$ is cause-relevant.
- **Minimal for** X $(\textsc{Min}(X))$: every $\mathbf{p} \in \Pi$ is *min*(X).

Finally, we also have:

- **Determinism** (D): for every query result $\mathbf{r}$, there is exactly one provenance result $(\mathbf{r}, B)$ in Π.

$\Rightarrow$ We say a set of postulates X is *satisfiable* if there exists a provenance relation that satisfies all postulates from X.

Satisfiability

We exhaustively investigated which sets of postulates are satisfiable.

For example:

- Π^{tok} satisfies $\mathrm{PP}, \mathrm{K}, \mathrm{PR}$ and D
- Π^{cf} satisfies $\mathrm{CP}, \mathrm{CC}, \mathrm{CR}$ and D
- $\Pi^{tokcf}_{\cap}$ satisfies PR, CR and D
- Π^{mon} satisfies $\textsc{Min}(\mathsf{k})$ and PR

Satisfiability: you can't have it all

Not every set of postulates is satisfiable!

Example (1)

The set $\{\mathrm{PR}, \mathrm{CC}\}$ is not satisfiable because:

- there must exist a provenance result for the total query result: $\mathbf{r} = (\{P, Q\}, \varphi)$ where φ is $P \vee (\neg P \wedge Q)$
- since $pol(\mathbf{r}) = P$, we have $tokens(\mathbf{r}) = \{P\}$
- since $\varphi \equiv P \vee Q$, there is only one cause: $\{P, Q\}$
- however, we require a provenance result with subinstance B to: satisfy: $\{P, Q\} \subseteq B \subseteq \{P\}$.

Satisfiability: you can't have it all

Example (2)

The set $\{\textsc{Min}(\mathrm{pp}, \mathrm{cc}), \mathrm{D}\}$ is not satisfiable because:

- there must exist a provenance result for the total query result: $\mathbf{r} = (\{P, Q, R\}, \varphi)$ where φ is $(P \wedge Q \wedge \neg R) \vee R$
- we have $pol(\mathbf{r}) = R$
- we have two causes: $\{P, R\}$ and $\{Q, R\}$
- both causes are sufficient subinstances satisfying pp and cc

However, without Determinism it is clearly satisfiable!

Concluding Remarks

- We deal with negation by considering three-valued provenance subinstances.
- Our postulates focus on incorporating different postulates from the literature and result in interesting provenance relations:
  - some deterministic: $\Pi^{tok}, \Pi^{cf}, \Pi^{tokcf}_{\cap}$
  - some not: Π^{mon}, minimally sufficient subinstances, ...
- We developed a general framework to study different provenance postulates.