Algorithmic Aspects of Semiring Provenance for Stratified Datalog

Matthias Naaf

Logic and Algebra for Query Evaluation, Berkeley 2023
Algorithmic Aspects

Semiring Provenance for Stratified Datalog

Computing Greatest Fixed Points
(in absorptive semirings)

Circuit Representations
Why Greatest Fixed Points?
Semiring Semantics for Datalog

Datalog

\[ T_{xy} : \neg E_{xy} \]
\[ T_{xy} : \neg E_{xz}, T_{zy} \]

\[ \begin{align*}
T_{uv} &= a \lor (a \land T_{vv}) \lor (d \land T_{wv}) \\
T_{uw} &= d \lor (d \land T_{ww}) \lor (a \land T_{vw}) \\
\end{align*} \]

Equation System

\( T_{uv} = a \lor (a \land T_{vv}) \lor (d \land T_{wv}) \)
\( T_{uw} = d \lor (d \land T_{ww}) \lor (a \land T_{vw}) \)
\( \vdots \)
Semiring Semantics for Datalog

Semantics: Least solution

- Power series: $T_{uw}^* = d + ac + abc + ab^2c + ab^3c + \ldots$
- PosBool: $T_{uw}^* = d \lor (a \land c)$

Equation System

$T_{uv} = a + (a \cdot T_{vv}) + (d \cdot T_{wv})$

$T_{uw} = d + (d \cdot T_{ww}) + (a \cdot T_{vw})$

$\vdots$
Semiring Semantics for Datalog

Datalog

\[ T_{xy} :- E_{xy} \]
\[ T_{xy} :- E_{xz}, T_{zy} \]

Equation System

\[
T_{uv} = a + (a \cdot T_{vv}) + (d \cdot T_{wv}) \\
T_{uw} = d + (d \cdot T_{ww}) + (a \cdot T_{vw}) \\
\vdots
\]

Semantics: Least solution

- Power series: \[ T^*_{uw} = d + ac + abc + ab^2c + ab^3c + \ldots \]
- PosBool: \[ T^*_{uw} = d \lor (a \land c) \]
- Tropical: \[ T^*_{uw} = \min(12, 2 + 5) = 7 \]

(\text{Green, Karvounarakis, Tannen, PODS'07})
Semiring Semantics for Stratified Datalog

**Stratified Datalog**

\[ T_{xy} :\neg E_{xy} \]
\[ T_{xy} :\neg E_{xz}, T_{zy} \]
\[ N_{xy} :\neg T_{xy} \]

**Negation:** can be defined in some semirings

- **PosBool:**
  \[ T_{uw}^* = d \lor (a \land c) \]
  \[ N_{uw}^* = \overline{T_{uw}^*} = \overline{d} \land (\overline{a} \lor \overline{c}) \]

- **Polynomials:** \( \overline{a^2} = ? \)
- **Tropical:** \( \overline{7} = ? \)

but not clear how do to it in general
Semiring Semantics for Stratified Datalog

Stratified Datalog

\[ T_{xy} := E_{xy} \]
\[ T_{xy} := E_{xz}, T_{zy} \]
\[ N_{xy} := \neg T_{xy} \]

Equation System

\[ T_{uv} = a + (a \cdot T_{vv} + d \cdot T_{wv}) \]
\[ T_{uw} = d + (d \cdot T_{ww} + a \cdot T_{vw}) \]

\[ \Rightarrow \text{Least solution} \]

Dualized System

\[ N_{uv} = \overline{a} \cdot (\overline{a} + N_{vv}) \cdot (\overline{d} + N_{wv}) \]
\[ N_{uw} = \overline{d} \cdot (\overline{d} + N_{ww}) \cdot (\overline{a} + N_{vw}) \]

\[ \Rightarrow \text{Greatest solution} \]
Motivation II: Fixed-point Logic

$$\text{gfp } R_x. \exists y(Exy \land Ry)](v)$$

“there is an infinite path from \(v\)"

\[R_a = 1 + R_a\]

\[R_b = \min(1 + R_a, 20 + R_c)\]

\[R_c = 0 + R_c\]
\[ \text{gfp } R_x. \exists y (E_{xy} \land R_y) \] \( (v) \) 

"there is an infinite path from \( v \)"

\[ Ra = 1 + Ra \]
\[ Rb = \min(1 + Ra, 20 + Rc) \]
\[ Rc = 0 + Rc \]

\[ Ra^* = \infty \]
\[ Rb^* = 20 \]
\[ Rc^* = 0 \]

Greatest Solution
Computing Greatest Fixed Points
Naive Approach

Goal:
Compute greatest fixed point of a polynomial operator

\[
\begin{bmatrix}
R_a \\
R_b \\
R_c
\end{bmatrix}
= \begin{bmatrix}
1 + R_a \\
\min(1 + R_a, 20 + R_c) \\
0 + R_c
\end{bmatrix}
\]

Iteration:

\[
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\rightarrow
\begin{bmatrix}
1 \\
1 \\
0
\end{bmatrix}
\rightarrow
\begin{bmatrix}
2 \\
2 \\
0
\end{bmatrix}
\rightarrow
\cdots
\rightarrow
\begin{bmatrix}
\infty \\
20 \\
0
\end{bmatrix}
\]
Naive Approach

**Goal:** Compute greatest fixed point of a polynomial operator

\[
F : \begin{pmatrix}
R_a \\
R_b \\
R_c \\
\end{pmatrix} \mapsto \begin{pmatrix}
1 + R_a \\
\min(1 + R_a, 20 + R_c) \\
0 + R_c \\
\end{pmatrix}
\]
Naive Approach

Goal: Compute greatest fixed point of a polynomial operator

\[ F : \begin{pmatrix} R_a \\ R_b \\ R_c \end{pmatrix} \mapsto \begin{pmatrix} 1 + R_a \\ \min(1 + R_a, 20 + R_c) \\ 0 + R_c \end{pmatrix} \]

Iteration:

\[
\begin{pmatrix}
0 \\
0 \\
0
\end{pmatrix} \mapsto \begin{pmatrix}
1 \\
1 \\
0
\end{pmatrix} \mapsto \begin{pmatrix}
2 \\
2 \\
0
\end{pmatrix} \mapsto \cdots \mapsto \begin{pmatrix}
20 \\
20 \\
0
\end{pmatrix} \mapsto \begin{pmatrix}
21 \\
20 \\
0
\end{pmatrix} \mapsto \begin{pmatrix}
22 \\
20 \\
0
\end{pmatrix} \mapsto \cdots \mapsto \begin{pmatrix}
\infty \\
20 \\
0
\end{pmatrix}
\]
Faster Computation

**Main Result**

Let \((K, +, \cdot, 0, 1)\) be an absorptive, fully-continuous semiring. For a polynomial operator \(F: K^n \to K^n\),

\[
\text{lfp}(F) = F^n(0), \quad \text{gfp}(F) = F^n(F^n(1)^\infty).
\]

We only need a *polynomial number* of semiring operations:

\[
\begin{align*}
(0) \to (1) \to (2) \to (3) \to \infty \\
0 \to 1 \to 2 \to 3 \to \infty
\end{align*}
\]
Which Semirings?

1. **Fully continuous**
   - Natural order: \( a \leq a + b \)
   - Each chain has supremum \( \bigcup C \) and infimum \( \bigcap C \), these commute with \(+/\cdot\)

2. **Absorption**
   - \( a + a \cdot b = a \) \(\iff\) \( 1 \) is greatest element \(\iff\) \( a \cdot b \leq a \)
Which Semirings?

1. **Fully continuous**
   - Natural order: \( a \leq a + b \)
   - Each chain has supremum \( \bigcup C \) and infimum \( \bigcap C \), these commute with \(+/\cdot\).

2. **Absorption**
   - \( a + a \cdot b = a \iff 1 \) is greatest element \( \iff a \cdot b \leq a \)

**Infinitary Power**

For \( a \in K \) we define: \( a^\infty := \bigcap_{n<\omega} a^n \)

**Remember:**

Decreasing multiplication
Proof Overview

Main Result

Let \((K, +, \cdot, 0, 1)\) be an absorptive, fully-continuous semiring. For a polynomial operator \(F: K^n \to K^n\),

\[
\lfp(F) = F^n(0), \quad \gfp(F) = F^n(F^n(1)\infty).
\]

Proof sketch:

derivation trees + absorption
Derivation Trees

\[
\begin{pmatrix}
X \\
Y \\
Z
\end{pmatrix} \mapsto \begin{pmatrix}
\min(5, 1+X+Y) \\
2+Z+Z \\
\min(3, 1+Z)
\end{pmatrix}
\]

inspired by Newton's method
(Esparza, Kiefer, Luttenberger, JACM'10)
Derivation Trees

\[
\begin{pmatrix}
X \\
Y \\
Z
\end{pmatrix} \rightarrow \begin{pmatrix}
\min(5, 1 + X + Y) \\
2 + Z + Z \\
\min(3, 1 + Z)
\end{pmatrix}
\]

inspired by Newton’s method
(Esparza, Kiefer, Luttenberger, JACM’10)

\[
\text{lfp} = \min\{\text{cost}(\triangle) \mid \text{finite } \triangle\}
\]

\[
\text{gfp} = \min\{\text{cost}(\triangle) \mid \text{finite } \triangle, \text{infinite } \triangle\}
\]
**Observation:** Prefixes of \( \_ \) correspond to iteration steps.

\[
\min\left( F(1), F^2(1), F^3(1), \ldots \right) = \bigcap_{n<\omega} \min\left\{ \text{cost}(\_), \text{finite/infinite} \_ \right\} = \text{gfp}(F)
\]
Absorption on Derivation Trees

If each coefficient \( ɔ_2 \) occurs more often in \( \mathcal{T} \) than in \( \mathcal{S} \), then \( \text{cost}(\mathcal{T}) \) is absorbed by \( \text{cost}(\mathcal{S}) \).

complicated tree \( \mathcal{T} \)

nice tree \( \mathcal{S} \)
Absorption on Derivation Trees

If each coefficient $2$ occurs more often in $\bullet$ than in $\star$, then cost($\bullet$) is absorbed by cost($\star$).

complicated tree $\bullet$  
ultimately periodic  
nice tree $\star$

\[
\begin{align*}
\text{cost} & \geq \text{cost} \\
\text{deterministic} & \geq \text{deterministic}
\end{align*}
\]

\[
\begin{align*}
\text{deterministic} & \geq \text{deterministic} \\
& \quad \text{for } n
\end{align*}
\]
Computing Nice Trees

**Main Result**

\[
gfp(F) = \min \left\{ \text{cost}(A) \mid \text{nice } \right\} = \ldots
\]
Computing Nice Trees

Main Result

\[
gfp(F) = \min \{ \text{cost}(\text{nice tree}) \mid \text{nice tree} \} = F^n(F^n(1)^\infty)
\]
Back to Datalog: Circuits
Circuits for Datalog Provenance

**Problem:** Provenance in polynomial semirings can become large

**Datalog**

\[
\begin{align*}
T_{xy} & : = Exy \\
T_{xy} & : = Exz, T_{zy}
\end{align*}
\]

**PosBool:** \( T_{uw}^* = ace + acf + ade + adf + bce + bcf + bde + bdf \)

**Solution:** Represent provenance computation by a small circuit
Circuits for Datalog Provenance

**Equation System**

\[
T_{uv} = a + (a \cdot T_{vv}) + (d \cdot T_{wv}) \\
T_{uw} = d + (d \cdot T_{ww}) + (a \cdot T_{vw}) \\
\vdots
\]

Recall

\[
lfp(F) = F^n(0)
\]
Circuits for Stratified Datalog

Strat. Datalog

\[ T_{xy} : \neg E_{xy} \]
\[ T_{xy} : \neg E_{xz}, T_{zy} \]
\[ N_{xy} : \neg T_{xy} \]

\[ \text{dualize circuit} \]
\[ a / \bar{a} \]

\[ \begin{array}{c}
  T^*_w \\
  T^*_u \\
  \vdots \\
  0 \\
  0 \\
  0 \\
  0
\end{array} \]
Circuits for Stratified Datalog

**Strat. Datalog**

\[ T_{xy} :\neg E_{xy} \]
\[ T_{xy} :\neg E_{xz}, T_{zy} \]

\[ N_{xy} := \neg T_{xy} \]

\[ \text{gfp}(F) = F^n( F^n(1)_{\infty} ) \]

\[ T^*_{uw}, T^*_{uw} \]

\[ N^*_{uw} N^*_{uw} \]

\[ n \]

\[ n \]

\[ \text{dualize circuit} \]

\[ + / \cdot \]

\[ a / \bar{a} \]

\[ 0 0 0 0 \]

\[ 1 1 1 1 \]
Summary

Computing greatest fixed points
- In absorptive semirings: $\text{gfp}(F) = F^n(F^n(1)\infty)$

Semiring provenance for stratified Datalog
- Negation: greatest solution to dual equation system (“negation normal form”)
- Circuit representations for Datalog can be generalized
Summary

Computing greatest fixed points
▶ In absorptive semirings: \( \text{gfp}(F) = F^n ( F^n(1) \infty ) \)

Semiring provenance for stratified Datalog
▶ Negation: greatest solution to dual equation system ("negation normal form")
▶ Circuit representations for Datalog can be generalized

Questions

1. Applications
   ▶ LFP: strategies in infinite games
   ▶ Stratified Datalog: ?

2. Alternating fixed points
   ▶ Is the main result applicable?
   ▶ Quasipolynomial time?