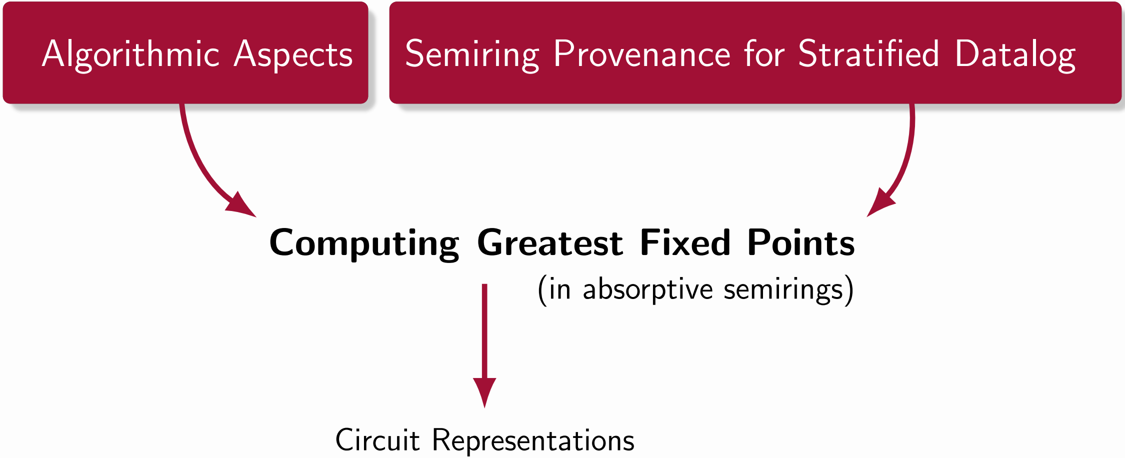
Algorithmic Aspects of Semiring Provenance for Stratified Datalog

Matthias Naaf

Logic and Algebra for Query Evaluation, Berkeley 2023

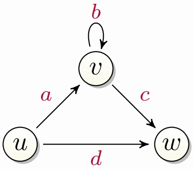

Algorithmic Aspects
Semiring Provenance for Stratified Datalog
Computing Greatest Fixed Points
(in absorptive semirings)
Circuit Representations

Why Greatest Fixed Points?

Semiring Semantics for Datalog

Datalog

$$Txy \;{:}\!-\; Exy$$
$$Txy \;{:}\!-\; Exz, Tzy$$



$\Rightarrow$

Equation System

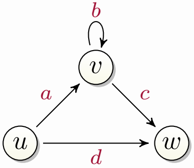
$$T_{uv} = a \vee (a \wedge T_{vv}) \vee (d \wedge T_{wv})$$
$$T_{uw} = d \vee (d \wedge T_{ww}) \vee (a \wedge T_{vw})$$
$$\vdots$$

Semiring Semantics for Datalog

Datalog

$Txy \mathrel{:-} Exy$

$Txy \mathrel{:-} Exz, Tzy$



$\Rightarrow$

Equation System

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

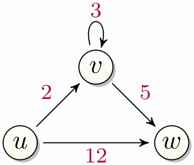
Semantics: Least solution

- Power series: $T^*_{uw} = d + ac + abc + ab^2c + ab^3c + \dots$
- PosBool: $T^*_{uw} = d \vee (a \wedge c)$

Semiring Semantics for Datalog

Datalog

$$Txy \;{:}{-}\; Exy$$
$$Txy \;{:}{-}\; Exz, Tzy$$



$\Rightarrow$

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: $T^*_{uw} = d \vee (a \wedge c)$
- Tropical: $T^*_{uw} = \min(12, 2 + 5) = 7$

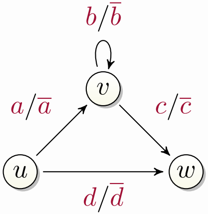
ω-continuous semirings

Semiring Semantics for Stratified Datalog

Stratified Datalog

$$Txy \;\text{:–}\; Exy$$
$$Txy \;\text{:–}\; Exz, Tzy$$

$$Nxy \;\text{:–}\; \neg Txy$$



Negation: can be defined in some semirings but not clear how do to it in general

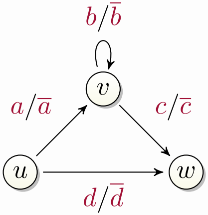
- PosBool: $T^*_{uw} = d \vee (a \wedge c)$
  $N^*_{uw} = \overline{T^*_{uw}} = \overline{d} \wedge (\overline{a} \vee \overline{c})$
- Polynomials: $\overline{a^2} = ?$
- Tropical: $\overline{7} = ?$

Semiring Semantics for Stratified Datalog

Stratified Datalog

$$Txy \;{:}\!-\; Exy$$
$$Txy \;{:}\!-\; Exz, Tzy$$

$$Nxy \;{:}\!-\; \neg Txy$$



Equation System

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

$\implies$ **Least solution**

$\rightsquigarrow$

Dualized System

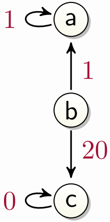
$$N_{uv} = \overline{a} \cdot (\overline{a} + N_{vv}) \cdot (\overline{d} + N_{wv})$$
$$N_{uw} = \overline{d} \cdot (\overline{d} + N_{ww}) \cdot (\overline{a} + N_{vw})$$

$\implies$ **Greatest solution**

Motivation II: Fixed-point Logic

$$\big[\mathbf{gfp}\, Rx.\ \exists y(Exy \wedge Ry)\big](v)$$

"there is an infinite path from v"



$$R_a = 1 + R_a$$

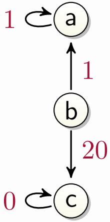
$$R_b = \min(1 + R_a,\ 20 + R_c)$$

$$R_c = 0 + R_c$$

Motivation II: Fixed-point Logic

$$\left[\textbf{gfp}\, Rx.\ \exists y(Exy \land Ry)\right](v)$$

"~~there is~~ cost of an infinite path from v"



$R_a = 1 + R_a$

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

$R_c = 0 + R_c$

${R_a}^* = \infty$

${R_b}^* = 20$

${R_c}^* = 0$

Greatest Solution

Computing Greatest Fixed Points

Naive Approach

$$
\begin{aligned}
R_a &= 1 + R_a \\
R_b &= \min(1 + R_a,\ 20 + R_c) \\
R_c &= 0 + R_c
\end{aligned}
$$

Naive Approach

Goal: Compute greatest fixed point of a polynomial operator

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

Naive Approach

Goal: Compute greatest fixed point of a polynomial operator

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

Iteration:

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix} \mapsto \cdots \mapsto \begin{pmatrix} 20 \\ 20 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 21 \\ 20 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} 22 \\ 20 \\ 0 \end{pmatrix} \mapsto \cdots \qquad \begin{pmatrix} \infty \\ 20 \\ 0 \end{pmatrix}$$

😕 😴

Faster Computation

Main Result

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

$$\mathsf{lfp}(\mathbf{F}) = \mathbf{F}^n(\mathbf{0}), \qquad \mathsf{gfp}(\mathbf{F}) = \mathbf{F}^n\big(\mathbf{F}^n(\mathbf{1})^\infty\big).$$

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

$$\underbrace{\begin{pmatrix}0\\0\\0\end{pmatrix} \mapsto \begin{pmatrix}1\\1\\0\end{pmatrix} \mapsto \begin{pmatrix}2\\2\\0\end{pmatrix} \mapsto \begin{pmatrix}3\\3\\0\end{pmatrix}}_{\leq n} \overset{\infty}{\longmapsto} \underbrace{\begin{pmatrix}\infty\\\infty\\0\end{pmatrix} \mapsto \begin{pmatrix}\infty\\20\\0\end{pmatrix} \circlearrowleft}_{\leq n}$$

Which Semirings?

recall: $(K, +, \cdot, 0, 1)$

❶ **Fully continuous**

- Natural order: $a \leq a + b$
- Each chain has supremum $\bigsqcup C$ and infimum $\bigsqcap C$, these commute with $+/\cdot$

❷ **Absorption**

- $a + a \cdot b = a \quad \Leftrightarrow \quad 1$ is greatest element $\quad \Leftrightarrow \quad a \cdot b \leq a$

Which Semirings?

recall: $(K, +, \cdot, 0, 1)$

1 Fully continuous

- Natural order: $a \leq a + b$
- Each chain has supremum $\bigsqcup C$ and infimum $\bigsqcap C$, these commute with $+/\cdot$

2 Absorption

- $a + a \cdot b = a$ $\Leftrightarrow$ 1 is greatest element $\Leftrightarrow$ $\boldsymbol{a \cdot b \leq a}$

Infinitary Power

For $a \in K$ we define: $a^{\infty} := \bigsqcap_{n<\omega} a^n$

Remember:
Decreasing multiplication

Proof Overview

Main Result

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

$$\mathsf{lfp}(\mathbf{F}) = \mathbf{F}^n(\mathbf{0}), \qquad \mathsf{gfp}(\mathbf{F}) = \mathbf{F}^n\big(\mathbf{F}^n(\mathbf{1})^\infty\big).$$

Proof sketch:

derivation trees

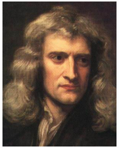
+

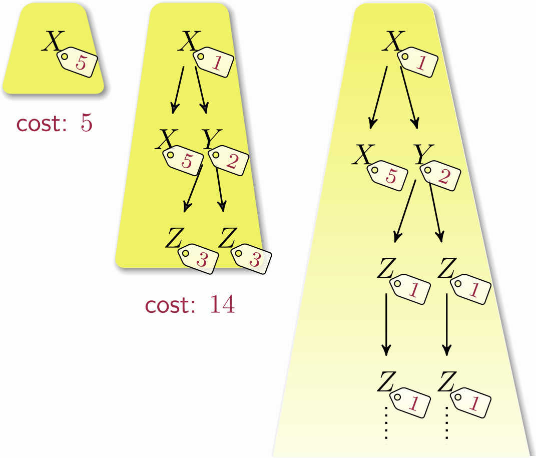
absorption

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)

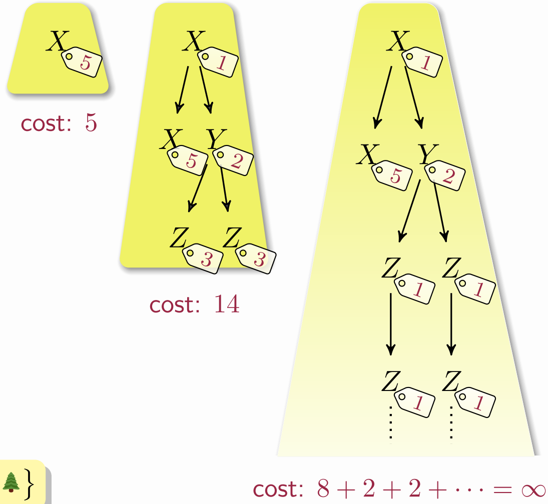




cost: $8+2+2+\cdots=\infty$

Derivation Trees

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

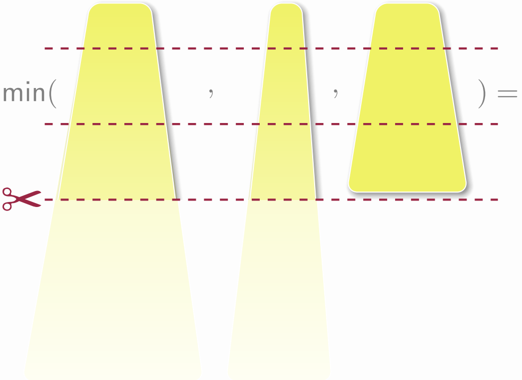


inspired by Newton's method
(Esparza, Kiefer, Luttenberger, JACM'10)

lfp = min{cost(🌲) | finite 🌲}

gfp = min{cost(🌲) | finite 🌲, infinite 🌲}

Observation: Prefixes of 🌲 correspond to iteration steps.



$\mathbf{F}(\mathbf{1})$

$\mathbf{F}^2(\mathbf{1})$

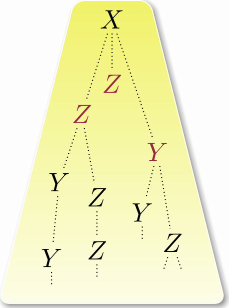
$\mathbf{F}^3(\mathbf{1})$

$\vdots$

$\bigsqcap_{n<\omega}$: $\quad \min\left\{ \mathrm{cost}(🌲) \mid \text{finite/infinite } 🌲 \right\} \quad = \quad \mathrm{gfp}(\mathbf{F})$ ■

Absorption on Derivation Trees

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



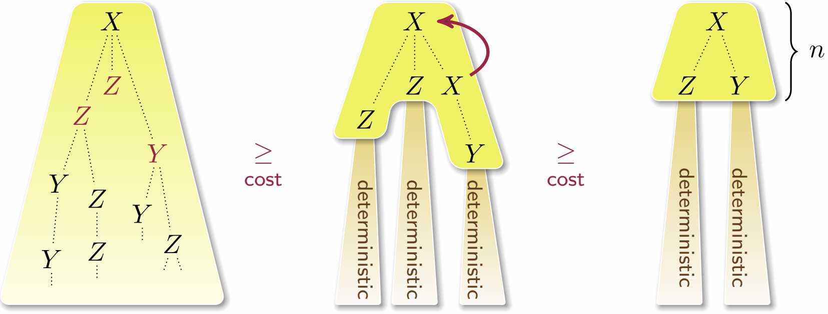
complicated tree

?

nice tree

Absorption on Derivation Trees

If each coefficient 🏷 2 occurs more often in 🌳 than in 🌲,
then cost(🌳) is absorbed by cost(🌲).

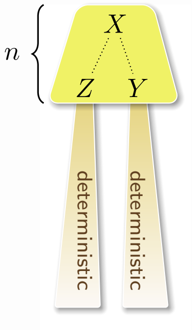


complicated tree 🌳 ultimately periodic nice tree 🌲

Computing Nice Trees

Main Result

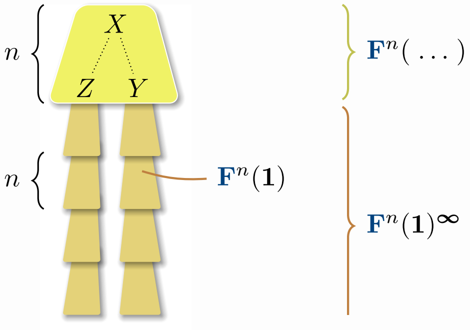
$$\mathrm{gfp}(\mathbf{F}) \;=\; \min\Big\{\mathrm{cost}(🌲)\;\Big|\;\text{nice } 🌲\Big\} \;=\; \ldots$$



Computing Nice Trees

Main Result

$$\mathrm{gfp}(\mathbf{F}) \;=\; \min\left\{\mathrm{cost}(\text{🌲}) \mid \text{nice } \text{🌲}\right\} \;=\; \mathbf{F}^n(\mathbf{F}^n(\mathbf{1})^\infty)$$



∎

Back to Datalog: Circuits

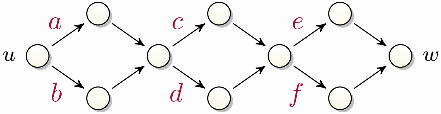
Circuits for Datalog Provenance

Problem: Provenance in polynomial semirings can become large

Datalog

Txy :– Exy

Txy :– Exz, Tzy



PosBool: $T^*_{uw} = ace + acf + ade + adf + bce + bcf + bde + bdf$

Solution: Represent provenance computation by a small circuit

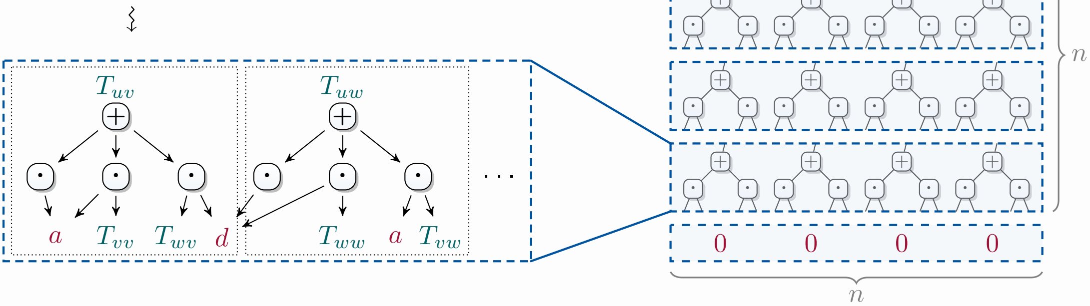
Circuits for Datalog Provenance

Recall

$\mathsf{lfp}(\mathbf{F}) = \mathbf{F}^n(0)$

Equation System

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

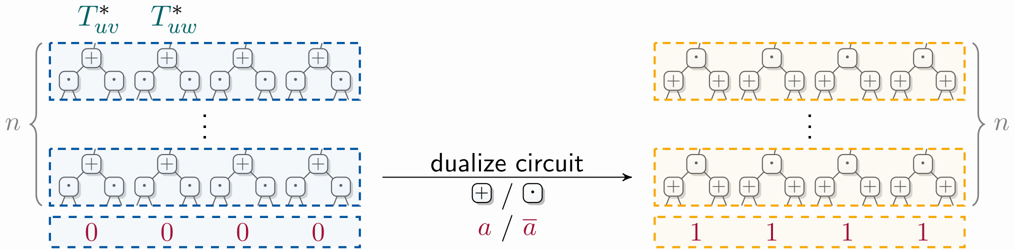


Circuits for Stratified Datalog

Strat. Datalog

$$Txy \;{:}\!-\; Exy$$
$$Txy \;{:}\!-\; Exz, Tzy$$

$$Nxy \;{:}\!-\; \neg Txy$$



Circuits for Stratified Datalog

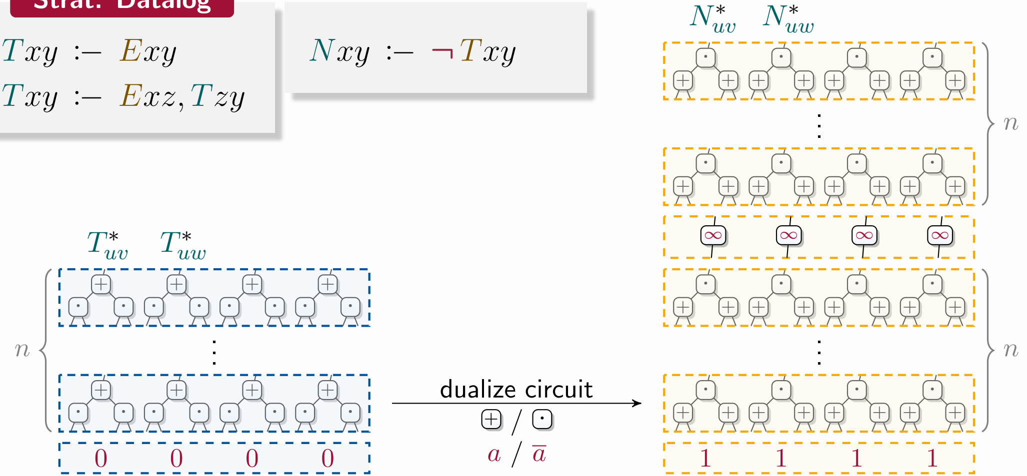
$$\mathsf{gfp}(\mathbf{F}) = \mathbf{F}^n\big(\,\mathbf{F}^n(1)^\infty\,\big)$$

Strat. Datalog

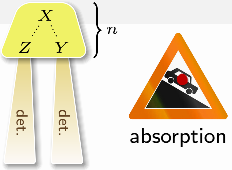
$Txy \;{:}{-}\; Exy$

$Txy \;{:}{-}\; Exz, Tzy$

$Nxy \;{:}{-}\; \neg Txy$



Summary

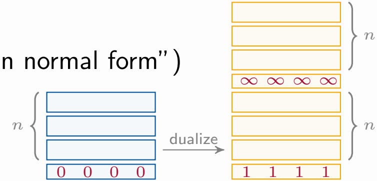


Computing greatest fixed points

- In absorptive semirings: $\mathrm{gfp}(\mathbf{F}) = \mathbf{F}^n\big(\mathbf{F}^n(1)^\infty\big)$

Semiring provenance for stratified Datalog

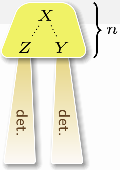
- Negation: greatest solution to dual equation system ("negation normal form")
- Circuit representations for Datalog can be generalized



Summary

Computing greatest fixed points

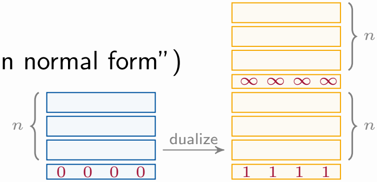
- In absorptive semirings: $\mathrm{gfp}(\mathbf{F}) = \mathbf{F}^n\big(\,\mathbf{F}^n(1)^\infty\,\big)$



absorption

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 ?