The Semiring Framework for Database Provenance

(: hindsight is great! :)
## Collaborators

<table>
<thead>
<tr>
<th>Category</th>
<th>Name(s)</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T of T award</strong></td>
<td>TJ Green</td>
<td>LogicBlox</td>
</tr>
<tr>
<td></td>
<td>Grigoris Karvounarakis</td>
<td>LogicBlox</td>
</tr>
<tr>
<td><strong>G of PODS paper</strong></td>
<td>TJ</td>
<td></td>
</tr>
<tr>
<td><strong>ORCHESTRA</strong></td>
<td>Zack Ives</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td></td>
<td>TJ, Grigoris</td>
<td></td>
</tr>
<tr>
<td><strong>Other core papers</strong></td>
<td>Nate Foster</td>
<td>Cornell University</td>
</tr>
<tr>
<td></td>
<td>Yael Amsterdamer</td>
<td>Bar-Ilan University</td>
</tr>
<tr>
<td></td>
<td>Daniel Deutch</td>
<td>Tel Aviv University</td>
</tr>
<tr>
<td></td>
<td>Tova Milo</td>
<td>Tel Aviv University</td>
</tr>
<tr>
<td></td>
<td>Sudeepa Roy</td>
<td>Duke University</td>
</tr>
<tr>
<td></td>
<td>Yuval Moskovitch</td>
<td>Tel Aviv University</td>
</tr>
<tr>
<td><strong>Recent work</strong></td>
<td>Erich Grädel</td>
<td>RWTH Aachen</td>
</tr>
<tr>
<td><strong>Much gratitude</strong></td>
<td>Peter Buneman</td>
<td>University of Edinburgh</td>
</tr>
</tbody>
</table>
Binary trust

Sue’s notes *

<table>
<thead>
<tr>
<th>cat</th>
<th>mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>rat</td>
</tr>
</tbody>
</table>

Val’s notes *

<table>
<thead>
<tr>
<th>mouse</th>
<th>gray</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>red</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
</tr>
</tbody>
</table>

Zack **

<table>
<thead>
<tr>
<th>cat</th>
<th>gray</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>red</td>
</tr>
</tbody>
</table>

** Sue and Val are noted zoologists. ** Zack is a noted computational zoologist

* * *
**Binary trust**

Sue’s notes *

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>mouse</td>
<td>Yes</td>
</tr>
<tr>
<td>cat</td>
<td>rat</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Val’s notes *

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>gray</td>
<td>No</td>
</tr>
<tr>
<td>mouse</td>
<td>red</td>
<td>No</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Zack **

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>gray</td>
<td>Yes</td>
</tr>
<tr>
<td>cat</td>
<td>red</td>
<td>No</td>
</tr>
</tbody>
</table>

* Sue and Val are noted zoologists.

** Zack is a noted *computational* zoologist
Access control

Sue’s notes

<table>
<thead>
<tr>
<th>cat</th>
<th>mouse</th>
<th>Pub</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>rat</td>
<td>Pub</td>
</tr>
</tbody>
</table>

Val’s notes

<table>
<thead>
<tr>
<th>mouse</th>
<th>gray</th>
<th>TSec</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>red</td>
<td>TSec</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>Conf</td>
</tr>
</tbody>
</table>

computation

Zack

<table>
<thead>
<tr>
<th>cat</th>
<th>gray</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>red</td>
<td>TSec</td>
</tr>
</tbody>
</table>

Pub < Conf < Sec < TSec
Confidence scores (non-binary trust)

<table>
<thead>
<tr>
<th></th>
<th>Sue's notes</th>
<th>Val's notes</th>
<th>Zack</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cat</td>
<td>mouse</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>cat</td>
<td>rat</td>
<td>0.9</td>
</tr>
<tr>
<td>mouse</td>
<td>gray</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>mouse</td>
<td>red</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

Computation:

\[
0.72 = \max(0.9 \times 0.8, 0.9 \times 0.6)
\]

\[
0.09 = 0.9 \times 0.1
\]
A simple model for data pricing

**Sue’s notes**

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>mouse</th>
<th>$10</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>cat</td>
<td>rat</td>
<td>$10</td>
</tr>
</tbody>
</table>

**Val’s notes**

<table>
<thead>
<tr>
<th></th>
<th>mouse</th>
<th>gray</th>
<th>$6</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>gray</td>
<td>$6</td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>red</td>
<td>$1</td>
<td></td>
</tr>
</tbody>
</table>

**Zack**

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>gray</th>
<th>$16</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>cat</td>
<td>red</td>
<td>$11</td>
</tr>
</tbody>
</table>

Computation:

\[
16 = \min(10 + 8, 10 + 6) \\
11 = 10 + 1
\]
Do it once and use it repeatedly: provenance

Label (annotate) input items abstractly with provenance tokens.

Provenance tracking: propagate expressions (involving tokens) (to annotate intermediate data and, finally, outputs)

Track two distinct ways of using data items by computation primitives:
• jointly (this alone is basically like keeping a log)
• alternatively (doing both is essential; think trust)

Input-output compositional; Modular (in the primitives)

Later, we want to evaluate the provenance expressions to obtain binary trust, access control, confidence scores, data prices, etc.
Algebraic interpretation for RDB

Set $X$ of provenance tokens.
Space of annotations, provenance expressions $\text{Prov}(X)$

**Prov($X$)-relations:**

- every tuple is annotated with some element from $\text{Prov}(X)$.

Binary operations on $\text{Prov}(X)$:

- corresponds to joint use (join, cartesian product),
- + corresponds to alternative use (union and projection).

Special annotations:

- “Absent” tuples are annotated with 0.
- 1 is a “neutral” annotation (data we do not track).
**$K$-Relational algebra**

Algebraic laws of $(\text{Prov}(X), +, \cdot, 0,1)$? More generally, for annotations from a structure $(K, +, \cdot, 0,1)$?

**$K$-relations.** Generalize RA+ to (positive) $K$-relational algebra.

Desired optimization equivalences of $K$-relational algebra iff $(K, +, \cdot, 0,1)$ is a **commutative semiring**.

Generalizes SPJU or UCQ or non-rec. Datalog

- set semantics $(\mathbb{B}, \vee, \wedge, \bot, T)$
- bag semantics $(\mathbb{N}, +, \cdot, 0, 1)$
- c-table-semantics [IL84] $(\text{BoolExp}(X), \vee, \wedge, \bot, T)$
- event table semantics [FR97,Z97] $(\mathcal{P}(\Omega), \cup, \cap, \emptyset, \Omega)$
What is a commutative semiring?

An algebraic structure \((K, +, \cdot, 0, 1)\) where:

- \(K\) is the domain
- \(+\) is associative, commutative, with 0 identity
- \(\cdot\) is associative, with 1 identity
- \(\cdot\) distributes over \(+\)
- \(a \cdot 0 = 0 \cdot a = 0\)

- \(\cdot\) is also commutative

Unlike ring, no requirement for inverses to \(+\)
Provenance: abstract semiring annotation

Sue’s notes

<table>
<thead>
<tr>
<th>cat</th>
<th>mouse</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>rat</td>
<td>q</td>
</tr>
</tbody>
</table>

Val’s notes

<table>
<thead>
<tr>
<th>mouse</th>
<th>gray</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>red</td>
<td>s</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>t</td>
</tr>
</tbody>
</table>

Zack(x,z):- Sue(x,y), Val(y,z)

Provenance polynomials $(\mathbb{N}[X], +, \cdot, 0, 1)$ semiring

<table>
<thead>
<tr>
<th>cat</th>
<th>gray</th>
<th>p\cdot r + q\cdot t</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>red</td>
<td>p\cdot s</td>
</tr>
</tbody>
</table>

Keep $X=\{p, q, r, s, t\}$ abstract.
Diagnostic for wrong answers; Deletion propagation.
E.g., $r=s=0$
Provenance polynomials

$((\mathbb{N}[X], +, \cdot, 0, 1) \text{ is the commutative semiring } \textit{freely generated} \text{ by } X)$
(universality property involving homomorphisms)

Provenance polynomials are \textit{PTIME}-computable (data complexity).
(query complexity depends on language and representation)
ORCHESTRA provenance (graph representation) about \textit{30\%} overhead

Monomials correspond to \textit{logical derivations} (proof trees in non-rec. Datalog)

\textbf{Provenance reading of polynomials:}
output tuple has provenance $2r^2 + rs$
three derivations of the tuple
- two of them use $r$, twice,
- the third uses $r$ and $s$, once each
Specialize provenance for access control

Sue’s notes

<table>
<thead>
<tr>
<th>cat</th>
<th>mouse</th>
<th>Pub</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>rat</td>
<td>Pub</td>
</tr>
</tbody>
</table>

Val’s notes

<table>
<thead>
<tr>
<th>mouse</th>
<th>gray</th>
<th>TSec</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>red</td>
<td>TSec</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>Conf</td>
</tr>
</tbody>
</table>

Zack(x,z):- Sue(x,y), Val(y,z)

Zack

<table>
<thead>
<tr>
<th>cat</th>
<th>gray</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>red</td>
<td>TSec</td>
</tr>
</tbody>
</table>

(A, min, max, 0, Pub) where A = Pub < Conf < Sec < TSec < 0

f: X → A  f(p)=f(q)=Pub  f(r)=f(s)=TSec  f(t)= Conf

Specialize provenance for confidence scores

<table>
<thead>
<tr>
<th>Sue’s notes</th>
<th>0.9</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat mouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat rat</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Val’s notes:

<table>
<thead>
<tr>
<th>mouse</th>
<th>gray</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse</td>
<td>red</td>
<td>0.1</td>
</tr>
<tr>
<td>rat</td>
<td>gray</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Zack(x,z):-
Sue(x,y), Val(y,z)

<table>
<thead>
<tr>
<th>Zack</th>
<th>0.72</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat gray</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cat red</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \forall = ([0,1], \max, \cdot, 0, 1) \text{ the Viterbi semiring} \]

\[ f : X \rightarrow [0,1] \quad f(p)=f(q)=0.9 \quad f(r)=0.6 \quad f(s)=0.1 \quad f(t)=0.8 \]

\[ \text{eval}(f): \mathbb{N}[X] \rightarrow \forall \quad \text{eval}(f)(pr+qt)=0.72 \quad \text{eval}(f)(ps)=0.09 \]
Some application semirings

\((\mathbb{B}, \land, \lor, T, \bot)\) \quad binary trust

\((\mathbb{N}, +, \cdot, 0, 1)\) \quad multiplicity (number of derivations)

\((\mathbb{A}, \text{min}, \text{max}, 0, \text{Pub})\) \quad access control

\(\mathbb{V} = ([0,1], \text{max}, \cdot, 0, 1)\) \quad Viterbi semiring (MPE) \quad confidence scores

\(\mathbb{T} = ([0, \infty], \text{min}, +, \infty, 0)\) \quad tropical semiring (shortest paths) \quad data pricing

\(\mathbb{F} = ([0,1], \text{max}, \text{min}, 0, 1)\) \quad “fuzzy logic” semiring
Two kinds of semirings in this framework

Provenance semirings, e.g.,

\((\mathbb{N}[X], +, \cdot, 0, 1)\) provenance polynomials [GKT07]

\((\text{Why}(X), \cup, \uplus, \emptyset, \{\emptyset\})\) witness why-provenance [BKT01]

Application semirings, e.g.,

\((\mathcal{A}, \text{min, max, 0, Pub})\) access control [FGT08]

\(\mathcal{V} = ([0,1], \text{max, \cdot, 0, 1})\) Viterbi semiring (MPE) [GKIT07]

Provenance specialization relies on

- Provenance semirings are freely generated by provenance tokens
- Query commutation with semiring homomorphisms
Query commutation with homomorphisms

query in $\mathcal{QL}$ homomorphism $h : K_1 \rightarrow K_2$

$K_1$-Rel $\xrightarrow{h} K_2$-Rel

$K_1$-Rel $\xrightarrow{h} K_2$-Rel

$\mathcal{QL} = RA+, \text{ Datalog} [GKT07]$

and extensions $[FGT08, GP10, ADT11a, T13, DMT15, GUKFC16, T17]$
A Hierarchy of Provenance Semirings [G09, DMRT14]

Example: $2x^2y + xy + 5y^2 + xz$

$\mathbb{N}[X]$  
+ idemp.  
$\mathbb{B}[X]$  
$\cdot$ idemp.

$x^2y + xy + y^2 + xz$  
absorption (ab+a=a)  
exy + y^2 + xz

Sorp($X$)  
$x^2y + xy + y^2 + xz$  
absorption (ab+a=a)  
exy + x^2 + xz

B[$X$]  
+ idemp.

Trio($X$)  
$3xy + 5y + xz$

most informative

Why($X$)  
$xy + y + xz$

least informative

Why($X$)  
$xy + y + xz$

Which($X$)  
$xyz$

$y + xz$  
PosBool($X$)

surjective semiring homomorphism, identity on $X$
A Hierarchy of Provenance Semirings [G09, DMRT14]
A menagerie of provenance semirings

(Which($X$), $\cup$, $\cup^*$, $\emptyset$, $\emptyset^*$) sets of contributing tuples “Lineage” (1) [CWW00]

(Why($X$), $\cup$, $\cup$, $\emptyset$, $\{\emptyset\}$) sets of sets of ... Witness why-provenance [BKT01]

(PosBool($X$), $\wedge$, $\vee$, $T$, $\bot$) minimal sets of sets of... Minimal witness why-provenance [BKT01] also “Lineage” (2) used in probabilistic dbs [SORK11]

(Trio($X$), $+$, $\cdot$, 0, 1) bags of sets of ... “Lineage” (3) [BDHT08,G09]

($\mathbb{B}[X]$,$+$, $\cdot$, 0, 1) sets of bags of ... Boolean coeff. polynomials [G09]

(Sorp($X$),$+$, $\cdot$, 0, 1) minimal sets of bags of ... absorptive polynomials [DMRT14]

($\mathbb{N}[X]$, $+$, $\cdot$, 0, 1) bags of bags of... universal provenance polynomials [GKT07]
From RA+ to Datalog

Immediate consequence operator $F$ of a Datalog program. Incorporates the edb predicates, maps idb predicates to idb predicates.

It’s expressible in RA+. E.g., transitive closure $F(T) = E \cup \pi_{1,3}(E \bowtie T)$

Generalize to $F: (K\text{-Rel})^n \to (K\text{-Rel})^n$ (n=# of idb predicates)

Solve certain (systems) of least fixed point equations over $K$-relations. $T = F(T)$

Equivalently:
- introduce unknowns $Z$ for the annotations of idb tuples
- solve system of fixed point equations over $K$;
  right-hand sides are polynomials in $K[Z]$.

Additional structure on $K$ for these to have (unique) solutions?
\( \omega \)-continuous semirings

Semirings \( K \) such that the immediate consequence operator of any Datalog program has a least fixpoint on \( K \)-relations.

**Naturally ordered** when

\[
x \leq y \quad \text{iff} \quad \text{there exists } z \quad \text{s.t.} \quad x + z = y
\]

is an order relation (all semirings seen here are naturally ordered)

**\( \omega \)-complete** also \( x_0 \leq x_1 \leq \ldots \leq x_n \leq \ldots \) have l.u.b.’s (sup’s)

**\( \omega \)-continuous** moreover \( + \) and \( \cdot \) preserve those l.u.b.’s
Among our examples

Many of the semirings that interest us
\( \mathbb{B}, \mathbb{T}, \mathbb{V}, \mathbb{A}, \mathbb{F} \) are already \( \omega \)-continuous.

\((\mathbb{N}, +, \cdot, 0, 1)\) is not, but its “completion” \((\mathbb{N}^\infty = \mathbb{N} \cup \{\infty\}, +, \cdot, 0, 1)\) is.

For provenance, the completion of \( \mathbb{N}[X] \) is not \( \mathbb{N}^\infty[X] \).
Instead of (finite) polynomials we need (possibly infinite)
formal power series.
They form an \( \omega \)-continuous semiring \( \mathbb{N}^\infty[[X]] \).
Monomials still correspond to derivations trees.
(Even transitive closure has infinitely many derivation trees if \( E \) has loops.)

The completion of \( \mathbb{B}[X] \) is \( \mathbb{B}[[X]] \).
Absorptive polynomials

Most informative provenance semiring for Datalog: \((\mathbb{N}_\leq[[X]], +, \cdot, 0,1)\)
(Infinite power series have finite representations as systems of polynomial equations.)

Absorption \( a + a \cdot b = a \)

Absorptive polynomials \( \text{Sorp}(X) \):

- boolean coefficients but only minimal degree monomials
  \[ x^2y + xy + y^2 + xz \rightarrow xy + y^2 + xz \]

Absorptive power series same as absorptive polynomials!

Why? Order monomials by degree of each variable.
In this infinite poset all antichains are finite! (Dickson’s Lemma)

\( \text{Sorp}(X) \) is already \( \omega \)-continuous: provides provenance polynomials for Datalog.

So is \( \text{PosBool}(X) \), but \( \text{Sorp}(X) \) provenance also supports tropical and Viterbi semiring applications
Further aspects of the framework

Extension to tree data (Nested Relational Calculus, structural recursion on trees, unordered XQuery) [FGT08]

Study of CQ/UCQ on provenance-annotated relations [G09]

Extension to aggregates (poly-size overhead) [ADT11a]

Poly-size provenance for Datalog (circuits; PosBool(X), Sorp(X)...) [DMRT14]

Extension to data-dependent finite state processes [DMT15]

Connections to semiring monad [FGT08, T13]
  to semimodules [ADT11a]
  to tensor products [ADT11a, DMT15]
Negative information; non-monotone operations (difference)

Boolean expressions [IL84]. Limited.

Add a binary operation corresponding to difference
m-semirings (common gen. of set and bag difference) [GP10]
spm-semirings (OPTIONAL in SPARQL) [GUKFC16]

Encode difference by aggregation [ADT11a]

Different equational theories, different algebraic optimizations [ADT11b]

Still not clear how to track negative information.
useful: non-answers (why not?), insertion propagation.

Logical model checking ("provenance of ... truth?")
eg negation as duality (NNFs), logical games
ongoing work with Grädel and Ives [T16, T17]
Current targets

ANALYTICS COMPUTATIONS

“Fine-grained provenance for linear algebra operators”
Yan, T., Ives TaPP 16

DISTRIBUTED SYSTEMS/NETWORK PROVENANCE

“Time-aware provenance for distributed systems”,
Zhou, Ding, Haeberlen, Ives, Loo TaPP 11

“Diagnosing missing events in distributed systems with negative provenance”,
Wu, Zhao, Haeberlen, Zhou, Loo SIGCOMM 14

STATIC ANALYSIS OF SOFTWARE

“On abstraction refinement for program analyses in Datalog”
Zhang, Mangal, Grigore, Naik PLDI 14
Framework references (I) *

[GKT07]  "Provenance semirings"  Green, Karvounarakis, Tannen  PODS 07.

[GKIT07]  "Update exchange with mappings and provenance"  Green, Karvounarakis, Ives, Tannen  VLDB 07.

[FGT08]  "Annotated XML: queries and provenance"  Foster, Green, Tannen  PODS 08.

[G09]  "Containment of conjunctive queries on annotated relations"  Green  ICDT 09.


* See also companion paper in  PODS 2017 proceedings.
Framework references (II)

[ADT11a]
“Provenance for aggregate queries”, Amsterdamer, Deutch, Tannen  PODS 11.

[ADT11b]
“On the limitations of provenance for queries with difference”,
Amsterdamer, Deutch, Tannen  TaPP 11

[T13]
“Provenance propagation in complex queries”
Tannen  Buneman Festschrift 2013

[DMRT14]

[DMT15]
“Provenance-based analysis of data-centric processes”
Deutch, Moskovitch, Tannen  VLDB J. 2015
Framework references (III)

[GUKFC16]  
"Algebraic structures for capturing the provenance of SPARQL queries”  
Geerts, Unger, Karvounarakis, Fundulaki, Christophides  JACM 2016

[T16]  
"About the provenance of truth”  Tannen  Simons Inst. Website 16  
https://simons.berkeley.edu/talks/val-tannen-2016-12-09

[T17]  
“Provenance analysis for FOL model checking”  Tannen  SIGLOG News 2017
Other references

[IL84]
“Incomplete information in relational databases” Imieliński, Lipski JACM 1984

[FR97]
“A probabilistic relational algebra” Fuhr, Röllecke TOIS 1997

[Z97]
“Query evaluation in probabilistic relational databases” Zimányi DDS 1997

[CWW00]
“Tracing the lineage of view data in a warehousing environment” Cui, Widom, Wiener TODS 2000

[BKT01]
“Why and where: a characterization of data provenance” Buneman, Khanna, Tan ICDT 2001

[BDHTW08]
“Databases with uncertainty and lineage” Benjelloun, Das Sarma, Halevy, Theobald, Widom VLDB J. 2008

[SORK11]
“Probabilistic databases” Suciu, Olteanu, Ré, Koch SLDM 2011
Thank you!