Provenance and Uncertainty in Databases

Todd J. Green

Department of Computer and Info. Science
University of Pennsylvania
`tjgreen@cis.upenn.edu`

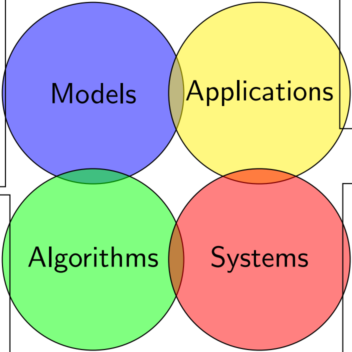
Uncertainty Management in Info. Systems @ Dagstuhl

October 14, 2008

Managing Uncertainty in Databases

kinds of uncertainty:
- incomplete
- probabilistic
- fuzzy
- . . .

representing uncertainty:
- *c*-tables, *v*-tables, etc.
- prob. event tables
- prob. XML, etc.



DB/IR integration
data exchange
dirty streams
data cleaning
. . .

reasoning about uncertainty:
- $\mathcal{RA}$ on *c*-tables, etc.
- the chase
- computing certain answers
- computing probabilities
- query optimization
- . . .

ProbView
MystiQ
Trio
MayBMS
Lahar
Clio
. . .

A major challenge: how to compare/relate/organize all of this!

Outline of Talk

I'll survey recent work at Penn that aims to unify some of these areas and exploit the connections among them:

- semiring-annotated relations (K-relations) which capture incomplete data, probabilistic data, fuzzy data, and others as special cases;
- semiring-based models of data provenance;
- query containment and equivalence results for various provenance models and for bag semantics (another instantiation of semiring framework).

Thanks to collaborators: Greg Karvounarakis, Nate Foster, Nick Taylor, Zack Ives, Val Tannen, and others in the Penn DB group.

Queries on c-tables [Imielinski,Lipski 84]

R

a	b	c	p
d	b	e	r
f	g	e	s

union of conjunctive queries (UCQ)

$$\begin{aligned} q(x,z) &\leftarrow R(x,u,z),\ R(v,w,z) \\ q(x,z) &\leftarrow R(x,y,u),\ R(v,y,z) \end{aligned}$$

Queries on c-tables [Imielinski,Lipski 84]

R

a	b	c	p
d	b	e	r
f	g	e	s

union of conjunctive queries (UCQ)

$$\begin{aligned} q(x,z) &\leftarrow R(x,u,z),\ R(v,w,z) \\ q(x,z) &\leftarrow R(x,y,u),\ R(v,y,z) \end{aligned}$$

$q(R)$

a	c	$(p \wedge p) \vee (p \wedge p)$
a	e	$p \wedge r$
d	c	$r \wedge p$
d	e	$(r \wedge r) \vee (r \wedge r) \vee (r \wedge s)$
f	e	$(s \wedge s) \vee (s \wedge s) \vee (s \wedge r)$

$=$

p
$p \wedge r$
$p \wedge r$
r
s

Queries on *c*-tables [Imielinski,Lipski 84]

R

a	b	c	p
d	b	e	r
f	g	e	s

union of conjunctive queries (UCQ)

$$q(x,z) \leftarrow R(x,u,z),\ R(v,w,z)$$
$$q(x,z) \leftarrow R(x,y,u),\ R(v,y,z)$$

$q(R)$

a	c	$(p \wedge p) \vee (p \wedge p)$	=	p
a	e	$p \wedge r$		$p \wedge r$
d	c	$r \wedge p$		$p \wedge r$
d	e	$(r \wedge r) \vee (r \wedge r) \vee (r \wedge s)$		r
f	e	$(s \wedge s) \vee (s \wedge s) \vee (s \wedge r)$		s

Valuation

$p \mapsto$ true

$r \mapsto$ false

$s \mapsto$ true

yields possible world

a	c
f	e

Set of possible worlds: $\text{rep}(R) := \{\text{Eval}_\nu(R) \mid \nu : \{p,r,s\} \to \mathbb{B}\}$

An Analogy With Bag Semantics

table with multiplicities

R

a	b	c	2
d	b	e	5
f	g	e	1

same query

$q(R)$

a	c	8
a	e	10
d	c	10
d	e	55
f	e	7

c-table calculations

a	c	$(p \wedge p) \vee (p \wedge p)$
a	e	$p \wedge r$
d	c	$r \wedge p$
d	e	$(r \wedge r) \vee (r \wedge r) \vee (r \wedge s)$
f	e	$(s \wedge s) \vee (s \wedge s) \vee (s \wedge r)$

multiplicity calculations

a	c	$2 \cdot 2 + 2 \cdot 2$
a	e	$2 \cdot 5$
d	c	$5 \cdot 2$
d	e	$5 \cdot 5 + 5 \cdot 5 + 5 \cdot 1$
f	e	$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 5$

The structure of the calculations is the same!

Abstracting the Structure of These Calculations

db ops	c-tables	bags	abstract
join	$\wedge$	$\cdot$	$\cdot$
union	$\vee$	$+$	$+$

Abstract calculations

a	c	$(p \cdot p) + (p \cdot p)$
a	e	$p \cdot r$
d	c	$r \cdot p$
d	e	$(r \cdot r) + (r \cdot r) + (r \cdot s)$
f	e	$(s \cdot s) + (s \cdot s) + (s \cdot r)$

These expressions capture the abstract structure of the calculations.

We will end up using these expressions as provenance!

Semiring-Annotated Relations [G.,Karvounarakis,Tannen 07]

Associate each tuple in database with an annotation from a commutative semiring $(K, +, \cdot, 0, 1)$

- $+$ and $\cdot$ are abstract operations
- 0 “not present”
- *n*-ary K-relation: a mapping $R : \mathbb{D}^n \to K$ from tuples to their annotations; 0 on all but finitely many tuples

Semiring-Annotated Relations [G.,Karvounarakis,Tannen 07]

Associate each tuple in database with an annotation from a commutative semiring $(K, +, \cdot, 0, 1)$

- $+$ and $\cdot$ are abstract operations
- 0 “not present”
- n-ary K-relation: a mapping $R : \mathbb{D}^n \rightarrow K$ from tuples to their annotations; 0 on all but finitely many tuples

Combine and propagate annotations during (positive) relational query processing

Semiring-Annotated Relations [G.,Karvounarakis,Tannen 07]

Associate each tuple in database with an annotation from a commutative semiring $(K, +, \cdot, 0, 1)$

- $+$ and $\cdot$ are abstract operations
- 0 “not present”
- n-ary K-relation: a mapping $R : \mathbb{D}^n \rightarrow K$ from tuples to their annotations; 0 on all but finitely many tuples

Combine and propagate annotations during (positive) relational query processing

Extensions to Datalog in [GKT 07] and generalization to XQuery/XML in [Foster,G.,Tannen 08] (won't discuss)

Positive Relational Algebra ($\mathcal{RA}^+$) on K-Relations

natural join $[R_1 \bowtie R_2](t) := R_1(t_1) \cdot R_2(t_2)$
where t on $\text{atts}(R_1) = t_1$, t on $\text{atts}(R_2) = t_2$

union $[R_1 \cup R_2](t) := R_1(t) + R_2(t)$

projection $[\pi_V R](t) := \sum_{t'=t \text{ on } V \text{ and } R(t')\neq 0} R(t')$

selection $[\sigma_P R](t) := R(t) \cdot P(t)$
where $P(t) = 0$ or 1

Semirings for Domain Applications

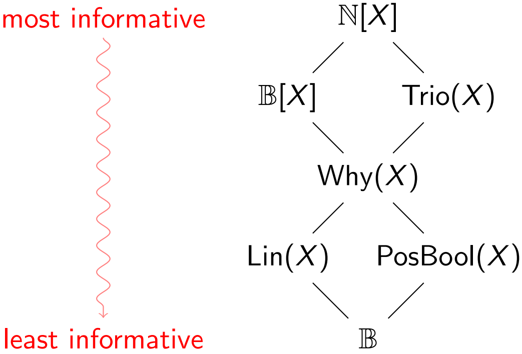
$(\mathbb{B}, \vee, \wedge, \bot, \top)$	Set semantics
$(\mathbb{N}, +, \cdot, 0, 1)$	Bag semantics
$(W, \mathsf{inf}, \mathsf{sup}, \{\bot\}, \{\top\})$ where $W = \{\emptyset, \{\bot\}, \{\top\}, \{\bot, \top\}\}$	Dunn-Belnap 4-valued logic
$([0,1], \max, \min, 0, 1)$	Fuzzy databases
$(\mathcal{C}, \min, \max, 0, \mathsf{P})$ where $\mathcal{C} = \mathsf{P} < \mathsf{C} < \mathsf{S} < \mathsf{T} < 0$	Access control policies
$(\mathbb{N}^{\infty}, \min, +, \infty, 0)$	Tropical semiring (costs)

Semirings for Provenance

X a set of variables, can be thought of as tuple ids

$(\mathrm{PosBool}(X), \vee, \wedge, \bot, \top)$	Positive Boolean *c*-tables
$(\mathrm{Lin}(X), \cup, \cup^\star, \emptyset, \emptyset^\star)$ sets of contributing tuples	Lineage [Cui,Widom 00]
$(\mathrm{Why}(X), \cup, \Cup, \emptyset, \{\emptyset\})$ sets of sets of contributing tuples	Proof Why-Provenance [Buneman+ 01]
$(\mathrm{Trio}(X), +, \cdot, 0, 1)$ bags of sets of contributing tuples	Trio (also called lineage in [Das Sarma+ 08])
$(\mathbb{N}[X], +, \cdot, 0, 1)$ "most informative" (universal)	Provenance polynomials
$(\mathbb{B}[X], +, \cdot, 0, 1)$	Boolean prov. polynomials

A Hierarchy of Provenance



most informative

least informative

$\mathbb{N}[X]$

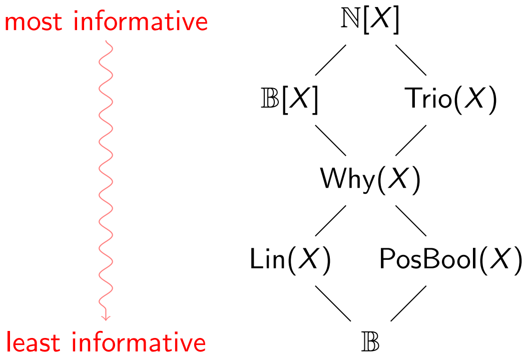
$\mathbb{B}[X]$ Trio(X)

Why(X)

Lin(X) PosBool(X)

$\mathbb{B}$

A Hierarchy of Provenance



most informative

$\mathbb{N}[X]$

$\mathbb{B}[X]$ Trio(X)

Why(X)

Lin(X) PosBool(X)

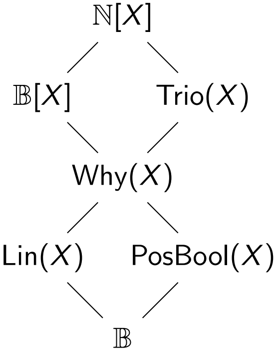
$\mathbb{B}$

least informative

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \rightarrow K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

$\mathbb{B}[X]$ Trio(X)

Why(X)

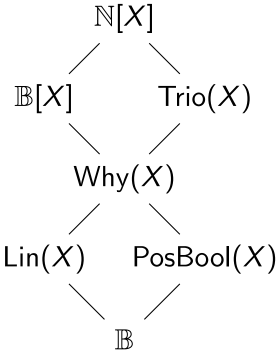
Lin(X) PosBool(X)

$\mathbb{B}$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \to K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

$\mathbb{B}[X]$ $\quad$ $\mathrm{Trio}(X)$

$\mathrm{Why}(X)$

$\mathrm{Lin}(X)$ $\quad$ $\mathrm{PosBool}(X)$

$\mathbb{B}$

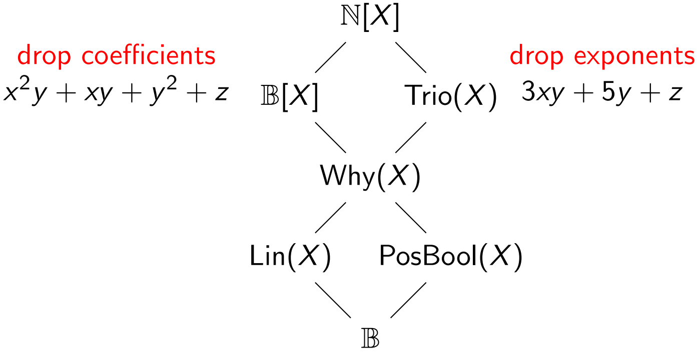
drop coefficients

$x^2y + xy + y^2 + z$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \rightarrow K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$

$\mathbb{B}[X]$

$\text{Trio}(X)$

drop exponents

$3xy + 5y + z$

$\text{Why}(X)$

$\text{Lin}(X)$

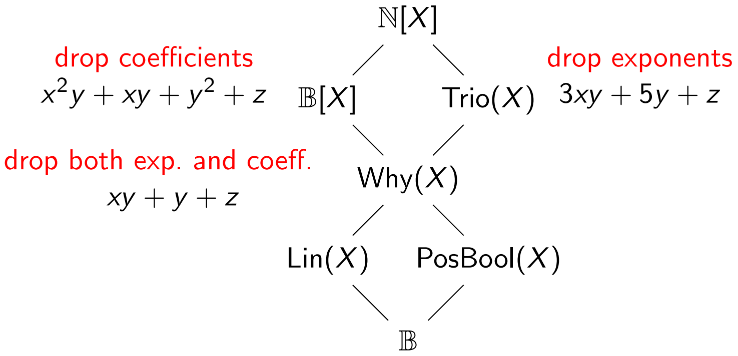
$\text{PosBool}(X)$

$\mathbb{B}$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \rightarrow K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$

$\mathbb{B}[X]$

$\mathsf{Trio}(X)$

drop exponents

$3xy + 5y + z$

drop both exp. and coeff.

$xy + y + z$

$\mathsf{Why}(X)$

$\mathsf{Lin}(X)$

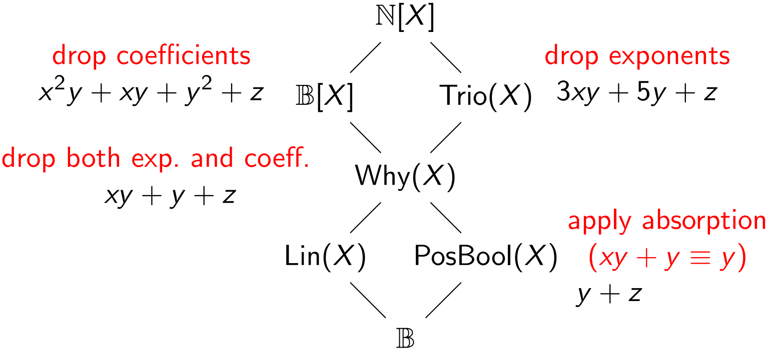
$\mathsf{PosBool}(X)$

$\mathbb{B}$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \to K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$

$\mathbb{B}[X]$

$\mathsf{Trio}(X)$

drop exponents

$3xy + 5y + z$

drop both exp. and coeff.

$xy + y + z$

$\mathsf{Why}(X)$

$\mathsf{Lin}(X)$

$\mathsf{PosBool}(X)$

apply absorption

$(xy + y \equiv y)$

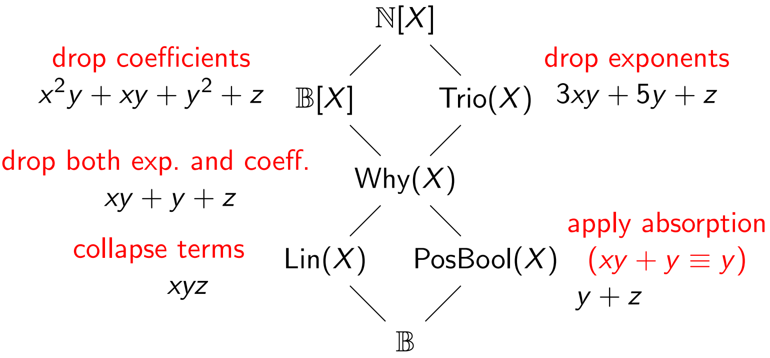
$y + z$

$\mathbb{B}$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \to K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$

$\mathbb{B}[X]$

$\mathsf{Trio}(X)$

drop exponents

$3xy + 5y + z$

drop both exp. and coeff.

$xy + y + z$

$\mathsf{Why}(X)$

collapse terms

xyz

$\mathsf{Lin}(X)$

$\mathsf{PosBool}(X)$

apply absorption

$(xy + y \equiv y)$

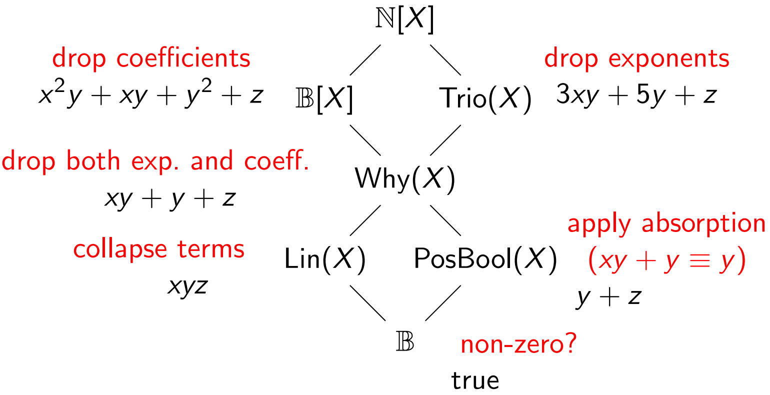
$y + z$

$\mathbb{B}$

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \to K_2$

A Hierarchy of Provenance

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



$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$

$\mathbb{B}[X]$

$\mathrm{Trio}(X)$

drop exponents

$3xy + 5y + z$

drop both exp. and coeff.

$xy + y + z$

$\mathrm{Why}(X)$

collapse terms

xyz

$\mathrm{Lin}(X)$

$\mathrm{PosBool}(X)$

apply absorption

$(xy + y \equiv y)$

$y + z$

$\mathbb{B}$

non-zero?

true

A path downward from K_1 to K_2 indicates that there exists a surjective semiring homomorphism $h : K_1 \rightarrow K_2$

Commutation with Homomorphisms

Recall the notion of strong representation system for incomplete dbs – for any PosBool(X)-relation R, query $q \in \mathcal{RA}^+$, following diagram should commute:

c-tables

$$\begin{tikzcd}
R \arrow[r, "\text{rep}"] \arrow[d, "q"] & \text{rep}(R) \arrow[d, "q"] \\
q(R) \arrow[r, "\text{rep}"] & \begin{array}{c} \text{rep}(q(R)) \\ = q(\text{rep}(R)) \end{array}
\end{tikzcd}$$

possible worlds

Commutation with Homomorphisms

Recall the notion of strong representation system for incomplete dbs – for any PosBool(X)-relation R, query $q \in \mathcal{RA}^+$, following diagram should commute:

c-tables

$$\begin{tikzcd}
R \arrow[r, "\text{rep}"] \arrow[d, "q"] & \text{rep}(R) \arrow[d, "q"] \\
q(R) \arrow[r, "\text{rep}"] & \begin{array}{c} \text{rep}(q(R)) \\ = q(\text{rep}(R)) \end{array}
\end{tikzcd}$$

possible worlds

Follows from a very general property of $\mathcal{RA}^+$ on K-relations!

Theorem (commutation with homomorphisms)

Let $h : K_1 \to K_2$ be a semiring homomorphism. Then for any K_1-relation R and query $q \in \mathcal{RA}^+$, we have $h(q(R)) = q(h(R))$.

Commutation with Homomorphisms

Recall the notion of strong representation system for incomplete dbs – for any PosBool(X)-relation R, query $q \in \mathcal{RA}^+$, following diagram should commute:

c-tables

$$\begin{tikzcd}
R \arrow[r, "\mathrm{rep}"] \arrow[d, "q"] & \mathrm{rep}(R) \arrow[d, "q"] \\
q(R) \arrow[r, "\mathrm{rep}"] & \begin{array}{c}\mathrm{rep}(q(R)) \\ = q(\mathrm{rep}(R))\end{array}
\end{tikzcd}$$

possible worlds

Follows from a very general property of $\mathcal{RA}^+$ on K-relations!

Theorem (commutation with homomorphisms)

Let $h : K_1 \to K_2$ be a semiring homomorphism. Then for any K_1-relation R and query $q \in \mathcal{RA}^+$, we have $h(q(R)) = q(h(R))$.

- Ex: $\forall$ valuations $\nu : X \to \mathbb{B}$ $\quad \mathrm{Eval}_\nu(q(R)) = q(\mathrm{Eval}_\nu(R))$

Provenance is Universal

Corollary (factoring)

The semantics of $\mathcal{RA}^+$ query answering on K-relations for any commutative semiring K factors through evaluation using provenance polynomials.

bag relation

R

a	b	c	2
d	b	e	5
f	g	e	1

$\xrightarrow{q}$

a	c	8
a	e	10
d	c	10
d	e	55
f	e	7

$q(R)$

↓ tag abstractly

↑ evaluate polynomials

$\mathbb{N}[X]$-relation

R'

a	b	c	p
d	b	e	r
f	g	e	s

$\xrightarrow{q}$

a	c	$2p^2$
a	e	pr
d	c	pr
d	e	$2r^2 + rs$
f	e	$2s^2 + rs$

$q(R')$

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations

$$\begin{tikzcd}
 & & \mathbb{N}[\mathsf{X}] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
 & \mathbb{B}[\mathsf{X}] \arrow[dl] \arrow[dr] & & \mathsf{Trio}(X) \arrow[dl] \arrow[dr, "{\{0,1\}}"] & \\
\mathbb{N}^{\infty} & & \mathsf{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
 & \mathsf{Lin}(X) & & \mathsf{PosBool}(X) \arrow[dl] \arrow[d] \arrow[dr] & \\
 & & \mathbb{B} & \mathcal{C} & {[0,1]}
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathsf{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^{\infty}$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations

$$\begin{tikzcd}
 & & \mathbb{N}[X] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
 & \mathbb{B}[X] \arrow[dl] \arrow[dr] & & \mathrm{Trio}(X) \arrow[dl] \arrow[dr, "\{0,1\}"] & \\
\mathbb{N}^\infty & & \mathrm{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
 & \mathrm{Lin}(X) & & \mathbf{PosBool}(X) \arrow[dl] \arrow[d] \arrow[dr] & \\
 & & \mathbb{B} & \mathcal{C} & [0,1]
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathrm{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^\infty$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations

$$\begin{tikzcd}
& & \mathbb{N}[X] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
& \mathbb{B}[X] \arrow[dl] \arrow[dr] & & \mathrm{Trio}(X) \arrow[dl] \arrow[dr, "\{0,1\}"] & \\
\mathbb{N}^{\infty} & & \mathbf{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
& \mathrm{Lin}(X) & & \mathrm{PosBool}(X) \arrow[dl] \arrow[dd] \arrow[dr] & \\
& & \mathbb{B} & & [0,1] \\
& & & \mathcal{C} &
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathrm{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^{\infty}$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations

$$\begin{tikzcd}
 & & \mathbb{N}[\mathrm{X}] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
 & \mathbb{B}[\mathrm{X}] \arrow[dl] \arrow[dr] & & \mathbf{Trio}(X) \arrow[dl] \arrow[dr, "\{0,1\}"] & \\
\mathbb{N}^\infty & & \mathrm{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
 & \mathrm{Lin}(X) & & \mathrm{PosBool}(X) \arrow[dl] \arrow[dd] \arrow[dr] & \\
 & & \mathbb{B} & & [0,1] \\
 & & & \mathcal{C} &
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

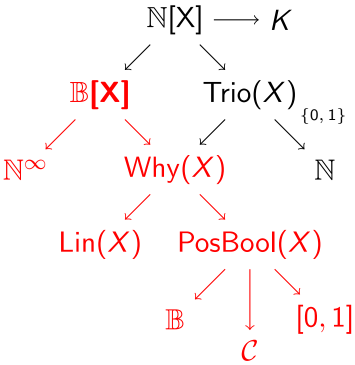
- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathrm{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^\infty$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations



$$\begin{tikzcd}
 & & \mathbb{N}[X] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
 & \mathbb{B}[\mathbf{X}] \arrow[dl] \arrow[dr] & & \mathrm{Trio}(X) \arrow[dl] \arrow[dr, "\{0,1\}"] & \\
\mathbb{N}^{\infty} & & \mathrm{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
 & \mathrm{Lin}(X) & & \mathrm{PosBool}(X) \arrow[dl] \arrow[d] \arrow[dr] & \\
 & & \mathbb{B} & \mathcal{C} & {[0,1]}
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathrm{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^{\infty}$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

Provenance and Uncertainty in Databases

The provenance models can be viewed as representation systems for incomplete K-relations

$$\begin{tikzcd}
 & & \mathbb{N}[\mathbf{X}] \arrow[r] \arrow[dl] \arrow[dr] & K & \\
 & \mathbb{B}[X] \arrow[dl] \arrow[dr] & & \mathsf{Trio}(X) \arrow[dl] \arrow[dr, "\{0,1\}"] & \\
\mathbb{N}^{\infty} & & \mathsf{Why}(X) \arrow[dl] \arrow[dr] & & \mathbb{N} \\
 & \mathsf{Lin}(X) & & \mathsf{PosBool}(X) \arrow[dl] \arrow[d] \arrow[dr] & \\
 & & \mathbb{B} & \mathcal{C} & {[0,1]}
\end{tikzcd}$$

Path from K_1 to K_2 means K_1 is a strong representation system for uncertain K_2-relations (for $\mathcal{RA}^+$)

- any $\nu : X \to K_2$ extends uniquely to sem. hom. $\mathsf{Eval}_\nu : K_1 \to K_2$

$\mathcal{C}$	access control semiring
$[0,1]$	fuzzy semiring
$\mathbb{N}^{\infty}$	tropical semiring (costs)

This is the connection between provenance and uncertainty!

From Incomplete to Probabilistic Databases

A *c*-table can be lifted to a probabilistic *c*-table by assigning probabilities to valuations $\nu : X \to \mathbb{B}$:

R

a	b	p
d	b	$p \wedge r$
f	g	$r \vee s$

Pr

p	0.6
r	0.3
s	1.0

$$\Pr[W] := \sum_{\nu \text{ s.t. } \mathrm{Eval}_\nu(R)=W} \Pr[\nu]$$

From Incomplete to Probabilistic Databases

A *c*-table can be lifted to a probabilistic *c*-table by assigning probabilities to valuations $\nu : X \to \mathbb{B}$:

R

a	b	p
d	b	$p \wedge r$
f	g	$r \vee s$

Pr

p	0.6
r	0.3
s	1.0

$$\Pr[W] := \sum_{\nu \text{ s.t. } \mathrm{Eval}_\nu(R)=W} \Pr[\nu]$$

Same can be done for incomplete K-relations – e.g., $\mathbb{N}[X]$ with probabilities on valuations $\nu' : X \to \mathbb{N}$ is a prob. bag relation:

R'

a	b	$2p$
d	b	pr
f	g	$r^2 + 3s$

$\Pr'$

$p \mapsto 3$	0.6
$p \mapsto 5$	0.4
$r \mapsto 1$	1.0
$s \mapsto 2$	1.0

$\Pr'[W]$:= as above

From Incomplete to Probabilistic Databases

A *c*-table can be lifted to a probabilistic *c*-table by assigning probabilities to valuations $\nu : X \to \mathbb{B}$:

R

a	b	p
d	b	$p \wedge r$
f	g	$r \vee s$

Pr

p	0.6
r	0.3
s	1.0

$$\Pr[W] := \sum_{\nu \text{ s.t. } \mathsf{Eval}_\nu(R)=W} \Pr[\nu]$$

Same can be done for incomplete *K*-relations – e.g., $\mathbb{N}[X]$ with probabilities on valuations $\nu' : X \to \mathbb{N}$ is a prob. bag relation:

R'

a	b	$2p$
d	b	pr
f	g	$r^2 + 3s$

Pr′

$p \mapsto 3$	0.6
$p \mapsto 5$	0.4
$r \mapsto 1$	1.0
$s \mapsto 2$	1.0

$\Pr'[W]$:= as above

Correctness: again, just use commutation with homomorphisms

Interlude

So that's K-relations. . .

. . . what about query containment/equivalence for K-relations?

Containment/Equivalence for K-Relations [G. 08]

Observation: the semiring annotations affect "the usual" (i.e., set semantics) query equivalences!

- e.g., the CQs

  $$q_1(x,y) \leftarrow R(x,y), R(x,z) \qquad q_2(u,v) \leftarrow R(u,v)$$

  are inequivalent under bag semantics, but equivalent under set semantics [Chaudhuri,Vardi 93]

Containment/Equivalence for K-Relations [G. 08]

Observation: the semiring annotations affect "the usual" (i.e., set semantics) query equivalences!

- e.g., the CQs

  $$q_1(x,y) \leftarrow R(x,y), R(x,z) \qquad q_2(u,v) \leftarrow R(u,v)$$

  are inequivalent under bag semantics, but equivalent under set semantics [Chaudhuri,Vardi 93]

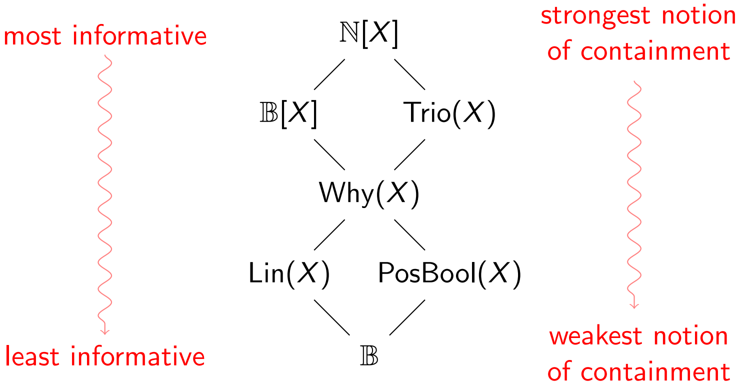
Theorem ([GKT 07])

If K is a distributive lattice, then for any $q_1, q_2 \in \mathcal{RA}^+$, we have $q_1 \sqsubseteq_K q_2 \iff q_1 \sqsubseteq_{\mathbb{B}} q_2$ (i.e., cont. holds under set semantics)

- PosBool(X), the fuzzy semiring, Dunn-Belnap 4-valued logic, and the access control semiring are all distributive lattices

Provenance Hierarchy and Query Containment

Recall the provenance hierarchy:



most informative

least informative

$\mathbb{N}[X]$

$\mathbb{B}[X]$ Trio(X)

Why(X)

Lin(X) PosBool(X)

$\mathbb{B}$

strongest notion of containment

weakest notion of containment

Can show that a path downward from K_1 to K_2 also indicates that for $\mathcal{RA}^+$queries, K_1-containment implies K_2-containment.

Logical Implications of Containment and Equivalence

$$\begin{tikzcd}
\mathbb{B}[X] \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \mathbb{N}[X] \\
\mathrm{Why}(X) \arrow[d, Rightarrow] & \\
\mathbb{N} \arrow[d, Rightarrow] & \\
\mathrm{Lin}(X) \arrow[d, Rightarrow] & \\
\mathrm{PosBool}(X) \arrow[r, Leftrightarrow] & \mathbb{B}
\end{tikzcd}$$

CQs, cont.

$$\begin{tikzcd}
\mathbb{B}[X] \arrow[r, Leftrightarrow] \arrow[d, Leftrightarrow] & \mathbb{N}[X] \arrow[d, Leftrightarrow] \\
\mathrm{Why}(X) \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \mathbb{N} \arrow[d, Leftrightarrow] \\
\mathrm{Lin}(X) \arrow[d, Rightarrow] & \mathrm{Trio}(X) \\
\mathrm{PosBool}(X) \arrow[r, Leftrightarrow] & \mathbb{B}
\end{tikzcd}$$

CQs, equiv.

$$\begin{tikzcd}
\mathbb{B}[X] \arrow[d, Rightarrow] & \mathbb{N}[X] \arrow[l, Rightarrow] \arrow[dd, Rightarrow] \\
\mathrm{Why}(X) \arrow[d, Rightarrow] & \\
\mathrm{Lin}(X) \arrow[d, Rightarrow] & \mathbb{N} \arrow[l, Rightarrow] \\
\mathrm{PosBool}(X) \arrow[r, Leftrightarrow] & \mathbb{B}
\end{tikzcd}$$

UCQs, cont.

$$\begin{tikzcd}
\mathbb{N} \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \mathbb{N}[X] \arrow[d, Leftrightarrow] \\
\mathbb{B}[X] \arrow[d, Rightarrow] & \mathrm{Trio}(X) \\
\mathrm{Why}(X) \arrow[d, Rightarrow] & \\
\mathrm{Lin}(X) \arrow[d, Rightarrow] & \\
\mathrm{PosBool}(X) \arrow[r, Leftrightarrow] & \mathbb{B}
\end{tikzcd}$$

UCQs, equiv.

$K_1 \Rightarrow K_2$ means K_1-containment (equivalence) implies K_2-containment (equivalence). All implications not marked $\Leftrightarrow$ are strict.

Diagrams all compatible with previous slide (but also include $\mathbb{N}$).

Complexity of Containment and Equivalence

		$\mathbb{B}$	PosBool(X)	Lin(X)	Why(X)	$\mathbb{B}[X]$	$\mathbb{N}[X]$	$\mathbb{N}$
CQs	cont	NP	NP	NP	NP	NP	NP	?(Π_2^p-hard)
	equiv	NP	NP	NP	GI	GI	GI	GI
UCQs	cont	NP	NP	NP	NP	NP	in PSPACE	undec
	equiv	NP	NP	NP	NP	NP	GI	GI

NP stands for NP-complete; GI is graph isomorphism-complete.
Red results are from [GKT 07]; blue results are from [G 08].

Decidability of bag-equivalence of UCQs was a longstanding open problem [Chaudhuri,Vardi 93]!

- Key: UCQs are bag-equivalent iff they are $\mathbb{N}[X]$-equivalent

Related Work

- The label systems of [Ioannidis,Ramakrishnan 95] – similar in spirit to semiring annotations
- The routes of [Chiticariu,Tan 06] – like minimal finite portions of our formal provenance polynomials (used for Datalog)
- Containment of CQs with where-provenance [Tan 03]
- Extending K-relations to support negation, and expressiveness of $\mathcal{RA}^+$ and $\mathcal{RA}$ on K-relations [Geerts,Poggi 08]
- Soft CSPs [Bistarelli+ 95] – coincides with project-join queries over (a restricted subclass of) K-relations

Conclusion

- K-relations: unifying framework for various kinds of uncertainty and provenance/lineage annotations
- Provenance polynomials – most "informative" kind of provenance annotations in this framework
- Positive decidability results for K-containment/equivalence of UCQs for various K used for provenance, also bag semantics

Didn't have time to discuss (but see overview in SIGMOD record):

- ORCHESTRA, a collaborative data sharing system incorporating semiring-based provenance

Ideas for Future Work

- Jointly recording provenance, security, multiplicities, uncertainty, etc. (product of semirings is also a semiring!)
- Query language extensions
    - e.g., generalized “duplicate elimination”

$$[\delta_K R](t) := \text{if } R(t) = 0 \text{ then } 0 \text{ else } 1$$

$\delta_{\mathbb{N}}$ is SQL-style duplicate elimination;
$\delta_{\mathsf{PosBool}(X)}$ is essentially the possible operator of MayBMS

- Containment/equivalence for K-UXML – wide open!
    - In [FGT 08] we show that K-equivalence is the same as $\mathbb{B}$-equivalence when K is a distributive lattice

Trio [Das Sarma+ 08]

Combines SQL-style duplicates and lineage annotations – Boolean combinations of tuple ids. (For $\mathcal{RA}^+$ queries, only use $\wedge$.)

table with duplicates

R

a	b	c	t_1
a	b	c	t_2
d	b	e	t_3
d	b	e	t_4
d	b	e	t_5
f	g	e	t_6

$\swarrow$ 2 indep. copies of (a b c)

8 indep. copies of (a c) $\nearrow$

same query

$q(R)$

a	c	t_1
a	c	$t_1 \wedge t_2$
a	c	$t_1 \wedge t_2$
	$\dots$	
a	c	t_2
a	e	$t_1 \wedge t_3$
a	e	$t_1 \wedge t_4$
	$\dots$	
f	e	$t_7 \wedge t_8$

A possible world is a bag relation obtained by choosing Boolean valuation of tuple ids, and keeping a given copy of tuple iff lineage annotation is satisfied.

Trio, Alternate View

The same example, with an algebraic notation ($\cdot$ replaces $\wedge$):

table with duplicates

R

a	b	c	$t_1 + t_2$
d	b	e	$t_3 + t_4 + t_5$
f	g	e	t_6

same query

$q(R)$

a	c	$2t_1 + 4t_1t_2 + 2t_2$
a	e	$t_1t_3 + t_1t_4 + t_1t_5$ $+t_2t_3 + t_2t_4 + t_2t_5$
	$\dots$	
f	e	$2t_6 + t_3t_6$ $+t_4t_6 + t_4t_6$

A possible world is obtained by applying $\{0,1\}$-valuation to variables, evaluating annotations to compute multiplicities

Trio, Alternate View

The same example, with an algebraic notation ($\cdot$ replaces $\wedge$):

table with duplicates

R

a	b	c	$t_1 + t_2$
d	b	e	$t_3 + t_4 + t_5$
f	g	e	t_6

same query

$q(R)$

a	c	$2t_1 + 4t_1t_2 + 2t_2$
a	e	$t_1t_3 + t_1t_4 + t_1t_5$ $+t_2t_3 + t_2t_4 + t_2t_5$
	…	
f	e	$2t_6 + t_3t_6$ $+t_4t_6 + t_4t_6$

A possible world is obtained by applying $\{0, 1\}$-valuation to variables, evaluating annotations to compute multiplicities

Annotations resemble provenance polynomials, but without exponents!

Modeling Trio with a Semiring

- Let $f : \mathbb{N}[X] \to \mathbb{N}[X]$ be the mapping which drops exponents, e.g.,

$$2x^2y + xy + 5y^2 \qquad \mapsto \qquad 3xy + 5y$$

Modeling Trio with a Semiring

- Let $f : \mathbb{N}[X] \rightarrow \mathbb{N}[X]$ be the mapping which drops exponents, e.g.,

$$2x^2y + xy + 5y^2 \qquad \mapsto \qquad 3xy + 5y$$

- Define equivalence relation

$$a \approx_f b \iff f(a) = f(b)$$

Can show $\approx_f$ is a congruence relation on $\mathbb{N}[X]$

Modeling Trio with a Semiring

- Let $f : \mathbb{N}[X] \to \mathbb{N}[X]$ be the mapping which drops exponents, e.g.,
$$2x^2y + xy + 5y^2 \qquad \mapsto \qquad 3xy + 5y$$
- Define equivalence relation
$$a \approx_f b \iff f(a) = f(b)$$
Can show $\approx_f$ is a congruence relation on $\mathbb{N}[X]$
- Trio-style lineage corresponds to the quotient semiring
$$\mathrm{Trio}(X) := \mathbb{N}[X]/\approx_f$$

Modeling Trio with a Semiring

- Let $f : \mathbb{N}[X] \to \mathbb{N}[X]$ be the mapping which drops exponents, e.g.,

$$2x^2y + xy + 5y^2 \qquad \mapsto \qquad 3xy + 5y$$

- Define equivalence relation

$$a \approx_f b \iff f(a) = f(b)$$

  Can show $\approx_f$ is a congruence relation on $\mathbb{N}[X]$

- Trio-style lineage corresponds to the quotient semiring

$$\mathrm{Trio}(X) := \mathbb{N}[X]/\approx_f$$

Theorem (correctness of $\mathcal{RA}^+$ on Trio(X)-relations)

For any $q \in \mathcal{RA}^+$, Trio(X)-relation R, and valuation $\nu : X \to \{0,1\}$, we have $Eval_\nu(q(R)) = q(Eval_\nu(R))$.

Proof. Observe that $\mathrm{Eval}_\nu : \mathrm{Trio}(X) \to \mathbb{N}$ is a semiring homomorphism, then use commutation with homomorphisms. ☐