Lecture 18

Data Provenance

Announcement

Project presentations:

- Tuesday, May 29, 8-1:30pm
- Presentation: 15'
- Presentation order on the Website
- Two Awards!
  - Best Project: Diploma + Amazon Gift Certificate
  - Best Presentation: Diploma + Amazon Gift Certificate
  - Voting instructions to be sent by email

Next lecture:

- Friday, 5/25, 10:30am, CSE403

Project Presentations Guidelines

What to include:

- A description of the problem: why is it important, why is it non-trivial
- An overview of prior approaches, and related work
- Your approach
- Your results (theoretical, empirical, experimental)
- A brief discussion on the significance of the results (do they work ? do they solve the problem you set out to do ? do they improve over existing work ?)
- Conclusions

Rule of thumb: 1 slide / minute, then subtract slack

You have 15’ → 12 slides.

Outline

Sources:

- Karnouvarakis et al., *Provenance Semirings*, PODS 2007
- Cheney, Chiticariu,Tan, *Provenance in Databases: Why, How, and Where*, 2007
- Tannen, Tutorial on Provenance in EDBT 2010

Data Provenance

Cheney, Chiticariu,Tan, *Provenance in Databases: Why, How, and Where*, 2007

- Provenance information describes the origins and the history of data in its life cycle. Such information (also called lineage) is important to many data management tasks.

Data Provenance

- Provenance inside the DBMS
  - Will discuss today
- Provenance outside of the DBMS
  - Much more messy; there is a standard, OPM (Open Provenance Model)

Provenance Annotations

- Some query produces an output table T(A,B,C)
- We store it over some period of time
- Later we ask: “where did this tuple come from?”
- The “provenance annotation” answers this.

A	B	C	
a1	b1	c1	provenance1
a2	b1	c1	provenance2
a2	b2	c2	provenance3
a2	b2	c3	provenance4

Provenance Annotations

- Start by annotating each tuple in the original database with a unique identifier; can be the Tuple Id (TID)

A	B	
a1	b1	X1
a2	b1	X2
a2	b2	X3

- Next, compute the provenance expression inductively, based on the query plan

Join Operator

$\bowtie$

A	B	
a1	b1	X1
a2	b1	X2
a2	b2	X3

B	C	
b1	c1	Y1
b2	c2	Y2
b2	c3	Y3

=

A	B	C	
a1	b1	c1	$X1 \cdot Y1$
a2	b1	c1	$X2 \cdot Y1$
a2	b2	c2	$X3 \cdot Y2$
a2	b2	c3	$X3 \cdot Y3$

Projection Operator

Π

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	
a1	X1+X2
a2	X3+X4+X5

Union Operator

$\cup$

A	B	
a1	b1	X1
a2	b2	X2

A	B	
a2	b2	Y1
a3	b3	Y2

=

A	B	
a1	b1	X1
a2	b2	X2+Y1
a3	b3	X3

Selection Operator

$\sigma_{A=a1}$

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	B	
a1	b1	X1
a1	b2	X2

We could simply remove the tuples filtered out.
But it's better to keep them around (we'll see why).
What is their annotation?

Selection Operator

$\sigma_{A=a1}$

A	B	
a1	b1	X1
a1	b2	X2
a2	b1	X3
a2	b2	X4
a2	b3	X5

=

A	B	
a1	b1	$X1 \cdot 1$
a1	b2	$X2 \cdot 1$
a2	b1	$X3 \cdot 0$
a2	b2	$X4 \cdot 0$
a2	b3	$X5 \cdot 0$

We could simply remove the tuples filtered out.
But it's better to keep them around (we'll see why).
What is their annotation?

Complex Example

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$(X \cdot X + X \cdot X) \cdot 0 = 2 \cdot X^2$
a	e	$X \cdot Y \cdot 1 = X \cdot Y$
d	c	$Y \cdot X \cdot 0 = 0$
d	e	$(Y \cdot Y + Y \cdot Z + Y \cdot Y) \cdot 1 = 2 \cdot Y^2 + Y \cdot Z$
f	e	$(Z \cdot Z + Z \cdot Y + Z \cdot Z) \cdot 1 = 2 \cdot Z^2 + Y \cdot Z$

Discuss in class what these annotations mean

K-Relations

Definition. A K-relation is a relation where each tuple is annotated with an element from the set K.

What we have described so far is an extension of the positive operations of the relational algebra to K-relations

We assumed that K has the operators +, $\cdot$

Identities on Provenance Expressions

The problem:

- We have defined the provenance expressions for query plans P
- Given a query Q, we want the provenance of its answers to be the same, no matter what plan we use: P1, P2, …
- What we need: if P1=P2, then
  the provenance expressions for P1 =
  the provenance expressions for P2

Identities on Provenance Expressions

Definition. A structure $(K, +, \cdot, 0, 1)$ is called a *commutative semiring* if:

1. $(K,+,0)$ is a commutative monoid:
   a. $+$ is associative: $(x+y)+z=x+(y+z)$
   b. $+$ is commutative: $x+y=y+x$
   c. 0 is the identity for $+$: $x+0=0+x=x$
2. $(K, \cdot, 1)$ is a commutative monoid:
   a. … (similar identities)
3. $\cdot$ distributes over $+$: $x \cdot (y+z) = x \cdot y + x \cdot z$
4. For all x, $x \cdot 0 = 0 \cdot x = 0$

Identities on Provenance Expressions

Definition. A structure $(K, +, \cdot, 0, 1)$ is called a *commutative semiring* if:

1. $(K,+,0)$ is a commutative monoid:
   a. $+$ is associative: $(x+y)+z=x+(y+z)$
   b. $+$ is commutative: $x+y=y+x$
   c. 0 is the identity for $+$: $x+0=0+x=x$
2. $(K, \cdot, 1)$ is a commutative monoid:
   a. … (similar identities)
3. $\cdot$ distributes over $+$: $x\cdot(y+z) = x\cdot y + x\cdot z$
4. For all x, $x\cdot 0 = 0\cdot x = 0$

Theorem. The standard identities of the Bag algebra hold for K-relations iff $(K, +, \cdot, 0, 1)$ is a commutative semiring.

Identities on Provenance Expressions

Discuss in class:

$$\boxed{q(x,y) := R(x,y),\ S(y,z),\ T(z,u)}$$

Given two plans, why are the annotations equal?

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

Q: Suppose we delete the tuple (d,b,e) from R. Which tuple(s) disappear from the result?

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

=

A	C	
a	c	$2 \cdot X^2$
a	e	0
d	e	0
f	e	$2 \cdot Z^2$

Q: Suppose we delete the tuple (d,b,e) from R. Which tuple(s) disappear from the result?

A: Set Y=0

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

Q: Suppose each tuple in R occurs 3 times (bag semantics). How many times occurs each tuple in the answer?

Applications

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

A	C	
a	c	18
a	e	9
d	e	27
f	e	27

Q: Suppose each tuple in R occurs 3 times (bag semantics). How many times occurs each tuple in the answer?

A. Set X=Y=Z=3

Lineage

Lineage = set of contributing tuples

- Terminology alert: provenance and lineages are not used consistently in the literature
- The PODS'2007 paper calls this *why-provenance* , Fig. 5; I will call it *lineage*

Lineage

$$\sigma_{C=e}\Pi_{AC}(\ \Pi_{AC}(R) \bowtie \Pi_{BC}(R) \cup \Pi_{AB}(R) \bowtie \Pi_{BC}(R)) =$$

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

A	C	
a	c	$2 \cdot X^2$
a	e	$X \cdot Y$
d	e	$2 \cdot Y^2 + Y \cdot Z$
f	e	$2 \cdot Z^2 + Y \cdot Z$

➔

A	C	
a	c	X
a	e	X, Y
d	e	Y, Z
f	e	Y, Z

Lineage = traces only the set of input tuples that contributed to an output tuple

This is also a semi-ring! Which one?

Semirings for various models of provenance (1)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	Y,Z

Lineage [CuiWidomWiener 00 etc.]

Sets of contributing tuples

Semiring: $(\mathrm{Lin}(X), \cup, \cup^*, \varnothing, \varnothing^*)$

Source: Tannen, EDBT 2010

Semirings for various models of provenance (2)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	{{Y},{Y,Z}}

(Witness, Proof) **why-provenance**
[BunemanKhannaTan 01] & [Buneman+ PODS08]

Sets of witnesses (w. =set of contributing tuples)

Semiring: $(\mathrm{Why}(X), \cup, \Cup, \varnothing, \{\varnothing\})$

Source: Tannen, EDBT 2010

Semirings for various models of provenance (3)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	{Y}

Minimal witness **why-provenance**
[BunemanKhannaTan 01]

Sets of minimal witnesses

Semiring: $(\mathrm{PosBool}(X), \wedge, \vee, \top, \perp)$

Source: Tannen, EDBT 2010

Semirings for various models of provenance (4)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	[{Y}, {Y}, {Y,Z}]

Notation:

{ } set

[ ] bag

Trio lineage [Das Sarma+ 08]

Bags of sets of contributing tuples (of witnesses)

Semiring: (Trio(X), +, $\cdot$, 0, 1) (defined in [Green, ICDT 09])

Source: Tannen, EDBT 2010

Semirings for various models of provenance (5)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	{[Y,Y], [Y,Z]}

Notation:

{ } set

[ ] bag

Polynomials with boolean coefficients [Green, ICDT 09]

(B[X]-provenance)

Sets of bags of contributing tuples

Semiring: (B[X], +, $\cdot$, 0, 1)

Source: Tannen, EDBT 2010

Semirings for various models of provenance (6)

R =

A	B	C	
a	b	c	X
d	b	e	Y
f	g	e	Z

Q =

A	C	
d	e	[[Y,Y], [Y,Y], [Y,Z]]

Provenance polynomials [GKT, PODS 07]

(N[X]-provenance)

Bags of bags of contributing tuples

Semiring: (N[X], +, $\cdot$, 0, 1)

Source: Tannen, EDBT 2010

Application

Discretionary Access Control [LaPadula]

- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

R =

A	B	C
a	b	c
d	b	e
f	g	e

X=C

Y=P

Z=C

A	C	
a	c	$2 \cdot X^2 = ?$
a	e	$X \cdot Y = ?$
d	e	$2 \cdot Y^2 + Y \cdot Z = ?$
f	e	$2 \cdot Z^2 + Y \cdot Z = ?$

Application

Discretionary Access Control [LaPadula]

- Public = P
- Confidential = C
- Secret = S
- Top Secret = T
- No Such Thing… = 0

R =

A	B	C	
a	b	c	X=C
d	b	e	Y=P
f	g	e	Z=T

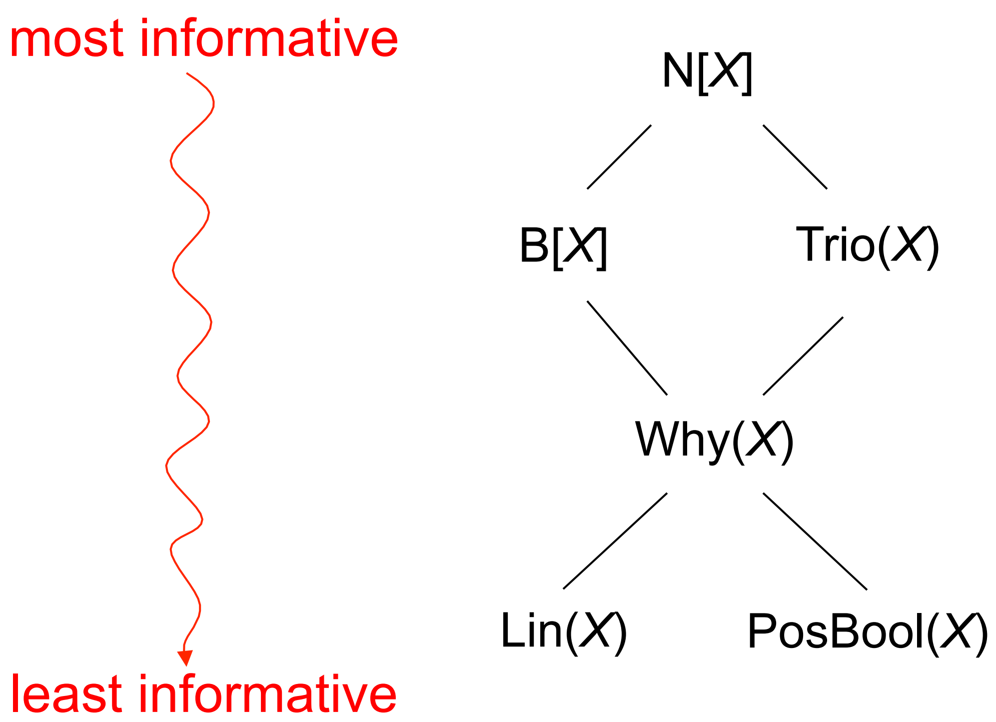
A	C	
a	c	$2 \cdot X^2 = C$
a	e	$X \cdot Y = C$
d	e	$2 \cdot Y^2 + Y \cdot Z = C$
f	e	$2 \cdot Z^2 + Y \cdot Z = T$

(A, min, max, 0, P), where A = P < C < S < T < 0

But are there useful commutative semirings?

$(B, \wedge, \vee, \top, \bot)$	Set semantics
$(\mathbb{N}, +, \cdot, 0, 1)$	Bag semantics
$(P(\Omega), \cup, \cap, \varnothing, \Omega)$	Probabilistic events [FuhrRölleke 97]
$(\text{BoolExp}(X), \wedge, \vee, \top, \bot)$	Conditional tables (c-tables) [ImielinskiLipski 84]
$(R_+^\infty, \min, +, 1, 0)$	Tropical semiring (cost/distrust score/confidence need)
$(A, \min, \max, 0, P)$ where $A = P < C < S < T < 0$	Access control levels [PODS8]

A provenance hierarchy



One semiring to rule them all… (apologies!)

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

$\mathbb{N}[X]$

drop coefficients

$x^2y + xy + y^2 + z$ $\mathbb{B}[X]$

drop exponents

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

drop both exp. and coeff.

$xy + y + z$

Why(X)

collapse terms

xyz

Lin(X)

PosBool(X)

apply absorption ($ab + b = b$)

$y + z$

A path downward from K_1 to K_2 indicates that there exists an **onto (surjective) semiring homomorphism** $h : K_1 \rightarrow K_2$

Using homomorphisms to relate models

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

N[X]

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

B[X]

Trio(X)

drop exponents
$3xy + 5y + z$

drop both exp. and coeff.
$xy + y + z$

Why(X)

collapse terms
xyz

Lin(X)

PosBool(X)

apply absorption
($ab + b = b$)
$y + z$

Homomorphism?

$h(x+y) = h(x)+h(y)$ $h(xy)=h(x)h(y)$ $h(0)=0$ $h(1)=1$

Moreover, for these homomorphisms $h(x)= x$