# Provenance Semirings

T.J. Green, G. Karvounarakis, V. Tannen University of Pennsylvania

Principles of Provenance (PrOPr)
Philadelphia, PA June 26, 2007

#### Provenance

- First studied in data warehousing
  - Lineage [Cui, Widom, Wiener 2000]
- Scientific applications (to assess quality of data)
  - Why-Provenance [Buneman, Khanna, Tan 2001]
- Our interest: P2P data sharing in the ORCHESTRA system (project headed by Zack Ives)
  - Trust conditions based on provenance
  - Deletion propagation

#### Annotated relations

- Provenance: an annotation on tuples
- Our observation: propagating provenance/lineage through views is similar to querying
  - Incomplete Databases (conditional tables)
  - Probabilistic Databases (independent tuple tables)
  - Bag Semantics Databases (tuples with multiplicities)
- Hence we look at queries on relations with annotated tuples



#### Incomplete databases: boolean C-tables

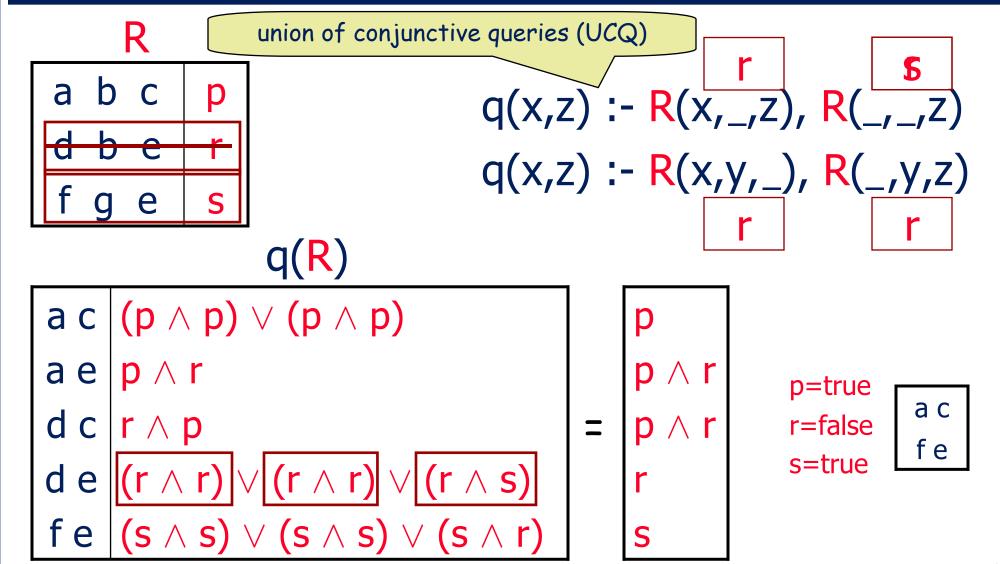
R
abc p
dbe r
fge s

boolean variables

semantics: a set of instances



# Imielinski & Lipski (1984): queries on C-tables

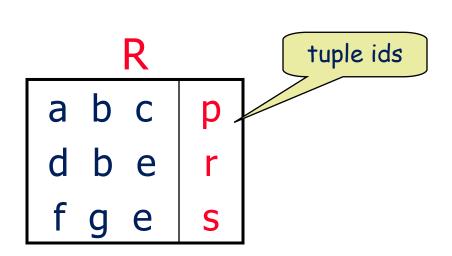


**PROPR 2007** 



### Why-provenance/lineage

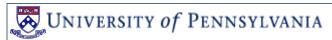
Which input tuples contribute to the presence of a tuple in the output?



q(R)

a c {p}
a e {p,r}
d c {p,r}
d e {r,s}
f e {r,s}

[Cui, Widom, Wiener 2000]
[Buneman, Khanna, Tan 2001]



### C-tables vs. Why-provenance

ac  $(p \land p) \lor (p \land p)$ ae  $p \land r$ dc  $r \land p$ de  $(r \land r) \lor (r \land r) \lor (r \land s)$ fe  $(s \land s) \lor (s \land s) \lor (s \land r)$ 

c-table calculations

Why-provenance calculations

```
a c ({p} ∪ {p}) ∪ ({p} ∪ {p})

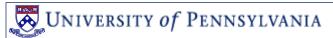
a e {p} ∪ {r}

d c {r} ∪ {p}

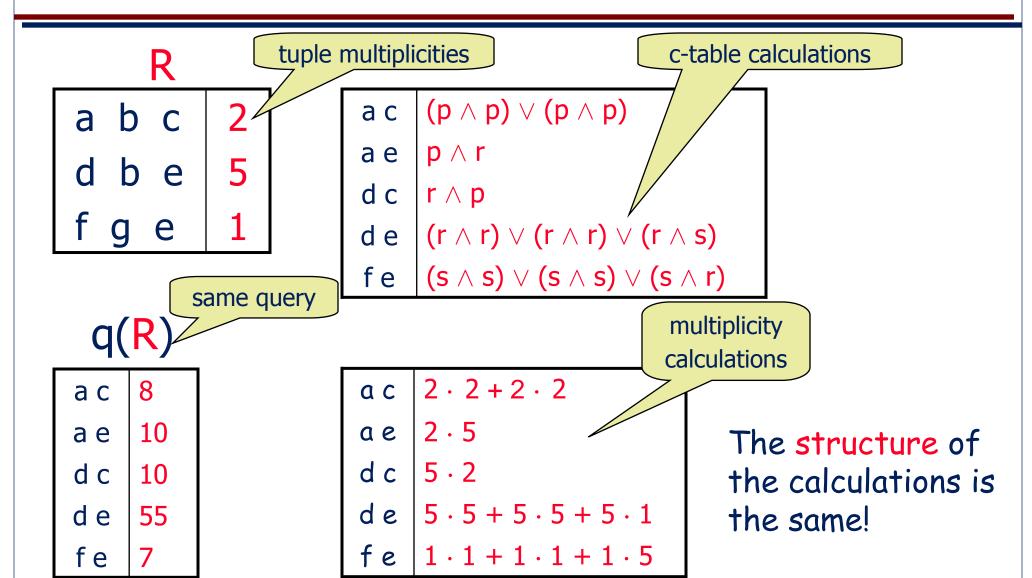
d e ({r} ∪ {r}) ∪ ({r} ∪ {r}) ∪ ({r} ∪ {s})

f e ({s} ∪ {s}) ∪ ({s} ∪ {s}) ∪ ({s} ∪ {r})
```

The structure of the calculations is the same!

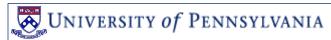


# Another analogy, with bag semantics



**PROPR 2007** 

-8



# Abstracting the structure of these calculations

	C-tables	Bags	Why-provenance	Abstract
join	<u> </u>	•	U	•
union	V	+	U	+

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, which encodes the logical derivation of the output tuples

We shall use these expressions as provenance

PROPR 2007



### Positive K-relational algebra

- We define an RA+ on K-relations:
  - The · corresponds to join:
  - The + corresponds to union and projection
  - 0 and 1 are used for selection predicates
  - Details in the paper (but recall how we evaluated the UCQ q earlier and we will see another example later)

PROPR 2007



### RA+ identities imply semiring structure!

- Common RA+ identities
  - Union and join are associative, commutative
  - Join distributes over union
  - etc. (but not idempotence!)

These identities hold for RA+ on K-relations iff

 $(K, +, \cdot, 0, 1)$  is a commutative semiring

(K,+,0) is a commutative monoid
(K, ⋅,1) is a commutative monoid
· distributes over +, etc



#### Calculations on annotated tables are particular cases

 $(\mathbb{B}, \vee, \wedge, \text{ false, true})$ 

usual relational algebra

 $(\mathbb{N}, +, \cdot, 0, 1)$ 

bag semantics

(PosBool(B),  $\vee$ ,  $\wedge$ , false, true) boolean C-tables

 $(\mathcal{P}(\Omega), \cup, \cap, \emptyset, \Omega)$ 

probabilistic event tables

 $(\mathcal{P}(X), \cup, \cup, \emptyset, \emptyset)$ 

lineage/why-provenance

#### Provenance Semirings

- X = {p, r, s, ...}: indeterminates (provenance "tokens" for base tuples)
- ullet N[X]: multivariate polynomials with coefficients in N and indeterminates in X
- (N[X], +, ·, 0, 1) is the most "general" commutative semiring: its elements abstract calculations in all semirings
- ullet  $\mathbb{N}[X]$  -relations are the relations with provenance!
  - The polynomials capture the propagation of provenance through (positive) relational algebra

#### A provenance calculation

$$q(x,z) := R(x, _,z), R(_, _,z)$$

$$q(x,z) := R(x,y, _), R(_,y,z)$$

R

q(R)

Why-provenance

abc p
dbe r
fges

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

a c  $\{p\}$ a e  $\{p,r\}$ d c  $\{p,r\}$ d e  $\{r,s\}$ same why-provenance,

Not just why- but also how-provenance (encodes derivations)!

different polynomials

• More informative than why-provenance



#### Further work

- Application: P2P data sharing in the ORCHESTRA system:
  - Need to express trust conditions based on provenance of tuples
  - Incremental propagation of deletions
  - Semiring provenance itself is incrementally maintainable
- Future extensions:
  - full relational algebra: For difference we need semirings with "proper subtraction"
  - richer data models: nested relations/complex values, XML