Preference-Driven Querying of Inconsistent Relational Databases

Slawomir Staworko[1] Jan Chomicki[1] Jerzy Marcinkowski[2]

[1]Department of Computer Science and Engineering
University at Buffalo, SUNY

[2]Institute of Informatics
Wroclaw University

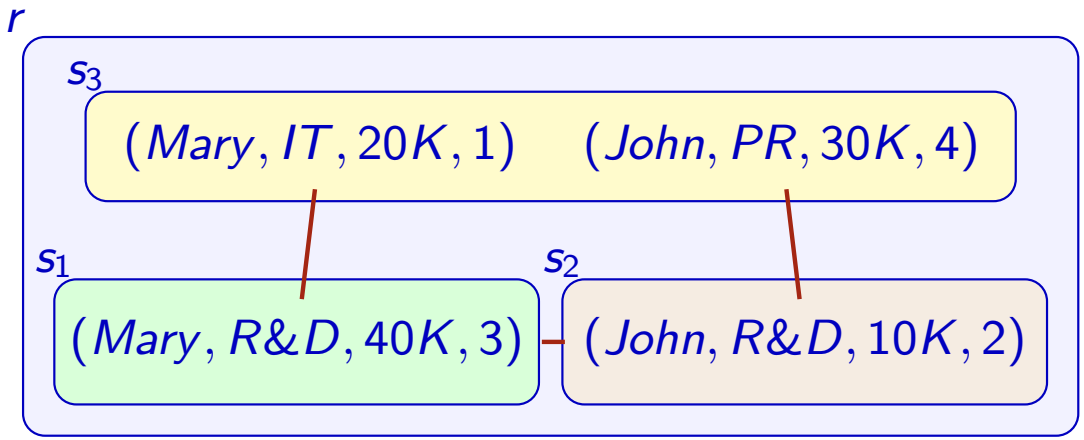
IIDB, March 26, 2006

Motivation

Schema

$Mgr(Name, Dept, Salary, Reports)$

$Key_1 : Name \qquad Key_2 : Dept$



Q_1: John earns more than Mary?

?- $\mathtt{Mgr(John, \_, s_1, \_), Mgr(Mary, \_, s_2, \_), s_1 > s_2.}$

$r \models Q_1$, but is Q_1 really *true*?

Consistent Query Answers

Repairs:

$r_1 = \{(Mary, R\&D, 40K, 3), (John, PR, 30K, 4)\}$

$r_2 = \{(Mary, IT, 20K, 1), (John, R\&D, 10K, 2)\}$

$r_3 = \{(Mary, IT, 20K, 1), (John, PR, 30K, 4)\}$

Q_1 is not consistently true in r!

What if ???

The user knows:

"s_1, s_2 better than s_3"

Motivation (cont.)

Schema

$Mgr(Name, Dept, Salary, Reports)$

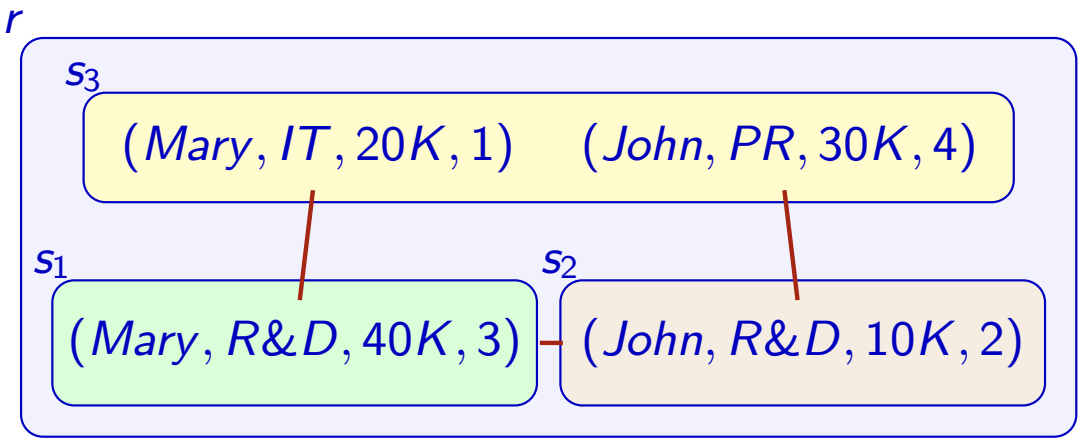
$Key_1 : Name \qquad Key_2 : Dept$

Data cleaning

- s_1,s_2 more reliable than s_3.
- the clean database:

$$r' = \left\{ \begin{array}{c} (Mary, R\&D, 40K, 3), \\ (John, R\&D, 10K, 2) \end{array} \right\}$$

- r' is inconsistent.



Preferred Repairs and CQA

Preferred repairs (maximizing reliablility):

$r_1 = \{(Mary, R\&D, 40K, 3), (John, PR, 30K, 4)\}$

$r_2 = \{(Mary, IT, 20K, 1), (John, R\&D, 10K, 2)\}$

~~$r_3 = \{(Mary, IT, 20K, 1), (John, PR, 30K, 4)\}$~~

Q_2: Mary earns more for less?

$\texttt{?- Mgr(Mary, \_, s}_1\texttt{, r}_1\texttt{), Mgr(John, \_, s}_2\texttt{, r}_2\texttt{), s}_1 > \texttt{s}_2\texttt{, r}_1 < \texttt{r}_2\texttt{.}$

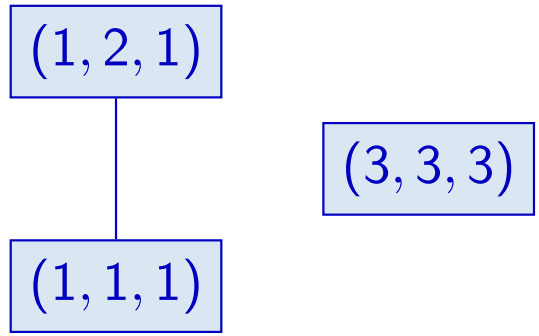
Repairs and Consistent Query Answers

Conflict graph:

- vertices = all tuples
- edges connect conflicting tuples

$R : A \rightarrow B$

A	B	C
1	1	1
1	2	1
3	3	3



Repair:

- a maximal consistent subset of the database
- Rep – all repairs of the database
- Rep = MIS

$r_1 = \{(1, 2, 1), (3, 3, 3)\}$
$r_2 = \{(1, 1, 1), (3, 3, 3)\}$

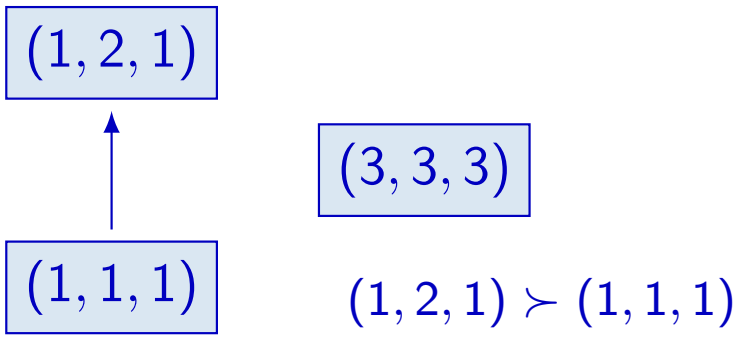
Consistent Query Answers:

answers present in every repair.

Priorities, Preferences, and Cleaning

Priority $\succ$

- an acyclic orientation of the conflict graph
- $\succ$ is total when all edges are oriented



$$\omega_{\succ}(r) = \{t \in r | \neg \exists t' \in r . t' \succ t\}$$

Preferred CQA

- $\mathcal{A}$-$Rep(\succ), \mathcal{B}$-$Rep(\succ), \ldots$ different families of preffered repairs w.r.t. $\succ$
- $\mathcal{X}$-preferred consistent answers w.r.t. $\succ$ are the answers present in every $\mathcal{X}$-preferred repair w.r.t $\succ$

Database cleaning with a total $\succ$

- $r' := \varnothing$
- **while** $\omega_{\succ}(r) \neq \varnothing$ **do**
  1. **choose any** $x \in \omega_{\succ}(r)$
  2. add x to r'
  3. remove x from r with neighbors
- **return** r'

Basic Characterization of Preferred Repairs

($\mathcal{P}1$) Non-emptiness

$$\mathcal{X}\text{-}Rep(\succ) \neq \varnothing$$

($\mathcal{P}2$) Monotonicity

$$\succ_1 \subseteq \succ_2$$
$$\Downarrow$$
$$\mathcal{X}\text{-}Rep(\succ_2) \subseteq \mathcal{X}\text{-}Rep(\succ_1)$$

($\mathcal{P}3$) Non-discrimination

$$\mathcal{X}\text{-}Rep(\varnothing) = Rep$$

($\mathcal{P}4$) Categoricity

$$\succ \text{ is total} \Rightarrow |\mathcal{X}\text{-}Rep(\succ)| = 1$$

Trvial family $\mathcal{T}_1$-$Rep(\succ)$:

1. 1^o if $\succ$ is total then return the clean database
2. 2^o otherwise return Rep

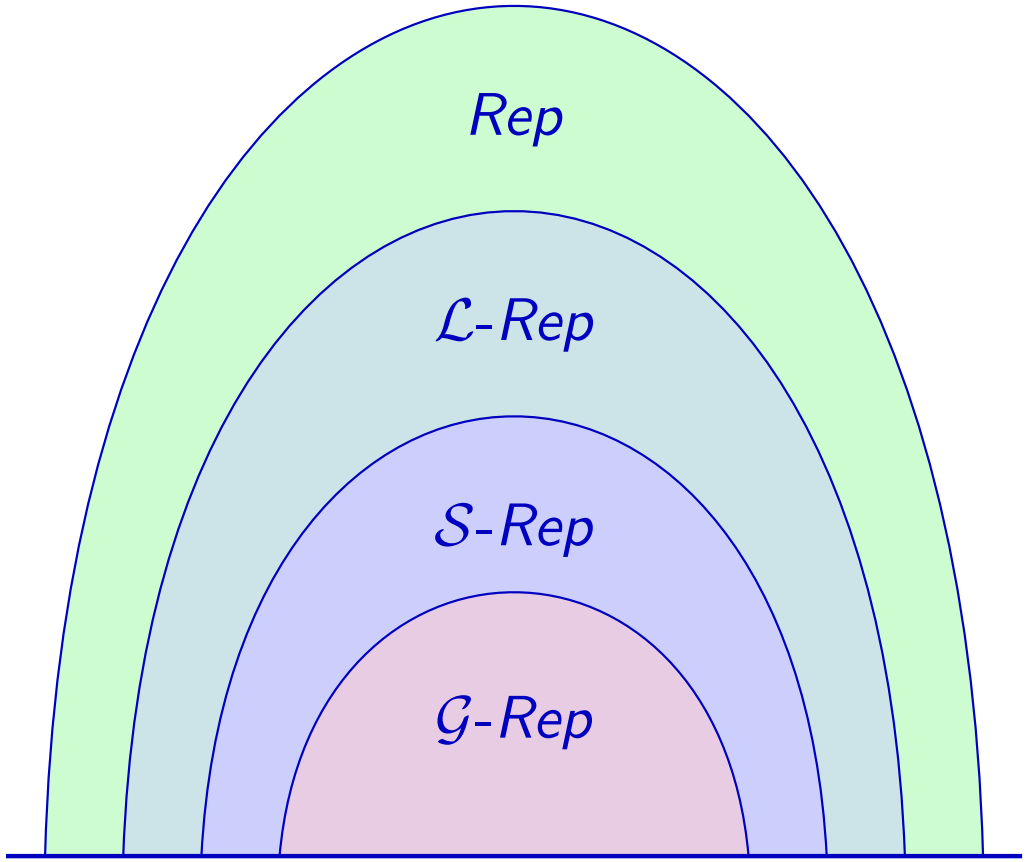
$\mathcal{T}_1$-Rep satisfies $\mathcal{P}1 - \mathcal{P}4$.

Optimal Use of Priorities

Complexity:
(of CQA)

Priority enforcement:

PTIME	None
co-NP-c	Local
co-NP-c	Semi-global
Π_2^p-c	Global



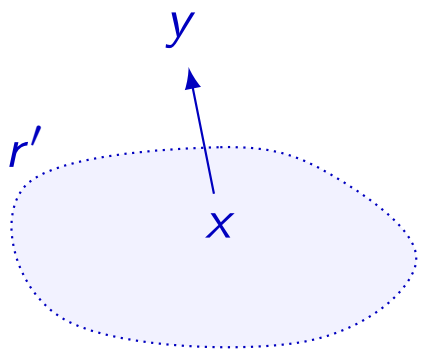
$\mathcal{L}$-*Rep*: Locally Optimal Repairs

r' is *locally optimal* iff

no tuple $x \in r'$ can be replaced with a tuple y such that:

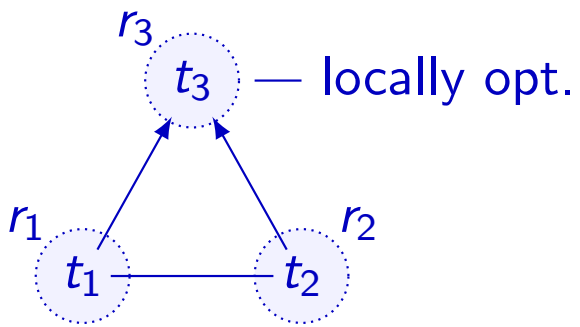
$$y \succ x.$$

(and the result is consistent)

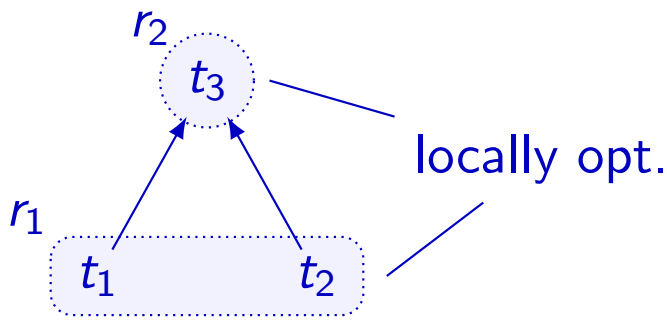


$\mathcal{L}$-*Rep* satisfies $\mathcal{P}1 - \mathcal{P}3$

$\mathcal{L}$-*Rep* is not categorical (not $\mathcal{P}4$)



Key dep.



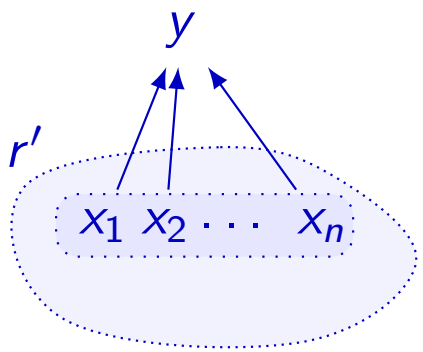
Non-key FD

$\mathcal{S}$-*Rep*: Semi-globally Optimal Repairs

r' is *semi-globally optimal* iff

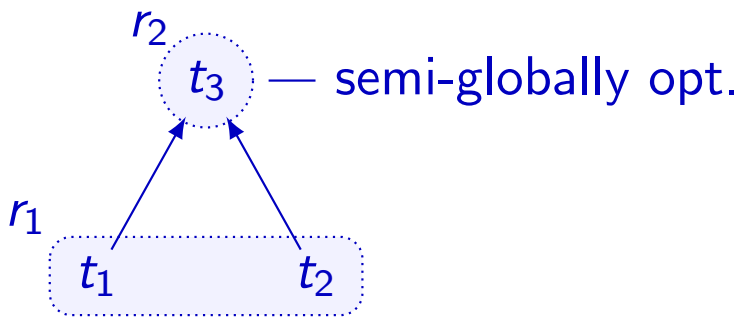
no set $X \subseteq r'$ can be replaced with a tuple y such that:

$$\forall x \in X . y \succ x.$$

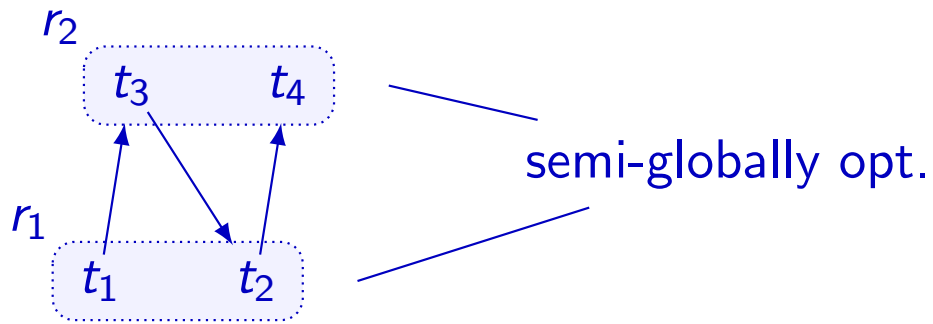


$\mathcal{S}$-*Rep* satisfies $\mathcal{P}1 - \mathcal{P}3$

$\mathcal{S}$-*Rep* is not categorical (not $\mathcal{P}4$)



Non-key FD



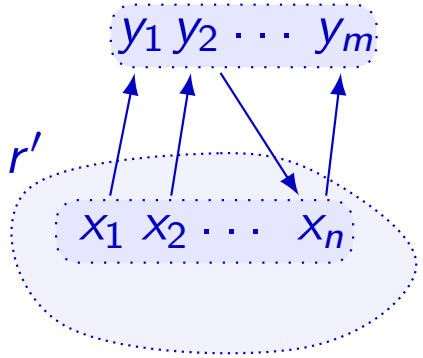
Many FDs

$\mathcal{G}$-*Rep*: Globally Optimal Repairs

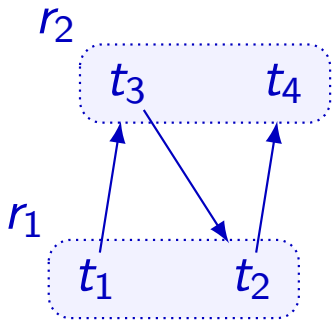
r' is *globally optimal* iff

no set $X \subseteq r'$ can be replaced with a set Y such that:

$$\forall x \in X . \exists y \in Y . y \succ x.$$



$\mathcal{G}$-*Rep* satisfies $\mathcal{P}1 - \mathcal{P}4$



— globally opt.

Many FDs

Alternative characterization

$$\mathcal{G}\text{-}Rep = \ll\text{-maximal repairs}$$

$$r_1 \ll r_2 \Leftrightarrow \forall x \in r_1 \setminus r_2 . \exists y \in r_2 \setminus r_1 . y \succ x.$$

Computational (data-)complexity

	Repair Check	Consistent Answers to	
		$\{\forall, \exists\}$-free queries	conjunctive queries
Rep	PTIME	PTIME	co-NP-complete
$\mathcal{L}$-*Rep*	PTIME	co-NP-complete	co-NP-complete
$\mathcal{S}$-*Rep*	PTIME	co-NP-complete	co-NP-complete
$\mathcal{G}$-*Rep*	co-NP-complete	Π_p^2-complete	Π_p^2-complete

$\mathcal{L}$-*Rep*, $\mathcal{S}$-*Rep*, and $\mathcal{G}$-*Rep*

For one FD computing consistent answers to $\{\exists, \forall\}$-queries is PTIME.

Computing preferred CQA with any family of (semi-globally) optimal repairs satisfying $\mathcal{P}1$ and $\mathcal{P}2$ is co-NP-hard.
(one atom and 2 FDs)

$\mathcal{C}$-*Rep*: Common optimal repairs

Desired properties:

- *optimality* to enforce priority use
- *monotonicity* ($\mathcal{P}2$) to prevent groundless elimination of repairs
- *non-emptiness* ($\mathcal{P}1$)

$\mathcal{C}$-*Rep* - repairs *common* for all families of (globally) optimal repairs satisfying $\mathcal{P}1$ and $\mathcal{P}2$

- $\mathcal{C}$-*Rep* satisfies $\mathcal{P}1 - \mathcal{P}4$
- $\mathcal{C}$-*Rep* $\subseteq$ $\mathcal{G}$-*Rep*
- $\mathcal{C}$-*Rep* $=$ $\mathcal{G}$-*Rep* for priorities that cannot be extended to a cyclic orientation.
- Repair check: PTIME; CQA: co-NP-c

Database cleaning

- $r' := \varnothing$
- **while** $\omega_{\succ}(r) \neq \varnothing$ **do**
  1. **choose any** $x \in \omega_{\succ}(r)$
  2. add x to r'
  3. remove x from r with neighbors
- **return** r'

Alternative characterization

$r' \in \mathcal{C}\text{-}Rep(\succ)$ iff r' can be a result of cleaning the database with $\succ$.

Conclusions

	Repair Check	Consistent Answers to $\{\forall, \exists\}$-free queries	Consistent Answers to conj. queries	Possible Applications
Rep	PTIME	PTIME	co-NP-c	no priorities given
$\mathcal{L}$-*Rep*	PTIME	co-NP-c		key (no duplicates)
$\mathcal{S}$-*Rep*	PTIME	co-NP-c		one FD (duplicates)
$\mathcal{G}$-*Rep*	co-NP-c	Π_p^2-c		many FDs with
$\mathcal{C}$-*Rep*	PTIME	co-NP-c		mutual conflicts

Related Work

S. Flesca, S. Greco, and E. Zumpano. *Active Integrity Constraints.*

$$S_{\succ}(r') = \{(x, y) \in r \times r \mid x \in r'\}$$
$$\mathcal{P}\text{-}Rep(\succ) = \{r' \in Rep \mid S_{\succ}(r') \text{ is maximal}\}$$

- CQA: Π_3^p-complete
- satisfies $\mathcal{P}1$ and $\mathcal{P}3$
- handles cyclic $\succ$, but then
- violates $\mathcal{P}2$ and $\mathcal{P}4$

G. Greco and D. Lembo *Data Integration with Preferences among Sources.*

- repairing a relation by removing tuples has to be *justified* by removing *similar* tuples from other relations.
- satisfies $\mathcal{P}2$, but not $\mathcal{P}1$ (non-emptiness)
- *weakened* framework satisfies $\mathcal{P}1$ but $\mathcal{P}2$ is lost.