Querying Inconsistent Databases

[Some] Past Research and Future Challenges

Jef Wijsen

University of Mons

DBDBD, Ghent, December 21, 2023

Table of Contents

Table of Contents

Inconsistent Data

Perrey Reeves

文A 18 lan

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Perrey Reeves (born 1970 or 1971 (age 52–53))[1] is an American film and television actress. She is best known for her recurring role as Melissa Gold on the television series *Entourage* from 2004 to 2011 and Marissa Jones in the 2003 comedy *Old School*.

Perrey Reeves

Early life [edit]

Reeves was born in New York City and raised in New Hampshire,[2] the daughter of Dr. Alexander Reeves, a

Inconsistent Databases

ACTORS	Name	Gender	Age
	Jolie	F	48
	Pitt	M	59
	Pitt	M	60

Every actor has, at most, one gender and one age:

`ACTORS PRIMARY KEY(Name).`

Data cleaning takes time (and money). Can we already obtain "reliable" information by querying the inconsistent database?

Inconsistent Databases

ACTORS

<u>Name</u>	Gender	Age
Jolie	F	48
Pitt	M	59
Pitt	M	60

Every actor has, at most, one gender and one age:

`ACTORS PRIMARY KEY(Name).`

Data cleaning takes time (and money). Can we already obtain "reliable" information by querying the inconsistent database?

Querying Inconsistent Databases

For ease of presentation, all queries return a Boolean (true/false).

ACTORS

Name	Gender	Age
Jolie	F	48
Pitt	M	59
Pitt	M	60

- Is Pitt's age 60?

$\exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$ is "possibly false".

- Is Pitt older than Jolie?

$$\exists y \exists z \exists v \exists w \left(\begin{array}{l} \text{ACTORS}(\underline{\text{Pitt}}, y, z) \wedge \\ \text{ACTORS}(\underline{\text{Jolie}}, v, w) \wedge z > w \end{array} \right)$$ is "certainly true".

A block is a maximal set of tuples of the same relation that agree on their primary key (blocks are separated by dashed lines).
A repair (or possible world) is obtained by picking a single tuple from each block.

With this notion, "certainly true" means "true in every repair".

If 2 ages are stored for n actors, there are at least 2^n repairs.

Querying Inconsistent Databases

For ease of presentation, all queries return a Boolean (true/false).

ACTORS

<u>Name</u>	Gender	Age
Jolie	F	48
Pitt	M	59
Pitt	M	60

- Is Pitt's age 60?

$$\exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60)) \text{ is "possibly false".}$$

- Is Pitt older than Jolie?

$$\exists y \exists z \exists v \exists w \left(\begin{array}{l} \text{ACTORS}(\underline{\text{Pitt}}, y, z) \wedge \\ \text{ACTORS}(\underline{\text{Jolie}}, v, w) \wedge z > w \end{array} \right) \text{ is "certainly true".}$$

A block is a maximal set of tuples of the same relation that agree on their primary key (blocks are separated by dashed lines).
A repair (or possible world) is obtained by picking a single tuple from each block.

With this notion, "certainly true" means "true in every repair".

If 2 ages are stored for n actors, there are at least 2^n repairs.

Consistent Query Answering for Primary Keys

Given a Boolean query Q, define the following decision problem:

Problem CERTAINTY(Q)

Input: A database instance that may violate primary-key constraints.

Question: Is Q true in every repair?

Example

If $Q_{60} = \exists y\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, 60))$, then the answer to CERTAINTY(Q_{60}) is "no" on our example database.

Remark

We assume that each relation name has a fixed primary key. Primary-key positions will be underlined. Primary keys can thus be derived from the query.

Consistent Query Answering for Primary Keys

Given a Boolean query Q, define the following decision problem:

Problem CERTAINTY(Q)

Input: A database instance that may violate primary-key constraints.

Question: Is Q true in every repair?

Example

If $Q_{60} = \exists y\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, 60))$, then the answer to CERTAINTY(Q_{60}) is "no" on our example database.

Remark

We assume that each relation name has a fixed primary key. Primary-key positions will be underlined. Primary keys can thus be derived from the query.

Table of Contents

Solving CERTAINTY(Q)

- A general solution in exponential time:

  Input: a database D
  for *each repair R of D* **do**
    if *Q is false in R* **then**
      return "no" (and halt)
  return "yes"

  ☞ CERTAINTY(Q) is in coNP for first-order queries Q.

- A smarter solution for $Q_{60} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$:

  Input: a database D
  Let $\overline{Q}_{60} := \exists y \exists z\,(\text{ACTORS}(\underline{\text{Pitt}}, y, z) \land \neg(z = 60))$
  if *Q_{60} is true and $\overline{Q}_{60}$ is false in D* **then**
    return "yes"
  else
    return "no"

  ☞ CERTAINTY(Q_{60}) is in the low complexity class FO (i.e., solvable by a first-order logic formula).

Solving CERTAINTY(Q)

- A general solution in exponential time:

  Input: a database D
  for *each repair R of D* **do**
      if *Q is false in R* **then**
          return "no" (and halt)
  return "yes"

  ☞ CERTAINTY(Q) is in coNP for first-order queries Q.

- A smarter solution for $Q_{60} = \exists y\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, 60))$:

  Input: a database D
  Let $\overline{Q}_{60} := \exists y \exists z\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, z) \wedge \neg(z = 60))$
  if *Q_{60} is true and $\overline{Q}_{60}$ is false in D* **then**
      return "yes"
  else
      return "no"

  ☞ CERTAINTY(Q_{60}) is in the low complexity class FO (i.e., solvable by a first-order logic formula).

Solving CERTAINTY(Q)

- A general solution in exponential time:

  Input: a database D
  for *each repair R of D* **do**
      if *Q is false in R* **then**
          return "no" (and halt)
  return "yes"

  ☞ CERTAINTY(Q) is in coNP for first-order queries Q.

- A smarter solution for $Q_{60} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$:

  Input: a database D
  Let $\overline{Q}_{60} := \exists y \exists z\,(\text{ACTORS}(\underline{\text{Pitt}}, y, z) \land \neg(z = 60))$
  if Q_{60} *is true and* $\overline{Q}_{60}$ *is false in D* **then**
      return "yes"
  else
      return "no"

  ☞ CERTAINTY(Q_{60}) is in the low complexity class FO (i.e., solvable by a first-order logic formula).

Solving CERTAINTY(Q)

- A general solution in exponential time:

  Input: a database D
  for *each repair R of D* **do**
      if *Q is false in R* **then**
          return "no" (and halt)
  return "yes"

  ☞ CERTAINTY(Q) is in coNP for first-order queries Q.

- A smarter solution for $Q_{60} = \exists y\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, 60))$:

  Input: a database D
  Let $\overline{Q}_{60} := \exists y \exists z\,(\mathrm{ACTORS}(\underline{\mathrm{Pitt}}, y, z) \land \neg\,(z = 60))$
  if *Q_{60} is true and $\overline{Q}_{60}$ is false in D* **then**
      return "yes"
  else
      return "no"

  ☞ CERTAINTY(Q_{60}) is in the low complexity class FO (i.e., solvable by a first-order logic formula).

SQL Rewriting

```
SELECT ‘yes’
FROM ACTORS
WHERE Name = ‘Pitt’
AND Age = 60;
```

$\rightsquigarrow$

```
SELECT ‘yes’
FROM ACTORS
WHERE Name = ‘Pitt’
AND Age = 60
AND NOT EXISTS (SELECT *
                FROM ACTORS
                WHERE Name = ‘Pitt’
                AND Age <> 60);
```

The Good, the Bad and the Ugly

Theorem (DBDBD, 2023)

For $Q_{\text{good}} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$, *the decision problem* CERTAINTY(Q_{good}) *is in* FO.

Theorem ([W., 2010])

For $Q_{\text{bad}} = \exists x \exists y\,(R(\underline{x}, y) \wedge S(\underline{y}, x))$, *the decision problem* CERTAINTY(Q_{bad}) *is in* P $\setminus$ FO.

P is the class of decision problems solvable in polynomial time.

Theorem ([Chomicki and Marcinkowski, 2005])

For $Q_{\text{ugly}} = \exists x_1 \exists x_2 \exists z\,(\text{ACTORS}(\underline{x_1}, \text{M}, z) \wedge \text{ACTORS}(\underline{x_2}, \text{F}, z))$, *the decision problem* CERTAINTY(Q_{ugly}) *is* coNP*-complete.*

The Good, the Bad and the Ugly

Theorem (DBDBD, 2023)

For $Q_{\text{good}} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$*, the decision problem* $\text{CERTAINTY}(Q_{\text{good}})$ *is in* FO*.*

Theorem ([W., 2010])

For $Q_{\text{bad}} = \exists x \exists y\,(R(\underline{x}, y) \wedge S(\underline{y}, x))$*, the decision problem* $\text{CERTAINTY}(Q_{\text{bad}})$ *is in* $\mathsf{P} \setminus \mathsf{FO}$*.*

P is the class of decision problems solvable in polynomial time.

Theorem ([Chomicki and Marcinkowski, 2005])

For $Q_{\text{ugly}} = \exists x_1 \exists x_2 \exists z\,(\text{ACTORS}(\underline{x_1}, \text{M}, z) \wedge \text{ACTORS}(\underline{x_2}, \text{F}, z))$*, the decision problem* $\text{CERTAINTY}(Q_{\text{ugly}})$ *is* coNP*-complete.*

The Good, the Bad and the Ugly

Theorem (DBDBD, 2023)

For $Q_{\text{good}} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$*, the decision problem* CERTAINTY(Q_{good}) *is in* FO*.*

Theorem ([W., 2010])

For $Q_{\text{bad}} = \exists x \exists y\,\big(R(\underline{x}, y) \land S(\underline{y}, x)\big)$*, the decision problem* CERTAINTY(Q_{bad}) *is in* $\mathsf{P} \setminus \mathsf{FO}$*.*

P is the class of decision problems solvable in polynomial time.

Theorem ([Chomicki and Marcinkowski, 2005])

For $Q_{\text{ugly}} = \exists x_1 \exists x_2 \exists z\,(\text{ACTORS}(\underline{x_1}, \text{M}, z) \land \text{ACTORS}(\underline{x_2}, \text{F}, z))$*, the decision problem* CERTAINTY(Q_{ugly}) *is* coNP*-complete.*

The Good, the Bad and the Ugly

Theorem (DBDBD, 2023)

For $Q_{\text{good}} = \exists y\,(\text{ACTORS}(\underline{\text{Pitt}}, y, 60))$, *the decision problem* $\text{CERTAINTY}(Q_{\text{good}})$ *is in* FO.

Theorem ([W., 2010])

For $Q_{\text{bad}} = \exists x \exists y\,\big(R(\underline{x}, y) \wedge S(\underline{y}, x)\big)$, *the decision problem* $\text{CERTAINTY}(Q_{\text{bad}})$ *is in* $\text{P} \setminus \text{FO}$.

P is the class of decision problems solvable in polynomial time.

Theorem ([Chomicki and Marcinkowski, 2005])

For $Q_{\text{ugly}} = \exists x_1 \exists x_2 \exists z\,\big(\text{ACTORS}(\underline{x_1}, \text{M}, z) \wedge \text{ACTORS}(\underline{x_2}, \text{F}, z)\big)$, *the decision problem* $\text{CERTAINTY}(Q_{\text{ugly}})$ *is* coNP*-complete*.

Research Agenda

- We aim to go beyond the task of determining CERTAINTY(Q) for individual queries Q.
- For "reasonable" classes $\mathcal{C}$ of queries, write an algorithm for the following problem:

Complexity Classification Task

Input: A query Q in the class $\mathcal{C}$.

Task: The computational complexity of CERTAINTY(Q), in terms of complexity classes like FO, P, coNP-complete,...

Research Agenda

- We aim to go beyond the task of determining CERTAINTY(Q) for individual queries Q.
- For "reasonable" classes $\mathcal{C}$ of queries, write an algorithm for the following problem:

Complexity Classification Task

Input: A query Q in the class $\mathcal{C}$.

Task: The computational complexity of CERTAINTY(Q), in terms of complexity classes like FO, P, coNP-complete,...

Which Query Classes Are "Reasonable"?

- The class of (Boolean) conjunctive queries (a.k.a. Select-Project-Join queries):

$$\exists \vec{u} \left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \land R_2(\underline{\vec{x}_2}, \vec{y}_2) \land \cdots \land R_n(\underline{\vec{x}_n}, \vec{y}_n) \right). \qquad (1)$$

- The class of disjunctions of conjunctive queries (a.k.a. UCQ queries):

$$Q_1 \lor Q_2 \lor \cdots \lor Q_m,$$

where each Q_i is of the form (1).

Which Query Classes Are "Reasonable"?

- The class of (Boolean) conjunctive queries (a.k.a. Select-Project-Join queries):

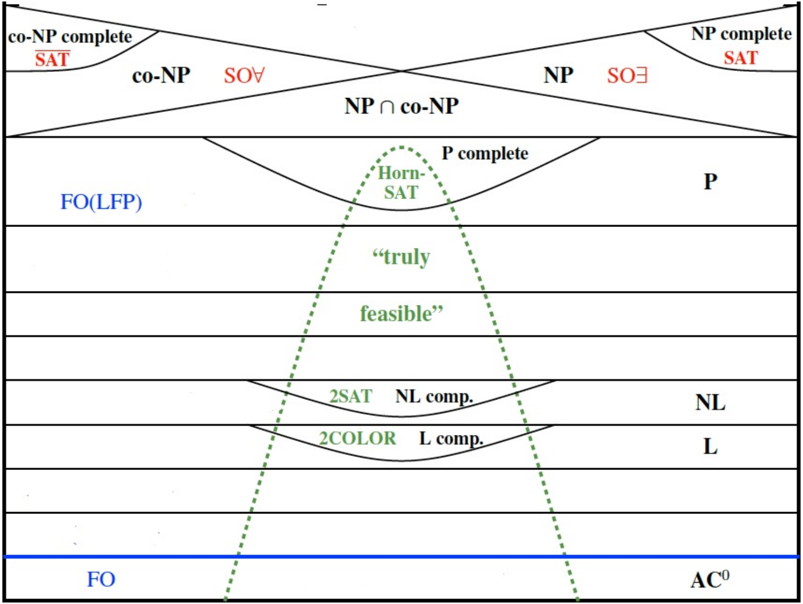
$$\exists \vec{u} \left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge R_2(\underline{\vec{x}_2}, \vec{y}_2) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n) \right). \qquad (1)$$

- The class of disjunctions of conjunctive queries (a.k.a. UCQ queries):

$$Q_1 \vee Q_2 \vee \cdots \vee Q_m,$$

  where each Q_i is of the form (1).

Which Complexity Classes?



Classifying CERTAINTY(Q) in P/coNP-complete is Hard

Recall

If P $\neq$ coNP, then some problems in coNP are neither in P nor coNP-complete.

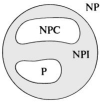


Figure 7.1 The world of NP, reprised (assuming P$\neq$NP).

Conjecture

If Q is a disjunction of conjunctive queries, then CERTAINTY(Q) *is in* P *or* coNP*-complete.*

Theorem ([Fontaine, 2015])

The above conjecture implies Bulatov's dichotomy theorem for the conservative constraint satisfaction problem (CSP).

Classifying CERTAINTY(Q) in P/coNP-complete is Hard

Recall

If P $\neq$ coNP, then some problems in coNP are neither in P nor coNP-complete.

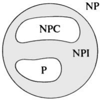


Figure 7.1 The world of NP, reprised (assuming P $\neq$ NP).

Conjecture

If Q is a disjunction of conjunctive queries, then CERTAINTY(Q) *is in* P *or* coNP*-complete.*

Theorem ([Fontaine, 2015])

The above conjecture implies Bulatov's dichotomy theorem for the conservative constraint satisfaction problem (CSP).

Classifying CERTAINTY(Q) in P/coNP-complete is Hard

Recall

If P $\neq$ coNP, then some problems in coNP are neither in P nor coNP-complete.

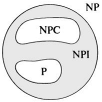


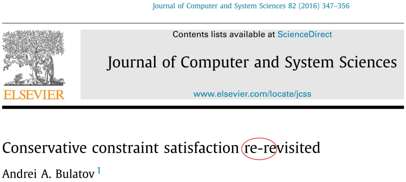
Figure 7.1 The world of NP, reprised (assuming P$\neq$NP).

Conjecture

If Q is a disjunction of conjunctive queries, then CERTAINTY(Q) *is in* P *or* coNP*-complete.*

Theorem ([Fontaine, 2015])

The above conjecture implies Bulatov's dichotomy theorem for the conservative constraint satisfaction problem (CSP).


Journal of Computer and System Sciences 82 (2016) 347–356

Contents lists available at ScienceDirect

ELSEVIER

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Conservative constraint satisfaction re-revisited

Andrei A. Bulatov [1]

Is it Easier for Conjunctive Queries?

Conjecture

If Q is of the form $\exists \vec{u} \left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n) \right)$, then CERTAINTY($Q$) *is in* P *or* coNP*-complete.*

Theorem ([Koutris and W., 2017])

The above conjecture holds under the assumption that $R_i \neq R_j$ whenever $i \neq j$.

Theorem ([Padmanabha et al., 2023])

The above conjecture holds under the assumption that $n = 2$.

Is it Easier for Conjunctive Queries?

Conjecture

If Q is of the form $\exists \vec{u}\left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n)\right)$, *then* CERTAINTY$(Q)$ *is in* P *or* coNP*-complete.*

Theorem ([Koutris and W., 2017])

The above conjecture holds under the assumption that $R_i \neq R_j$ whenever $i \neq j$.

Theorem ([Padmanabha et al., 2023])

The above conjecture holds under the assumption that $n = 2$.

Is it Easier for Conjunctive Queries?

Conjecture

If Q is of the form $\exists\vec{u}\left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n)\right)$, then CERTAINTY($Q$) *is in* P *or* coNP*-complete.*

Theorem ([Koutris and W., 2017])

The above conjecture holds under the assumption that $R_i \neq R_j$ whenever $i \neq j$.

Theorem ([Padmanabha et al., 2023])

The above conjecture holds under the assumption that $n = 2$.

Table of Contents

The Good Among The Good, the Bad and the Ugly

A directed graph, called attack graph, is defined for every conjunctive query.

Theorem ([Koutris and W., 2017])

Let $Q = \exists \vec{u} \left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n) \right)$ with $R_i \neq R_j$ for $i \neq j$. Then,

- *if Q's attack graph is acyclic, then* CERTAINTY(Q) *is in* FO;
- *if Q's attack graph is cyclic, then* CERTAINTY(Q) *is* L*-hard.*

The Good Among The Good, the Bad and the Ugly

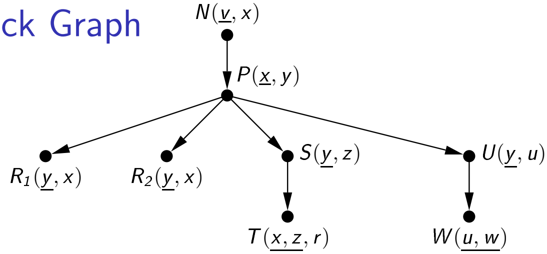
A directed graph, called attack graph, is defined for every conjunctive query.

Theorem ([Koutris and W., 2017])

Let $Q = \exists\vec{u}\left(R_1(\underline{\vec{x}_1}, \vec{y}_1) \land \cdots \land R_n(\underline{\vec{x}_n}, \vec{y}_n)\right)$ *with* $R_i \neq R_j$ *for* $i \neq j$*. Then,*

- *if Q's attack graph is acyclic, then* CERTAINTY(Q) *is in* FO*;*
- *if Q's attack graph is cyclic, then* CERTAINTY(Q) *is* L*-hard.*

Attack Graph



$N^+ = \{v\}$

$P^+ = \{x\}$

${R_1}^+ = \{y, x, z, r, u\}$

${R_2}^+ = \{y, x, z, r, u\}$

$S^+ = \{y, x, u\}$

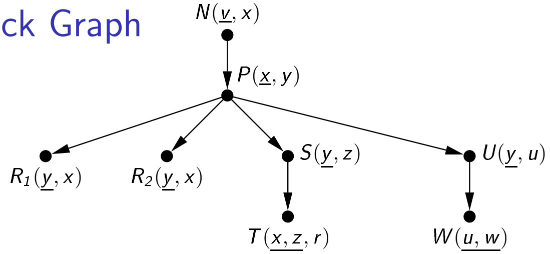
$U^+ = \{y, x, z, r\}$

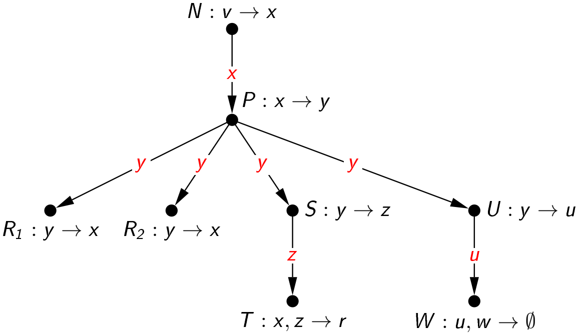
$T^+ = \{x, z, y, u\}$

$W^+ = \{u, w\}$

S^+, e.g., is the closure of S's key w.r.t. all other FDs.
S can attack with $z \notin S^+$.

Attack Graph





$N^+ = \{v\}$

$P^+ = \{x\}$

${R_1}^+ = \{y, x, z, r, u\}$

${R_2}^+ = \{y, x, z, r, u\}$

$S^+ = \{y, x, u\}$

$U^+ = \{y, x, z, r\}$

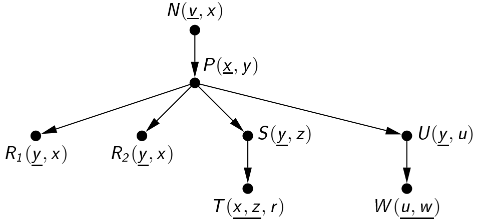
$T^+ = \{x, z, y, u\}$

$W^+ = \{u, w\}$

S^+, e.g., is the closure of S's key w.r.t. all other FDs.
S can attack with $z \notin S^+$.

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

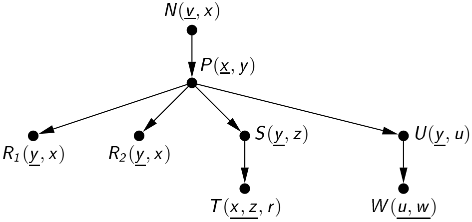
φ_N is true in the database $\iff$ Q is true in every repair.

$$\begin{aligned}
\varphi_N &:= \exists v\,(\exists x\,(N(\underline{v}, x)) \wedge \neg\exists x\,(N(\underline{v}, x) \wedge \neg\varphi_P(x))) \\
\varphi_P(x) &:= \exists y\,(P(\underline{x}, y)) \\
&\quad \wedge \neg\exists y\,(P(\underline{x}, y) \wedge \neg(\varphi_{R_1}(x, y) \wedge \varphi_{R_2}(x, y) \wedge \varphi_S(x, y) \wedge \varphi_U(y))) \\
\varphi_{R_i}(x, y) &:= R_i(\underline{y}, x) \wedge \neg\exists x'\,(R_i(\underline{y}, x') \wedge x' \neq x),\ 1 \leq i \leq 2 \\
\varphi_S(x, y) &:= \exists z\,(S(\underline{y}, z)) \wedge \neg\exists z\,(S(\underline{y}, z) \wedge \neg\exists r\,(T(\underline{x, z}, r))) \\
\varphi_U(y) &:= \exists u\,(U(\underline{y}, u)) \wedge \neg\exists u\,(U(\underline{y}, u) \wedge \neg\exists w\,(W(\underline{u, w})))
\end{aligned}$$

In SQL, we get 4 embedded NOT EXISTS...

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

φ_N is true in the database $\iff$ Q is true in every repair.

$$\varphi_N := \exists v\,(\exists x\,(N(\underline{v}, x)) \land \neg\exists x\,(N(\underline{v}, x) \land \neg\varphi_P(x)))$$

$$\varphi_P(x) := \exists y\,(P(\underline{x}, y)) \land \neg\exists y\,(P(\underline{x}, y) \land \neg(\varphi_{R_1}(x, y) \land \varphi_{R_2}(x, y) \land \varphi_S(x, y) \land \varphi_U(y)))$$

$$\varphi_{R_i}(x, y) := R_i(\underline{y}, x) \land \neg\exists x'\,(R_i(\underline{y}, x') \land x' \neq x)\,, 1 \leq i \leq 2$$

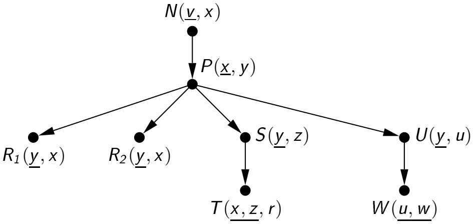
$$\varphi_S(x, y) := \exists z\,(S(\underline{y}, z)) \land \neg\exists z\,(S(\underline{y}, z) \land \neg\exists r\,(T(\underline{x, z}, r)))$$

$$\varphi_U(y) := \exists u\,(U(\underline{y}, u)) \land \neg\exists u\,(U(\underline{y}, u) \land \neg\exists w\,(W(\underline{u, w})))$$

In SQL, we get 4 embedded NOT EXISTS.

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

φ_N is true in the database $\iff$ Q is true in every repair.

$$\varphi_N := \exists v \left(\exists x \left(N(\underline{v}, x)\right) \land \neg\exists x \left(N(\underline{v}, x) \land \neg\varphi_P(x)\right)\right)$$

$$\varphi_P(x) := \exists y \left(P(\underline{x}, y)\right)$$

$$\land \neg\exists y \left(P(\underline{x}, y) \land \neg\left(\varphi_{R_1}(x, y) \land \varphi_{R_2}(x, y) \land \varphi_S(x, y) \land \varphi_U(y)\right)\right)$$

$$\varphi_{R_i}(x, y) := R_i(\underline{y}, x) \land \neg\exists x' \left(R_i(\underline{y}, x') \land x' \neq x\right), 1 \leq i \leq 2$$

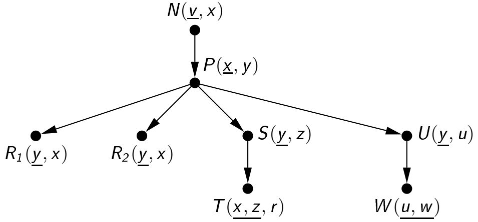
$$\varphi_S(x, y) := \exists z \left(S(\underline{y}, z)\right) \land \neg\exists z \left(S(\underline{y}, z) \land \neg\exists r \left(T(\underline{x, z}, r)\right)\right)$$

$$\varphi_U(y) := \exists u \left(U(\underline{y}, u)\right) \land \neg\exists u \left(U(\underline{y}, u) \land \neg\exists w \left(W(\underline{u, w})\right)\right)$$

In SQL, we get 4 embedded NOT EXISTS...

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

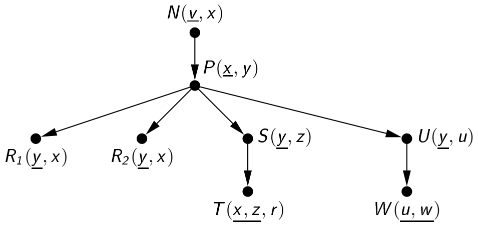
φ_N is true in the database $\iff$ Q is true in every repair.

$$\begin{aligned}
\varphi_N &:= \exists v\,(\exists x\,(N(\underline{v}, x)) \wedge \neg\exists x\,(N(\underline{v}, x) \wedge \neg\varphi_P(x)))\\
\varphi_P(x) &:= \exists y\,(P(\underline{x}, y))\\
&\quad \wedge \neg\exists y\,(P(\underline{x}, y) \wedge \neg(\varphi_{R_1}(x, y) \wedge \varphi_{R_2}(x, y) \wedge \varphi_S(x, y) \wedge \varphi_U(y)))\\
\varphi_{R_i}(x, y) &:= R_i(\underline{y}, x) \wedge \neg\exists x'\left(R_i(\underline{y}, x') \wedge x' \neq x\right), 1 \leq i \leq 2\\
\varphi_S(x, y) &:= \exists z\,(S(\underline{y}, z)) \wedge \neg\exists z\left(S(\underline{y}, z) \wedge \neg\exists r\left(T(\underline{x, z}, r)\right)\right)\\
\varphi_U(y) &:= \exists u\,(U(\underline{y}, u)) \wedge \neg\exists u\left(U(\underline{y}, u) \wedge \neg\exists w\left(W(\underline{u, w})\right)\right)
\end{aligned}$$

In SQL, we get 4 embedded NOT EXISTS...

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

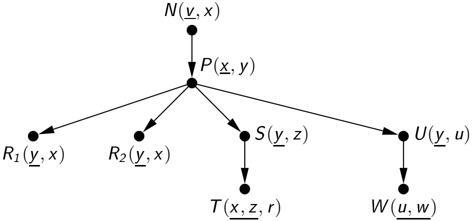
φ_N is true in the database $\iff$ Q is true in every repair.

$$
\begin{aligned}
\varphi_N &:= \exists v \, (\exists x \, (N(\underline{v}, x)) \wedge \neg \exists x \, (N(\underline{v}, x) \wedge \neg \varphi_P(x))) \\
\varphi_P(x) &:= \exists y \, (P(\underline{x}, y)) \\
&\quad \wedge \neg \exists y \, (P(\underline{x}, y) \wedge \neg (\varphi_{R_1}(x, y) \wedge \varphi_{R_2}(x, y) \wedge \varphi_S(x, y) \wedge \varphi_U(y))) \\
\varphi_{R_i}(x, y) &:= R_i(\underline{y}, x) \wedge \neg \exists x' \left(R_i(\underline{y}, x') \wedge x' \neq x \right), 1 \leq i \leq 2 \\
\varphi_S(x, y) &:= \exists z \left(S(\underline{y}, z) \right) \wedge \neg \exists z \left(S(\underline{y}, z) \wedge \neg \exists r \left(T(\underline{x, z}, r) \right) \right) \\
\varphi_U(y) &:= \exists u \left(U(\underline{y}, u) \right) \wedge \neg \exists u \left(U(\underline{y}, u) \wedge \neg \exists w \left(W(\underline{u, w}) \right) \right)
\end{aligned}
$$

In SQL, we get 4 embedded NOT EXISTS. . .

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

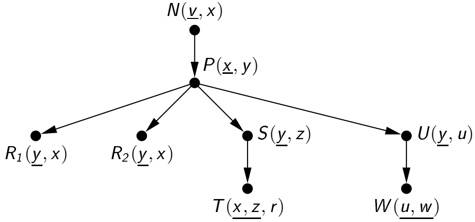
φ_N is true in the database $\iff$ Q is true in every repair.

$$
\begin{aligned}
\varphi_N &:= \exists v\,(\exists x\,(N(\underline{v}, x)) \land \neg\exists x\,(N(\underline{v}, x) \land \neg\varphi_P(x)))\\
\varphi_P(x) &:= \exists y\,(P(\underline{x}, y))\\
&\quad \land \neg\exists y\,(P(\underline{x}, y) \land \neg(\varphi_{R_1}(x, y) \land \varphi_{R_2}(x, y) \land \varphi_S(x, y) \land \varphi_U(y)))\\
\varphi_{R_i}(x, y) &:= R_i(\underline{y}, x) \land \neg\exists x'\left(R_i(\underline{y}, x') \land x' \neq x\right), 1 \leq i \leq 2\\
\varphi_S(x, y) &:= \exists z\,(S(\underline{y}, z)) \land \neg\exists z\left(S(\underline{y}, z) \land \neg\exists r\left(T(\underline{x, z}, r)\right)\right)\\
\varphi_U(y) &:= \exists u\,(U(\underline{y}, u)) \land \neg\exists u\left(U(\underline{y}, u) \land \neg\exists w\left(W(\underline{u, w})\right)\right)
\end{aligned}
$$

In SQL, we get 4 embedded NOT EXISTS...

Attack Graph and (Consistent) First-Order Rewriting

Q :



We construct a first-order formula φ_N such that for every database:

φ_N is true in the database $\iff$ Q is true in every repair.

$$
\begin{aligned}
\varphi_N &:= \exists v \, (\exists x \, (N(\underline{v}, x)) \land \neg \exists x \, (N(\underline{v}, x) \land \neg \varphi_P(x))) \\
\varphi_P(x) &:= \exists y \, (P(\underline{x}, y)) \\
&\quad \land \neg \exists y \, (P(\underline{x}, y) \land \neg (\varphi_{R_1}(x, y) \land \varphi_{R_2}(x, y) \land \varphi_S(x, y) \land \varphi_U(y))) \\
\varphi_{R_i}(x, y) &:= R_i(\underline{y}, x) \land \neg \exists x' \left(R_i(\underline{y}, x') \land x' \neq x \right), 1 \leq i \leq 2 \\
\varphi_S(x, y) &:= \exists z \left(S(\underline{y}, z) \right) \land \neg \exists z \left(S(\underline{y}, z) \land \neg \exists r \left(T(\underline{x, z}, r) \right) \right) \\
\varphi_U(y) &:= \exists u \left(U(\underline{y}, u) \right) \land \neg \exists u \left(U(\underline{y}, u) \land \neg \exists w \left(W(\underline{u, w}) \right) \right)
\end{aligned}
$$

In SQL, we get 4 embedded NOT EXISTS...

Observation Regarding Correctness

$$\varphi_{R_1}(x,y) := R_1(\underline{y},x) \wedge \neg\exists x' \left(R_1(\underline{y},x') \wedge x' \neq x\right)$$
$$\varphi_{R_2}(x,y) := R_2(\underline{y},x) \wedge \neg\exists x' \left(R_2(\underline{y},x') \wedge x' \neq x\right)$$

In words,

- $\varphi_{R_1}(x,y)$: there is a singleton block containing $R_1(\underline{y},x)$;
- $\varphi_{R_2}(x,y)$: there is a singleton block containing $R_2(\underline{y},x)$.

That is, R_i-blocks of size ≥ 2 can be ignored. For example,

R_1	$\underline{y}$	x
	a	c_1
	a	c_2

R_2	$\underline{y}$	x
	a	c_1
	a	c_2

To construct a repair that falsifies the query, pick $R_1(\underline{a},c_i)$ and $R_2(\underline{a},c_j)$ such that $c_i \neq c_j$.

Observation Regarding Correctness

$$\varphi_{R_1}(x,y) := R_1(\underline{y},x) \wedge \neg\exists x' \left(R_1(\underline{y},x') \wedge x' \neq x\right)$$
$$\varphi_{R_2}(x,y) := R_2(\underline{y},x) \wedge \neg\exists x' \left(R_2(\underline{y},x') \wedge x' \neq x\right)$$

In words,

- $\varphi_{R_1}(x,y)$: there is a singleton block containing $R_1(\underline{y},x)$;
- $\varphi_{R_2}(x,y)$: there is a singleton block containing $R_2(\underline{y},x)$.

That is, R_i-blocks of size ≥ 2 can be ignored. For example,

R_1

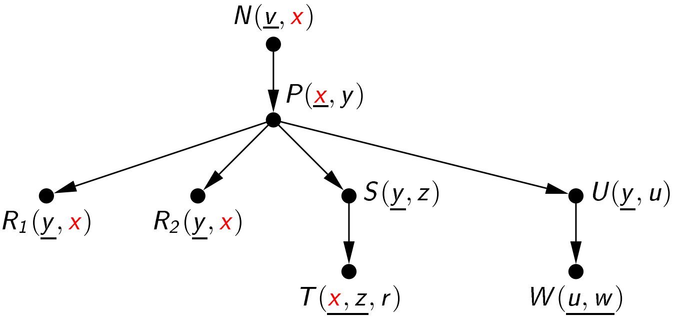
$\underline{y}$	x
a	c_1
a	c_2

R_2

$\underline{y}$	x
a	c_1
a	c_2

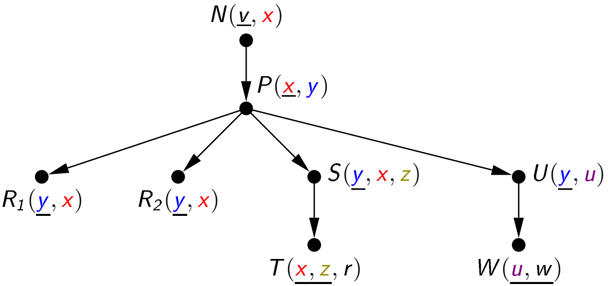
To construct a repair that falsifies the query, pick $R_1(\underline{a},c_i)$ and $R_2(\underline{a},c_j)$ such that $c_i \neq c_j$.

Attack Graph $\neq$ Join Tree



The subgraph induced by atoms that contain x is not connected.

Attack Graph that Is a Join Tree

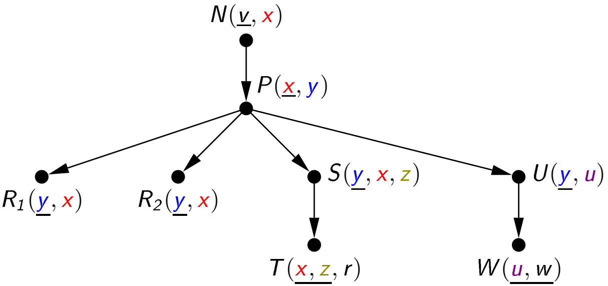


Moreover, every internal node V has zero indegree in the attack graph of the subquery rooted at V ($V \in \{P, S, U\}$).
Such a join tree is called a Pair-Pruning Join Tree (PPJT).
Yannakakis' algorithm extends to the inconsistent setting:

Theorem ([Fan et al., 2023])

If Q has a PPJT, then CERTAINTY(Q) *is in* LIN *(i.e., problems solvable in linear time).*

Attack Graph that Is a Join Tree

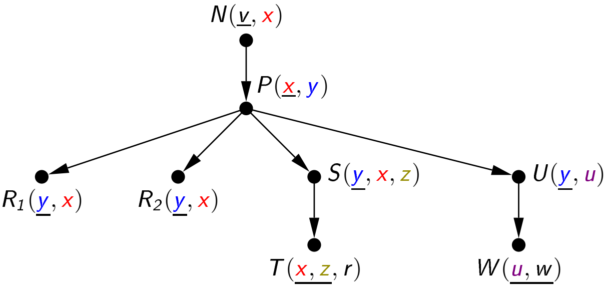


Moreover, every internal node V has zero indegree in the attack graph of the subquery rooted at V ($V \in \{P, S, U\}$).
Such a join tree is called a Pair-Pruning Join Tree (PPJT).
Yannakakis' algorithm extends to the inconsistent setting:

Theorem ([Fan et al., 2023])

If Q has a PPJT, then CERTAINTY(Q) *is in* LIN *(i.e., problems solvable in linear time).*

Attack Graph that Is a Join Tree



Moreover, every internal node V has zero indegree in the attack graph of the subquery rooted at V ($V \in \{P, S, U\}$).
Such a join tree is called a Pair-Pruning Join Tree (PPJT).
Yannakakis' algorithm extends to the inconsistent setting:

Theorem ([Fan et al., 2023])

If Q has a PPJT, then CERTAINTY(Q) *is in* LIN *(i.e., problems solvable in linear time).*

Yannakakis+Pruning

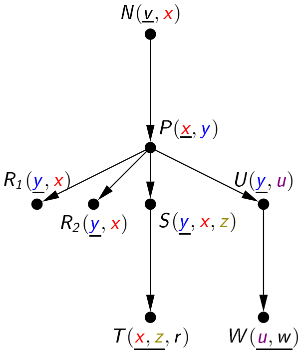
$$Answer(\text{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\text{fadingkey}}(v)$$
$$N^{\text{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\text{join}}(x)$$

$$P^{\text{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\text{fadingkey}}(x)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\text{join}}(y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\text{join}}(x, y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\text{join}}(x, y)$$

$$U^{\text{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\text{fadingkey}}(y)$$
$$U^{\text{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\text{join}}(u)$$

$$S^{\text{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\text{fadingkey}}(y)$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\text{join}}(x, z)$$

$$R_i^{\text{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\text{fadingkey}}(y)$$
$$R_i^{\text{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$
$$(1 \leq i \leq 2)$$

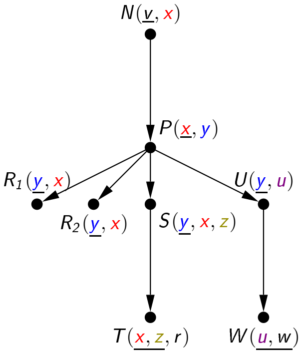


$$T^{\text{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$
$$W^{\text{join}}(u) \leftarrow W(\underline{u, w})$$

Yannakakis+Pruning

$$Answer(\text{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\text{fadingkey}}(v)$$
$$N^{\text{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\text{join}}(x)$$

$$P^{\text{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\text{fadingkey}}(x)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\text{join}}(y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\text{join}}(x, y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\text{join}}(x, y)$$



$$U^{\text{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\text{fadingkey}}(y)$$
$$U^{\text{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\text{join}}(u)$$

$$S^{\text{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\text{fadingkey}}(y)$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\text{join}}(x, z)$$

$$T^{\text{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$
$$W^{\text{join}}(u) \leftarrow W(\underline{u, w})$$

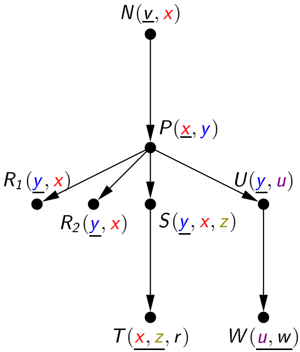
$$R_i^{\text{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\text{fadingkey}}(y)$$
$$R_i^{\text{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$
$$(1 \leq i \leq 2)$$

Yannakakis+Pruning

$$\mathit{Answer}(\text{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\text{fadingkey}}(v)$$
$$N^{\text{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\text{join}}(x)$$

$$P^{\text{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\text{fadingkey}}(x)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\text{join}}(y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\text{join}}(x, y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\text{join}}(x, y)$$

$$U^{\text{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\text{fadingkey}}(y)$$
$$U^{\text{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\text{join}}(u)$$



$$S^{\text{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\text{fadingkey}}(y)$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\text{join}}(x, z)$$

$$T^{\text{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$
$$W^{\text{join}}(u) \leftarrow W(\underline{u, w})$$

$$R_i^{\text{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\text{fadingkey}}(y)$$
$$R_i^{\text{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$
$$(1 \leq i \leq 2)$$

Yannakakis+Pruning

$$\mathit{Answer}(\mathrm{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\mathrm{fadingkey}}(v)$$

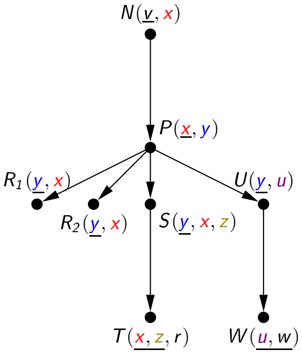
$$N^{\mathrm{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\mathrm{join}}(x)$$

$$P^{\mathrm{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\mathrm{fadingkey}}(x)$$

$$P^{\mathrm{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\mathrm{join}}(y)$$

$$P^{\mathrm{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\mathrm{join}}(x, y)$$

$$P^{\mathrm{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\mathrm{join}}(x, y)$$



$$U^{\mathrm{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\mathrm{fadingkey}}(y)$$

$$U^{\mathrm{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\mathrm{join}}(u)$$

$$S^{\mathrm{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\mathrm{fadingkey}}(y)$$

$$S^{\mathrm{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$

$$S^{\mathrm{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\mathrm{join}}(x, z)$$

$$T^{\mathrm{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$

$$W^{\mathrm{join}}(u) \leftarrow W(\underline{u, w})$$

$$R_i^{\mathrm{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\mathrm{fadingkey}}(y)$$

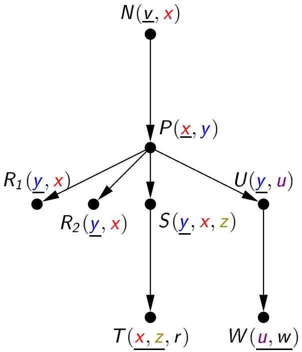
$$R_i^{\mathrm{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$

$$(1 \leq i \leq 2)$$

Yannakakis+Pruning

$$\mathit{Answer}(\text{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\text{fadingkey}}(v)$$

$$N^{\text{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\text{join}}(x)$$



$$P^{\text{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\text{fadingkey}}(x)$$

$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\text{join}}(y)$$

$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\text{join}}(x, y)$$

$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\text{join}}(x, y)$$

$$U^{\text{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\text{fadingkey}}(y)$$

$$U^{\text{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\text{join}}(u)$$

$$S^{\text{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\text{fadingkey}}(y)$$

$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$

$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\text{join}}(x, z)$$

$$T^{\text{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$

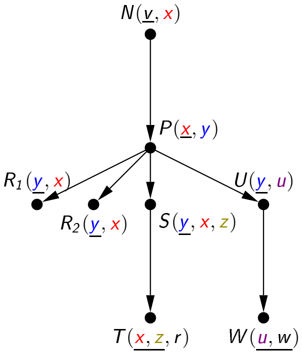
$$W^{\text{join}}(u) \leftarrow W(\underline{u, w})$$

$$R_i^{\text{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\text{fadingkey}}(y)$$

$$R_i^{\text{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$

$$(1 \leq i \leq 2)$$

Yannakakis+Pruning



$$T^{\text{join}}(x, z) \leftarrow T(\underline{x, z}, r)$$
$$W^{\text{join}}(u) \leftarrow W(\underline{u, w})$$

$$Answer(\text{yes}) \leftarrow N(\underline{v}, x) \wedge \neg N^{\text{fadingkey}}(v)$$
$$N^{\text{fadingkey}}(v) \leftarrow N(\underline{v}, x) \wedge \neg P^{\text{join}}(x)$$

$$P^{\text{join}}(x) \leftarrow P(\underline{x}, y) \wedge \neg P^{\text{fadingkey}}(x)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg U^{\text{join}}(y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg S^{\text{join}}(x, y)$$
$$P^{\text{fadingkey}}(x) \leftarrow P(\underline{x}, y) \wedge \neg R_i^{\text{join}}(x, y)$$

$$U^{\text{join}}(y) \leftarrow U(\underline{y}, u) \wedge \neg U^{\text{fadingkey}}(y)$$
$$U^{\text{fadingkey}}(y) \leftarrow U(\underline{y}, u) \wedge \neg W^{\text{join}}(u)$$

$$S^{\text{join}}(x, y) \leftarrow S(\underline{y}, x, z) \wedge \neg S^{\text{fadingkey}}(y)$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge S(\underline{y}, x', z) \wedge x \neq x'$$
$$S^{\text{fadingkey}}(y) \leftarrow S(\underline{y}, x, z) \wedge \neg T^{\text{join}}(x, z)$$

$$R_i^{\text{join}}(x, y) \leftarrow R_i(\underline{y}, x) \wedge \neg R_i^{\text{fadingkey}}(y)$$
$$R_i^{\text{fadingkey}}(y) \leftarrow R_i(\underline{y}, x) \wedge R_i(\underline{y}, x') \wedge x \neq x'$$
$$(1 \leq i \leq 2)$$

LinCQA

- LinCQA is a system that takes as input any query with a PPJT and outputs rewritings in both SQL and non-recursive Datalog with negation.
- `https://github.com/xiatingouyang/LinCQA/`
- See [Fan et al., 2023] for experiments.

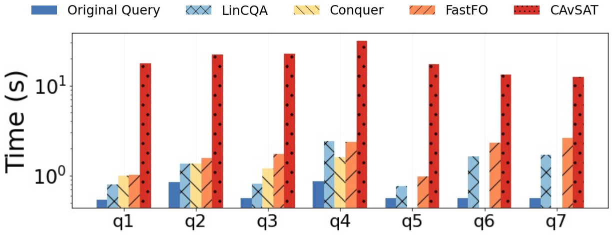


Table of Contents

Table of Contents

Range Consistent Query Answering [Arenas et al., 2001]

For queries returning numbers instead of Booleans.
For ease of presentation, all queries return a single number.

MOVIES

Title	Actor
Mr. & Mrs. Smith	Jolie
Mr. & Mrs. Smith	Pitt
Mr. & Mrs. Smith	Reeves

ACTORS

Name	Gender	Age
Jolie	F	48
Pitt	F	59
Pitt	M	60
Reeves	F	52
Reeves	F	53

Get the sum of ages of all actresses in Mr. & Mrs. Smith:

$$\mathtt{SUM}(z) \leftarrow \text{MOVIES}(\underline{\text{Mr. \& Mrs. Smith}, x}), \text{ACTORS}(\underline{x}, \text{F}, z).$$

- The lowest answer across all repairs is $48 + 52 = 100$;
- the greatest answer across all repairs is $48 + 59 + 53 = 160$;
- the interval $[100, 160]$ is called the range consistent answer.

Range Consistent Query Answering [Arenas et al., 2001]

For queries returning numbers instead of Booleans.
For ease of presentation, all queries return a single number.

MOVIES

<u>Title</u>	<u>Actor</u>
Mr. & Mrs. Smith	Jolie
Mr. & Mrs. Smith	Pitt
Mr. & Mrs. Smith	Reeves

ACTORS

<u>Name</u>	Gender	Age
Jolie	F	48
Pitt	F	59
Pitt	M	60
Reeves	F	52
Reeves	F	53

Get the sum of ages of all actresses in Mr. & Mrs. Smith:

$$\mathtt{SUM}(z) \leftarrow \text{MOVIES}(\underline{\text{Mr. \& Mrs. Smith}, x}), \text{ACTORS}(\underline{x}, \text{F}, z).$$

- The lowest answer across all repairs is $48 + 52 = 100$;
- the greatest answer across all repairs is $48 + 59 + 53 = 160$;
- the interval $[100, 160]$ is called the range consistent answer.

Range Consistent Query Answering [Arenas et al., 2001]

For queries returning numbers instead of Booleans.
For ease of presentation, all queries return a single number.

MOVIES

Title	Actor
Mr. & Mrs. Smith	Jolie
Mr. & Mrs. Smith	Pitt
Mr. & Mrs. Smith	Reeves

ACTORS

Name	Gender	Age
Jolie	F	48
Pitt	F	59
Pitt	M	60
Reeves	F	52
Reeves	F	53

Get the sum of ages of all actresses in Mr. & Mrs. Smith:

$$\texttt{SUM}(z) \leftarrow \mathsf{MOVIES}(\underline{\mathsf{Mr.\ \&\ Mrs.\ Smith}, x}), \mathsf{ACTORS}(\underline{x}, \mathsf{F}, z).$$

- The lowest answer across all repairs is $48 + 52 = 100$;
- the greatest answer across all repairs is $48 + 59 + 53 = 160$;
- the interval $[100, 160]$ is called the range consistent answer.

Formal Setting

- Numerical terms $f()$ expressible in the (safe) rule format

$$\texttt{AGG}(r) \leftarrow R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge R_2(\underline{\vec{x}_2}, \vec{y}_2) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n), \qquad (2)$$

  where r is either a numerical variable or a constant, and AGG is an aggregate operator (e.g., MAX, MIN, SUM, COUNT, AVG).
- Given a database instance, let $f^+()$ and $f^-()$ be, respectively, the greatest and smallest values of $f()$ across all repairs.
- Aggregate logic = first-order logic + aggregate operators.
- When can $f^+()$ and $f^-()$ be expressed in aggregate logic?
  - Not investigated since [Fuxman, 2007].
  - It is easily shown that $f^+()$ and $f^-()$ are not expressible in aggregate logic if the attack graph of the body of (2) has a cycle.
  - Does the converse hold?

Formal Setting

- Numerical terms $f()$ expressible in the (safe) rule format

$$\mathtt{AGG}(r) \leftarrow R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge R_2(\underline{\vec{x}_2}, \vec{y}_2) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n), \qquad (2)$$

  where r is either a numerical variable or a constant, and AGG is an aggregate operator (e.g., MAX, MIN, SUM, COUNT, AVG).
- Given a database instance, let $f^+()$ and $f^-()$ be, respectively, the greatest and smallest values of $f()$ across all repairs.
- Aggregate logic = first-order logic + aggregate operators.
- When can $f^+()$ and $f^-()$ be expressed in aggregate logic?
  - Not investigated since [Fuxman, 2007].
  - It is easily shown that $f^+()$ and $f^-()$ are not expressible in aggregate logic if the attack graph of the body of (2) has a cycle.
  - Does the converse hold?

Formal Setting

- Numerical terms $f()$ expressible in the (safe) rule format

$$\mathtt{AGG}(r) \leftarrow R_1(\underline{\vec{x}_1}, \vec{y}_1) \wedge R_2(\underline{\vec{x}_2}, \vec{y}_2) \wedge \cdots \wedge R_n(\underline{\vec{x}_n}, \vec{y}_n), \qquad (2)$$

  where r is either a numerical variable or a constant, and `AGG` is an aggregate operator (e.g., `MAX`, `MIN`, `SUM`, `COUNT`, `AVG`).
- Given a database instance, let $f^{+}()$ and $f^{-}()$ be, respectively, the greatest and smallest values of $f()$ across all repairs.
- Aggregate logic = first-order logic + aggregate operators.
- When can $f^{+}()$ and $f^{-}()$ be expressed in aggregate logic?
  - Not investigated since [Fuxman, 2007].
  - It is easily shown that $f^{+}()$ and $f^{-}()$ are not expressible in aggregate logic if the attack graph of the body of (2) has a cycle.
  - Does the converse hold?

Rewriting Example

$$\mathtt{SUM}(z) \leftarrow \mathsf{MOVIES}(\underline{\mathsf{Mr.\ \&\ Mrs.\ Smith}, x}), \mathsf{ACTORS}(\underline{x}, \mathsf{F}, z).$$

- Upper bound rewriting:

$$\mathsf{U}(x, \mathtt{MAX}(z)) \leftarrow \mathsf{MOVIES}(\underline{\mathsf{Mr.\ \&\ Mrs.\ Smith}, x}), \mathsf{ACTORS}(\underline{x}, \mathsf{F}, z)$$
$$\mathsf{UB}(\mathtt{SUM}(z)) \leftarrow \mathsf{U}(x, z)$$

- Lower bound rewriting:

$$\mathsf{POSSIBLE\_M}(x) \leftarrow \mathsf{ACTORS}(\underline{x}, \mathsf{M}, z)$$
$$\mathsf{CERTAIN\_F}(x, z) \leftarrow \mathsf{ACTORS}(\underline{x}, \mathsf{F}, z), \neg \mathsf{POSSIBLE\_M}(x)$$
$$\mathsf{L}(x, \mathtt{MIN}(z)) \leftarrow \mathsf{MOVIES}(\underline{\mathsf{Mr.\ \&\ Mrs.\ Smith}, x}), \mathsf{CERTAIN\_F}(\underline{x}, z)$$
$$\mathsf{LB}(\mathtt{SUM}(z)) \leftarrow \mathsf{L}(x, z)$$

Table of Contents

Counting

Given a Boolean query Q, define the following counting problem:

Problem $\sharp$CERTAINTY(Q)

Input: A database instance that may violate primary-key constraints.

Question: How many repairs of satisfy Q?

Complexity Classification Task

Input: A self-join-free Boolean conjunctive query Q.

Task: Determine lower and upper complexity bounds on the complexity of $\sharp$CERTAINTY(q), in terms of common complexity classes like FP and $\sharp$P.

- Solved in [Maslowski and W., 2013] and generalized to FDs in [Calautti et al., 2022].
- Same problem as query answering in block-independent disjoint (BID) probabilistic databases under the restriction that in every block $\mathbf{b}$, every tuple has probability $\frac{1}{|\mathbf{b}|}$.

Counting

Given a Boolean query Q, define the following counting problem:

Problem $\sharp$CERTAINTY(Q)

Input: A database instance that may violate primary-key constraints.

Question: How many repairs of satisfy Q?

Complexity Classification Task

Input: A self-join-free Boolean conjunctive query Q.

Task: Determine lower and upper complexity bounds on the complexity of $\sharp$CERTAINTY(q), in terms of common complexity classes like FP and $\sharp$P.

- Solved in [Maslowski and W., 2013] and generalized to FDs in [Calautti et al., 2022].
- Same problem as query answering in block-independent disjoint (BID) probabilistic databases under the restriction that in every block $\mathbf{b}$, every tuple has probability $\frac{1}{|\mathbf{b}|}$.

Counting

Given a Boolean query Q, define the following counting problem:

Problem $\sharp$CERTAINTY(Q)

Input: A database instance that may violate primary-key constraints.

Question: How many repairs of satisfy Q?

Complexity Classification Task

Input: A self-join-free Boolean conjunctive query Q.

Task: Determine lower and upper complexity bounds on the complexity of $\sharp$CERTAINTY(q), in terms of common complexity classes like FP and $\sharp$P.

- Solved in [Maslowski and W., 2013] and generalized to FDs in [Calautti et al., 2022].
- Same problem as query answering in block-independent disjoint (BID) probabilistic databases under the restriction that in every block $\mathbf{b}$, every tuple has probability $\frac{1}{|\mathbf{b}|}$.

BID Databases

Every input to CERTAINTY(Q) is a block-independent disjoint database without probabilities (or with uniform probabilities).

☞ Inconsistency is not only a burden, but also a chance. [1]

Researchers:

	Name	Affiliation	P
t_1^1	Fred	U. Washington	$p_1^1 = 0.3$
t_1^2		U. Wisconsin	$p_1^2 = 0.2$
t_1^3		Y! Research	$p_1^3 = 0.5$
t_2^1	Sue	U. Washington	$p_2^1 = 1.0$
t_3^1	John	U. Wisconsin	$p_3^1 = 0.7$
t_3^2		U. Washington	$p_3^2 = 0.3$
t_4^1	Frank	Y! Research	$p_4^1 = 0.9$
t_4^2		M. Research	$p_4^2 = 0.1$

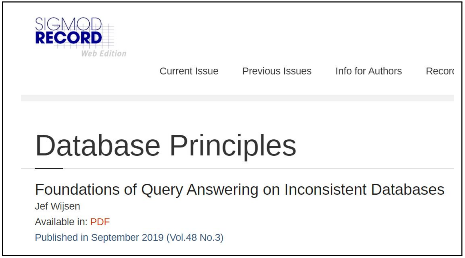
[1] Inspired by [Kern-Isberner and Lukasiewicz, 2017]. The image is from [Dalvi et al., 2009].

Table of Contents

Concluding Remarks

Consistent Query Answering is an active research area since [Arenas et al., 1999]:

- Database repairing w.r.t. different classes of constraints
- Database repairing and data exchange
- Database repairing and approximations
- Database repairing and preferences
- Database repairing and implementations
- Database repairing and database management systems
- Consistent query answering for queries with negation
- Consistent query answering in description logics
- Consistent query answering over graph databases
- ...



SIGMOD RECORD Web Edition

Current Issue Previous Issues Info for Authors Record

Database Principles

Foundations of Query Answering on Inconsistent Databases

Jef Wijsen

Available in: PDF

Published in September 2019 (Vol.48 No.3)

Thanks!

FYI, Brad Pitt celebrated his 60th birthday on December 18, 2023.

References I

Arenas, M., Bertossi, L. E., and Chomicki, J. (1999).
Consistent query answers in inconsistent databases.
In *PODS*, pages 68–79. ACM Press.

Arenas, M., Bertossi, L. E., and Chomicki, J. (2001).
Scalar aggregation in fd-inconsistent databases.
In *ICDT*, volume 1973 of *Lecture Notes in Computer Science*, pages 39–53. Springer.

Calautti, M., Livshits, E., Pieris, A., and Schneider, M. (2022).
Counting database repairs entailing a query: The case of functional dependencies.
In *PODS*, pages 403–412. ACM.

Chomicki, J. and Marcinkowski, J. (2005).
Minimal-change integrity maintenance using tuple deletions.
Inf. Comput., 197(1-2):90–121.

Dalvi, N. N., Ré, C., and Suciu, D. (2009).
Probabilistic databases: diamonds in the dirt.
Commun. ACM, 52(7):86–94.

Fan, Z., Koutris, P., Ouyang, X., and Wijsen, J. (2023).
LinCQA: Faster consistent query answering with linear time guarantees.
Proc. ACM Manag. Data, 1(1):38:1–38:25.

Fontaine, G. (2015).
Why is it hard to obtain a dichotomy for consistent query answering?
ACM Trans. Comput. Log., 16(1):7:1–7:24.

Fuxman, A. (2007).
Efficient query processing over inconsistent databases.
PhD thesis, University of Toronto.

References II

Kern-Isberner, G. and Lukasiewicz, T. (2017).
Many facets of reasoning under uncertainty, inconsistency, vagueness, and preferences: A brief survey.
KI, 31(1):9–13.

Koutris, P. and W., J. (2017).
Consistent query answering for self-join-free conjunctive queries under primary key constraints.
ACM Trans. Database Syst., 42(2):9:1–9:45.

Maslowski, D. and W., J. (2013).
A dichotomy in the complexity of counting database repairs.
J. Comput. Syst. Sci., 79(6):958–983.

Padmanabha, A., Segoufin, L., and Sirangelo, C. (2023).
A dichotomy in the complexity of consistent query answering for two atom queries with self-join.
CoRR, abs/2309.12059.

W., J. (2010).
A remark on the complexity of consistent conjunctive query answering under primary key violations.
Inf. Process. Lett., 110(21):950–955.