

On the Complexity of Repair Checking in Inconsistent Databases

Foto Afrati¹ Phokion G. Kolaitis²

¹National Technical University of Athens

²IBM Almaden Research Center

LID 2008
Rome, May 2008

Coping with Inconsistent Databases

- **Inconsistent databases** arise in a variety of contexts and for different reasons:
 - in data warehousing of heterogeneous data obeying different integrity constraints.
 - for lack of support of particular integrity constraints;
 - ...

Coping with Inconsistent Databases

- **Inconsistent databases** arise in a variety of contexts and for different reasons:
 - in data warehousing of heterogeneous data obeying different integrity constraints.
 - for lack of support of particular integrity constraints;
 - ...
- **Database repairs** provide a framework for coping with inconsistent databases in a principled way.

Database Repairs

Definition (Arenas, Bertossi, Chomicki – 1999)

Σ a set of integrity constraints and r an inconsistent database.
A database r' is a *repair* of r w.r.t. Σ if

- r' is a consistent database (i.e., $r' \models \Sigma$);
- r' differs from r in a **minimal** way.

Database Repairs

Definition (Arenas, Bertossi, Chomicki – 1999)

Σ a set of integrity constraints and r an inconsistent database.
A database r' is a *repair* of r w.r.t. Σ if

- r' is a consistent database (i.e., $r' \models \Sigma$);
- r' differs from r in a **minimal** way.

Fact

Several different types of repairs have been considered:

- Subset-repairs;
- \oplus -repairs (symmetric-difference-repairs);
- Cardinality-based repairs;
- Attribute-based repairs.

Algorithmic Problems about Inconsistent Databases

- **The Consistent Query Answering Problem:**
 - Compute the certain answers of a query q on an inconsistent database r , where the set of possible worlds is the set of repairs of r .

Algorithmic Problems about Inconsistent Databases

- **The Consistent Query Answering Problem:**
 - Compute the certain answers of a query q on an inconsistent database r , where the set of possible worlds is the set of repairs of r .
 - Consistent query answering has been investigated in depth.

Algorithmic Problems about Inconsistent Databases

- **The Consistent Query Answering Problem:**
 - Compute the certain answers of a query q on an inconsistent database r , where the set of possible worlds is the set of repairs of r .
 - Consistent query answering has been investigated in depth.
- **The Repair Checking Problem:**
 - Given r and r' , tell whether or not r' is a repair of r .

Algorithmic Problems about Inconsistent Databases

- **The Consistent Query Answering Problem:**

- Compute the certain answers of a query q on an inconsistent database r , where the set of possible worlds is the set of repairs of r .
- Consistent query answering has been investigated in depth.

- **The Repair Checking Problem:**

- Given r and r' , tell whether or not r' is a repair of r .
- Repair checking is a data cleaning problem that underlies consistent query answering
- So far, repair checking has received **less** attention than consistent query answering.

Aim of this Work

Embark on a systematic investigation of the algorithmic aspects of the repair checking problem

- Study classes of integrity constraints that have been considered in information integration and data exchange.
- Study subset-repairs and \oplus -repairs.
- Introduce and study *CC-repairs* (*component-cardinality repairs*), a new type of cardinality-based repairs that have a Pareto-optimality character.

Types of Constraints

Definition

- *Equality-generating dependency (egd)*: $\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow x_i = x_j)$, where $\phi(\mathbf{x})$ is a conjunction of atoms.
- *Denial constraint*: $\forall \mathbf{x} \neg(\alpha(\mathbf{x}) \wedge \beta(\mathbf{x}))$, where $\alpha(\mathbf{x})$ is a non-empty conjunction of atoms and $\beta(\mathbf{x})$ is a conjunction of comparison atoms $x_i = x_j$, $x_i \neq x_j$, $x_i < x_j$, $x_i \leq x_j$.

Example

- Every functional dependency is an egd, but **not** vice versa:
 $\forall x, y, z(\text{MOTHER}(z, x) \wedge \text{MOTHER}(w, x) \rightarrow z = w)$.
- Every egd is (logically equivalent) to a denial constraint, but **not** vice versa:
 $\forall x, y \neg(\text{MOTHER}(x, y) \wedge x = y)$

Types of Constraints

Definition

- *Tuple-generating dependency (tgd)*:

$$\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \exists \mathbf{y}\psi(\mathbf{x}, \mathbf{y})),$$

where $\phi(\mathbf{x})$ is a conjunction of atoms with vars. in \mathbf{x} , and $\psi(\mathbf{x}, \mathbf{y})$ is a conjunction of atoms with vars. in \mathbf{x} and \mathbf{y} .

- *Full tgd*: a tgd with no existential quantifiers in rhs.

$$\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \psi(\mathbf{x})),$$

where $\phi(\mathbf{x})$ and $\psi(\mathbf{x})$ are conjunctions of atoms.

- *LAV (local-as-view) tgd*: a tgd in which lhs is a single atom.

$$\forall \mathbf{x}(P(\mathbf{x}) \rightarrow \exists \mathbf{y}\psi(\mathbf{x}, \mathbf{y})).$$

Note: Every inclusion dependency is a LAV tgd, but **not** vice versa.

Types of Constraints

Example

(dropping universal quantifiers)

- The following is a tgd

$$(\text{MOTHER}(z, x) \wedge \text{MOTHER}(z, y) \rightarrow \exists u(\text{FATHER}(u, x) \wedge \text{FATHER}(u, y))) \quad .$$

- The following are full tgds:

$$(\text{SIBLING}(x, y) \rightarrow \text{SIBLING}(y, x))$$

$$(\text{MOTHER}(z, x) \wedge \text{MOTHER}(z, y) \rightarrow \text{SIBLING}(x, y))$$

- The following is a LAV tgd:

$$(\text{SIBLING}(x, y) \rightarrow \exists z(\text{MOTHER}(z, x) \wedge \text{MOTHER}(z, y)))$$

Types of Repairs

Definition

Σ a set of integrity constraints and r an inconsistent database.

- r' is a *subset-repair* of r w.r.t. Σ if $r' \subset r$, $r' \models \Sigma$, and there is **no** r'' such that $r' \subset r'' \subset r$ and $r'' \models \Sigma$.
- r' is a \oplus -*repair* of r w.r.t. Σ if $r' \models \Sigma$ and there is **no** r'' such that $r \oplus r'' \subset r \oplus r'$ and $r'' \models \Sigma$.

Fact

- If $r' \subset r$, then r' is a subset-repair of r if and only if r' is a \oplus -repair of r .
- If Σ is a set of denial constraints, then every \oplus -repair is also a subset-repair.

Earlier Work - Tractability Results

Theorem

- folklore

If Σ is a set of denial constraints, then the \oplus -repair checking problem w.r.t. Σ is in LOGSPACE.

Earlier Work - Tractability Results

Theorem

- folklore

If Σ is a set of denial constraints, then the \oplus -repair checking problem w.r.t. Σ is in LOGSPACE.

- Chomicki and Marcinkowski – 2005

If Σ is the union of an acyclic set of inclusion dependencies and a set of functional dependencies, then the subset-repair checking problem w.r.t. Σ is in PTIME; in fact, it is in LOGSPACE.

Earlier Work - Tractability Results

Theorem

- folklore

If Σ is a set of denial constraints, then the \oplus -repair checking problem w.r.t. Σ is in LOGSPACE.

- Chomicki and Marcinkowski – 2005

If Σ is the union of an acyclic set of inclusion dependencies and a set of functional dependencies, then the subset-repair checking problem w.r.t. Σ is in PTIME; in fact, it is in LOGSPACE.

- Staworko – 2007

If Σ is a set of full tgds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME.

Weakly Acyclic Sets of Tgds

Fact

- Acyclic sets of inclusion dependencies and set of full tgds are special cases of **weakly acyclic sets of tgds**.
- Weakly acyclic sets of inclusion dependencies are known to have good algorithmic behavior in data exchange and data integration.

Definition

- The *position graph* of a set Σ of tgds:
 - The nodes are the pairs (R, A) , where R is a relation symbol and A is an attribute of R . Such a pair (R, A) is called a *position*.
 - Let $\phi(\mathbf{x}) \rightarrow \exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y})$ be a tgd in Σ and let x in \mathbf{x} be a variable that also occurs in $\psi(\mathbf{x}, \mathbf{y})$. For every occurrence of x in $\phi(\mathbf{x})$ in position (R, A_i) , add the following edges:
 - (i) For every occurrence of x in $\psi(\mathbf{x}, \mathbf{y})$ in position (S, B_j) , add an edge $(R, A_i) \rightarrow (S, B_j)$;
 - (ii) In addition, for every existentially quantified variable y in \mathbf{y} and for every occurrence of y in $\psi(\mathbf{x}, \mathbf{y})$ in position (T, C_k) , add a *special edge* $(R, A_i) \xrightarrow{*} (T, C_k)$.
- Σ is *weakly acyclic* if the position graph has **no** cycle going through a special edge.
- A tgd θ is *weakly acyclic* if $\{\theta\}$ is weakly acyclic.

Weakly Acyclic Sets of Tgds

Fact

- Every acyclic set of inclusion dependencies is a weakly acyclic set (the position graph is acyclic)
- Every set of full tgds is weakly acyclic (the position graph has no special edges).

Example

$$\Sigma = \{D(e, m) \rightarrow M(m), M(m) \rightarrow \exists e D(e, m)\}$$

is a weakly acyclic, but cyclic, set of inclusion dependencies.

Position graph:

$$D.1 \xleftarrow{*} M.1 \rightleftarrows D.2$$

Weakly Acyclic Sets of Tgds

Fact

Weakly acyclic sets of tgds have good algorithmic behavior in data exchange and data integration. Specifically, there are PTIME algorithms for:

- Computing the certain answers of conjunctive queries;
- Computing a canonical universal solution;
- Computing the core of the universal solutions.

Weakly Acyclic Sets of Tgds

Fact

Weakly acyclic sets of tgds have good algorithmic behavior in data exchange and data integration. Specifically, there are PTIME algorithms for:

- Computing the certain answers of conjunctive queries;
- Computing a canonical universal solution;
- Computing the core of the universal solutions.

Problem

Does the good algorithmic behavior of weakly acyclic sets of tgds extend to the repair checking problem?

Weakly Acyclic Sets of Tgds: Intractability

Theorem

There is a weakly acyclic set Σ of tgds such that the subset-repair checking problem w.r.t. Σ is coNP-complete.

Weakly Acyclic Sets of Tgds: Intractability

Theorem

There is a weakly acyclic set Σ of tgds such that the subset-repair checking problem w.r.t. Σ is coNP-complete.

Proof.

coNP-hardness via a reduction from POSITIVE 1-IN-3-SAT

- Σ consists of the (non-LAV) tgd

$$A(w) \wedge P(x, y, z) \rightarrow$$

$$\exists u_1, u_2, u_3 (T(x, u_1) \wedge T(y, u_2) \wedge T(z, u_3) \wedge S(u_1, u_2, u_3))$$

and the two full tgds:

$$T(x, u) \wedge T(x, u') \wedge D(u, u') \rightarrow S(u, u, u), \quad T(x, u) \rightarrow A(u).$$

- Σ is weakly acyclic: all special edges are from pos. of P to pos. of T and S ; no position of P has an incoming edge.



Weakly Acyclic Sets of Tgds: Intractability

Theorem

There is a weakly acyclic set Σ of tgds such that the subset-repair checking problem w.r.t. Σ is coNP-complete.

Weakly Acyclic Sets of Tgds: Intractability

Theorem

There is a weakly acyclic set Σ of tgds such that the subset-repair checking problem w.r.t. Σ is coNP-complete.

Theorem (Chomicki and Marcinkowski - 2005)

There is a set Σ consisting of one inclusion dependency and one functional dependency such that the subset-repair checking problem w.r.t. Σ is coNP-complete.

Note: The inclusion dependency is

$$R(x_1, x_2, x_3, x_4) \rightarrow \exists y_1, y_2 y_3 R(y_1, y_2, x_4, y_3),$$

which is **not** weakly acyclic (**special** self-loop on R .4).

Weakly Acyclic Sets of LAV Tgds: Tractability

Theorem

If Σ is the union of a weakly acyclic set of LAV tgds and a set of egds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME; in fact, it is in NC.

Weakly Acyclic Sets of LAV Tgds: Tractability

Theorem

If Σ is the union of a weakly acyclic set of LAV tgds and a set of egds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME; in fact, it is in NC.

Proof Idea.

- **Key property of LAV tgds:** only single facts *fire* a tgd (and no combinations of facts).
- We design a PTIME algorithm that:
Exploits weak acyclicity to compute facts that do not *transitively fire* non-full tgds in Σ ;
- **Note:** The algorithm in the proceedings works for tgds in which every variable in the lhs occurs also in the rhs.

Full Tgds: PTIME-completeness

Theorem (Staworko – 2007)

If Σ is a set of full tgds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME.

Full Tgds: PTIME-completeness

Theorem (Staworko – 2007)

If Σ is a set of full tgds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME.

Theorem

There is a set Σ of full tgds such that the subset-repair problem w.r.t. Σ is PTIME-complete.

Full Tgds: PTIME-completeness

Theorem (Staworko – 2007)

If Σ is a set of full tgds, then the \oplus -repair checking problem w.r.t. Σ is in PTIME.

Theorem

There is a set Σ of full tgds such that the subset-repair problem w.r.t. Σ is PTIME-complete.

Proof (Hint).

- Logspace Reduction from HORN 3-SAT.
- Use full tgds to encode the **unit propagation algorithm** for HORN 3-SAT.



Complexity of Subset- and \oplus -Repair Checking

Constraints \ Semantics	Subset-repair	\oplus-repair
Denial	LOGSPACE	LOGSPACE
Acyc. set of IND & egds	LOGSPACE	LOGSPACE
Weak. acyc. LAV tgds & egds	NC	NC
Full tgds & egds	PTIME-comp.	PTIME-comp.
IND & egds	coNP-comp.	coNP-comp.
Weak. acyc. tgds & egds	coNP-comp.	coNP-comp.

Complexity of Subset- and \oplus -Repair Checking

Constraints \ Semantics	Subset-repair	\oplus -repair
Denial	LOGSPACE	LOGSPACE
Acyc. set of IND & egds	LOGSPACE	LOGSPACE
Weak. acyc. LAV tgds & egds	NC	NC
Full tgds & egds	PTIME-comp.	PTIME-comp.
IND & egds	coNP-comp.	coNP-comp.
Weak. acyc. tgds & egds	coNP-comp.	coNP-comp.

Note

New phenomenon:

Good algorithmic behavior of acyclic sets of inclusion dependencies and sets of full tgds for \oplus -repair checking does **not** extend to arbitrary weakly acyclic sets of tgds.

C-Repairs: Cardinality Repairs

Definition

Σ a set of integrity constraints and r an inconsistent database. r' is a *C-repair (cardinality-repair)* of r w.r./ Σ if $r'' \models \Sigma$ and there is **no** r'' such that $r'' \models \Sigma$ and $|r \oplus r''| < |r \oplus r'|$.

C-Repairs: Cardinality Repairs

Definition

Σ a set of integrity constraints and r an inconsistent database. r' is a *C-repair (cardinality-repair)* of r w.r.t. Σ if $r'' \models \Sigma$ and there is **no** r'' such that $r'' \models \Sigma$ and $|r \oplus r''| < |r \oplus r'|$.

Theorem (Lopatenko and Bertossi – 2007)

There is a denial constraint φ such that the C-repair checking problem w.r.t. φ is coNP-complete.

CC-Repairs: Component Cardinality Repairs

Definition

- $|r| \leq_{cc} |r'|$ if for every relation symbol P in the schema, we have that $|P^r| \leq |P^{r'}|$.
- $|r| <_{cc} |r'|$ if $|r| \leq_{cc} |r'|$ and there is at least one relation symbol P in the schema such that $|P^r| < |P^{r'}|$.
- r' is a *CC-repair (component-cardinality repair)* of r w.r.t. Σ if $r' \models \Sigma$ and there is **no** r'' such that $r'' \models \Sigma$ and $|r \oplus r''| <_{cc} |r \oplus r'|$.

CC-Repairs: Component Cardinality Repairs

Definition

- $|r| \leq_{cc} |r'|$ if for every relation symbol P in the schema, we have that $|P^r| \leq |P^{r'}|$.
- $|r| <_{cc} |r'|$ if $|r| \leq_{cc} |r'|$ and there is at least one relation symbol P in the schema such that $|P^r| < |P^{r'}|$.
- r' is a *CC-repair (component-cardinality repair)* of r w.r.t. Σ if $r' \models \Sigma$ and there is **no** r'' such that $r'' \models \Sigma$ and $|r \oplus r''| <_{cc} |r \oplus r'|$.

Fact

- Every C-repair is a CC-repair.
- Every CC-repair is a \oplus -repair.

Example

Let Σ be the set consisting of the following four tgds:

$$\begin{aligned} P(x) &\rightarrow R(x), & P'(x) &\rightarrow R'(x), \\ R(x) &\rightarrow R'(x), & P'(x) &\rightarrow Q'(x). \end{aligned}$$

- Inconsistent $r = \{P(1), P'(1)\}$; consistent r_1, r_2, r_3 :

$$r_1 = \emptyset; \quad 2; \quad (1, 1, 0, 0, 0)$$

$$r_2 = \{P'(1), R'(1), Q'(1)\}; \quad 3; \quad (1, 0, 0, 1, 1)$$

$$r_3 = \{P(1), R(1), R'(1), \}; \quad 3; \quad (0, 1, 1, 1, 0)$$

characteristic sequence under the order (P, P', R, R', Q')

- r_1, r_2, r_3 are CC-repairs.
- r_1 is the only C-repair among them.

CC-Repairs: Intractability

Theorem

- There is denial constraint θ such that the CC-repair checking problem w.r.t. χ is coNP-complete.
- There is a full tgd φ such that the CC-repair problem w.r.t. θ is coNP-complete.
- There is a LAV acyclic tgd ψ such that the CC-repair checking problem w.r.t. ψ is coNP-complete.
- There is an acyclic set Ψ of inclusion dependencies such that the CC-repair problem w.r.t. Ψ is coNP-complete.

CC-Repairs: Intractability

Theorem

There is an acyclic set Ψ of inclusion dependencies such that the CC-repair problem w.r.t. Ψ is coNP-complete.

Proof.

- coNP-hardness via a reduction from POSITIVE 1-IN-3-SAT
- Ψ is the following acyclic set of inclusion dependencies:

$$P(x, y, z) \rightarrow \exists u, v, w Q(x, y, z, u, v, w)$$

$$Q(x, y, z, u, v, w) \rightarrow S(u, v, w)$$

$$Q(x, y, z, u, v, w) \rightarrow T(x, u)$$

$$Q(x, y, z, u, v, w) \rightarrow T(y, v)$$

$$Q(x, y, z, u, v, w) \rightarrow T(z, w).$$



Synopsis

- \oplus -repair checking is in PTIME for weakly acyclic sets of LAV tgds.
- \oplus -repair checking can be coNP-complete for weakly acyclic sets of tgds.
- CC-repair checking can be coNP-complete for denial constraints, full tgds, and acyclic sets of inclusion dependencies.

Synopsis

- \oplus -repair checking is in PTIME for weakly acyclic sets of LAV tgds.
- \oplus -repair checking can be coNP-complete for weakly acyclic sets of tgds.
- CC-repair checking can be coNP-complete for denial constraints, full tgds, and acyclic sets of inclusion dependencies.
- **Open Problem:** Prove or disprove that a *dichotomy theorem* holds for the complexity of the \oplus -repair checking problem w.r.t. sets of tgds and egds.