Repair Checking in Inconsistent Databases: Algorithms and Complexity

Foto Afrati[1]   Phokion G. Kolaitis[2]

[1]National Technical University of Athens

[2]UC Santa Cruz and IBM Almaden Research Center

Oxford, January 2009

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.
  - In ETL applications, where data has to be “cleansed" before it can be processed.
  - 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.
  - In ETL applications, where data has to be “cleansed" before it can be processed.
  - For lack of support of particular integrity constraints.
  - ...
- Database repairs provide a framework for coping with inconsistent databases in a principled way and without “cleansing" dirty data first.

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.

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 $r'' \models \Sigma$.

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 functional dependencies, then every $\oplus$-repair is also a subset-repair.

Example

Relation schema R, instance $r = \{R(a,b), R(a,c), R(b,c)\}$

- $\Sigma = \{R(x,y) \wedge R(x,z) \rightarrow y = z\}$
  r has two $\oplus$-repairs (and subset repairs) w.r.t. Σ:
  - $r_1 = \{R(a,b), R(b,c)\}$
    and
  - $r_2 = \{R(a,c), R(b,c)\}$.

Example

Relation schema R, instance $r = \{R(a,b), R(a,c), R(b,c)\}$

- $\Sigma = \{R(x,y) \wedge R(x,z) \rightarrow y = z\}$
  r has two $\oplus$-repairs (and subset repairs) w.r.t. Σ:
  - $r_1 = \{R(a,b), R(b,c)\}$
    and
  - $r_2 = \{R(a,c), R(b,c)\}$.
- $\Sigma' = \{R(x,y) \rightarrow R(y,x)\}$
  r has eight $\oplus$-repairs w.r.t. Σ':
  - $r_1 = \emptyset$
  - $r_2 = \{R(a,b), R(b,a)\}$
  - $r_3 = \{R(a,b), R(b,a), R(a,c), R(c,a)\}$
  - $\ldots$

Possible Worlds and Certain Answers

Definition

Suppose that with every instance r over some schema **S**, we have associated a set $\mathcal{W}(r)$ of instances over some other (possibly different) schema **T** (the set of *possible worlds* of r).

If q is a query over **T**, then the *certain answers of q on r w.r.t. $\mathcal{W}(r)$* is

$$\text{certain}(q, r, \mathcal{W}(r)) = \bigcap\{q(r') : r' \in \mathcal{W}(r)\}.$$

Note

The certain answers is the standard semantics of queries in the context of *incomplete information*.

Repairs and Consistent Answers

Definition (Arenas, Bertossi, Chomicki)

Fix a particular type of repairs (say, subset repairs or $\oplus$-repairs)
Let Σ be a set of integrity constraints, let q be a query, and let r be an instance.
The *consistent answers of q on r w.r.t. Σ*, denoted by $\mathrm{cons}_\Sigma(q,r)$, is the set $\mathrm{certain}(q,r,\mathcal{W}(r))$, where $\mathcal{W}(r)$ is the set of all repairs of r, i.e.,

$$\mathrm{cons}_\Sigma(q,r) = \bigcap\{q(r') : r' \text{ is a repair of } r\}.$$

Example (Revisited)

Relation schema R, instance $r = \{R(a,b), R(a,c), R(b,c)\}$
$\Sigma = \{R(x,y) \wedge R(x,z) \rightarrow y = z\}$
Recall that r has two $\oplus$-repairs (and subset repairs) w.r.t. Σ:

- $r_1 = \{R(a,b), R(b,c)\}$
  and
- $r_2 = \{R(a,c), R(b,c)\}$.

Then

- If $q(x)$ is the query $\exists z R(x,z)$, then
$$\mathrm{cons}_\Sigma(q,r) = \{a,b\}.$$
- If $q'(x)$ is the query $\exists z R(z,x)$, then
$$\mathrm{cons}_\Sigma(q',r) = \{c\}.$$

Data Complexity of Consistent Query Answering

Theorem (Chomicki and Marcinkowski - 2003)

There exist a set Σ of two functional dependencies (in fact, key constraints) and a Boolean conjunctive query q such that the following problem is coNP-complete:
Given an instance r, is $\mathrm{cons}_{\Sigma}(q, r)$ true?

Data Complexity of Consistent Query Answering

Theorem (Chomicki and Marcinkowski - 2003)

There exist a set Σ of two functional dependencies (in fact, key constraints) and a Boolean conjunctive query q such that the following problem is coNP-complete:
Given an instance r, is $\mathrm{cons}_\Sigma(q, r)$ true?

Theorem (Staworko - 2007)

There exist a set Σ consisting of one functional dependency and two universal constraints, and a Boolean conjunctive query q such that the following problem is Π_2^p-complete:
Given an instance r, is $\mathrm{cons}_\Sigma(q, r)$ true?

Data Complexity of Consistent Query Answering

Extensive study over the past decade for various classes of integrity constraints and for different types of repairs.

- Intractability results (coNP-hardness, Π_2^p-hardness)
- Tractability results for restricted classes of conjunctive queries:
  - Polynomial-time algorithms.
  - Rewriting to first-order queries.
- Prototype systems for consistent query answering:
  - Hippo (Chomicki et al.)
  - ConQuer (Fuxman)

Note

For overviews, see the invited paper by J. Chomicki in ICDT 2007 and the Ph.D. theses of A. Fuxman (2007) and S. Staworko (2007).

Algorithmic Problems about Inconsistent Databases

- **The Consistent Query Answering Problem:**
  - 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.

Repair Checking vs. Consistent Query Answering

Proposition (Chomicki and Marcinkowski - 2003)

Let Σ be a set of integrity constraints containing all inclusion dependencies. There is a Boolean query q such that the $\oplus$-repair checking problem w.r.t. Σ has a logspace-reduction to the complement of the consistent query answering problem for q w.r.t. Σ

Note

Thus, in many cases, lower bounds for the complexity of the $\oplus$-repair checking problem yield lower bounds for the complexity of the consistent query answering problem.

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

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

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

- The following are full tgds:

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

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

- The following is a LAV tgd:

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

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 subset-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 subset-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 subset-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 and egds, then the subset-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:

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

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

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

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:
  $T(x,u) \wedge T(x,u') \wedge D(u,u') \rightarrow S(u,u,u),\ T(x,u) \rightarrow A(u).$
- 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 subset-repair checking problem w.r.t. Σ is in PTIME; in fact, it is in LOGSPACE.

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 subset-repair checking problem w.r.t. Σ is in PTIME; in fact, it is in LOGSPACE.

Proof Idea.

- Key property of LAV tgds: only single facts *fire* a tgd (and no combinations of facts). Hence, LAV tgds are preserved under unions of models.
- Key property of weakly acyclic sets of tgds: The solution aware chase terminates in polynomial time.

Note: The solution aware chase was used in the study of peer data exchange (Fuxman, K ..., Miller, Tan - 2005). □

Weakly Acyclic Sets of LAV Tgds: Tractability

Lemma

Let Σ be the union of a weakly acyclic set of LAV tgds and a set of egds. Then there is a constant c such that the following holds.
Let r, r' be two instances such that $r' \models \Sigma$, and let t be a fact in $r \setminus r'$ such that there is a non-empty set $A \subset r \setminus r'$ such that $t \in A$ and $r' \cup A \models \Sigma$. There there is a set A_t of facts such that

- $t \in A_t$
- $A_t \subseteq r \setminus r'$
- $|A_t| \leq c$
- $r' \cup A_t \models \Sigma$.

Weakly Acyclic Sets of LAV Tgds: Tractability

Algorithm for subset-repair checking w.r.t. a set Σ that is the union of a weakly acyclic set of LAV tgds and a set of egds.

Given r and r' with $r' \subset r$, $r \not\models \Sigma$, $r' \models \Sigma$:
Test whether there is a set A^* such that

1. A^* is non-empty
2. $|A^*| \leq c$
3. $A^* \subseteq r \setminus r'$
4. $r' \cup A^* \models \Sigma$.

- If such a set A^* exists, then r' is not a subset repair of r w.r.t. Σ;
- Otherwise, r' is a subset repair of r w.r.t. Σ.

Subset Repairs vs. $\oplus$-Repairs

Theorem

If Σ is a weakly acyclic set of LAV tgds, then the $\oplus$-repair problem w.r.t. Σ is in PTIME; in fact, it is in LOGSPACE.

Theorem

There is a weakly acyclic set Σ_1 of LAV tgds and a set Σ_2 of egds such that the $\oplus$-repair checking problem w.r.t. $\Sigma_1 \cup \Sigma_2$ is coNP-complete.

Theorem

There is a weakly acyclic set Σ_1 of LAV tgds and a set Σ_2 of egds such that the $\oplus$-repair checking problem w.r.t. $\Sigma_1 \cup \Sigma_2$ is coNP-complete.

Proof.

Reduction from POSITIVE 1-IN-3-SAT

$$P_{23}(w,w,x,y,z) \rightarrow$$
$$\exists u,v,w(T(x,u,x,y,z) \wedge T(y,v,x,y,z) \wedge T(z,w,x,y,z) \wedge S(u,v,w))$$
$$T(x,u,x',y',z') \wedge T(x,u',x'',y'',z'') \rightarrow u = u'$$
$$P_{23}(s,s,x,y,z) \wedge P_{23}(w,w',x',y',z') \rightarrow w = w'$$
$$P_1(x,y,z) \rightarrow \exists w P_{23}(w,w',x,y,z)$$
$$T(x',u,x,y,z) \rightarrow P_{23}(w,w,x,y,z).$$

$\square$

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	?
Weak. acyc. LAV tgds	LOGSPACE	LOGSPACE
Weak. acyc. LAV tgds & egds	LOGSPACE	coNP-comp.
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 ⊕-Repair Checking

Constraints \ Semantics	**Subset-repair**	**⊕-repair**
Denial	LOGSPACE	LOGSPACE
Acyc. set of IND & egds	LOGSPACE	?
Weak. acyc. LAV tgds	LOGSPACE	LOGSPACE
Weak. acyc. LAV tgds & egds	LOGSPACE	coNP-comp.
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 subset-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.t. 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:

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

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

$$\begin{array}{rrr} r_1 = \emptyset; & 2; & (1,1,0,0,0) \\ r_2 = \{P'(1), R'(1), Q'(1)\}; & 3; & (1,0,0,1,1) \\ r_3 = \{P(1), R(1), R'(1), \}; & 3; & (0,1,1,1,0) \end{array}$$

  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:

$$
\begin{aligned}
&P(x,y,z) \rightarrow \exists u,v,w Q(x,y,z,u,v,w)\\
&Q(x,y,z,u,v,w) \rightarrow S(u,v,w)\\
&Q(x,y,z,u,v,w) \rightarrow T(x,u)\\
&Q(x,y,z,u,v,w) \rightarrow T(y,v)\\
&Q(x,y,z,u,v,w) \rightarrow T(z,w).
\end{aligned}
$$

$\square$

Synopsis

- Subset repair checking is in PTIME for weakly acyclic sets of LAV tgds and egds.
- $\oplus$-repair checking is in PTIME for weakly acyclic sets of LAV tgds.
- $\oplus$-repair checking can be coNP-complete for weakly acyclic sets of LAV tgds and egds.
- $\oplus$-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.

Directions and Problems

- Open Problem: Prove or disprove that a *dichotomy theorem holds for the complexity of the ⊕-repair checking problem w.r.t. sets of tgds and egds.*
- Investigate the complexity of repair checking for other types of repairs (attribute-based repairs).
  Work in this direction has already been carried out by J. Wisden.
- Are there criteria for differentiating between repairs of the same type?