Project-Join-Repair

An Approach to Consistent Query Answering Under Functional Dependencies

Jef Wijsen

Université de Mons-Hainaut, Belgium

Preliminary I: Repairing

- Let Σ be a set of fd's and I an inconsistent relation. A repair of I is a maximal (under set inclusion) consistent subset of I.

I	Name	Birth	Sex	ZIP	City
	An	1964	F	7000	Mons
	Ed	1962	M	7000	Bergen

$\mathbf{Name} \rightarrow \mathbf{Birth}$

$\mathbf{Name} \rightarrow \mathbf{Sex}$

$\mathbf{Name} \rightarrow \mathbf{ZIP}$

$\mathbf{ZIP} \rightarrow \mathbf{City}$

Preliminary I: Repairing

- Let Σ be a set of fd's and I an inconsistent relation. A repair of I is a maximal (under set inclusion) consistent subset of I.

I

Name	**Birth**	**Sex**	**ZIP**	**City**
An	1964	F	7000	Mons
Ed	1962	M	7000	Bergen

$\mathbf{Name} \rightarrow \mathbf{Birth}$

$\mathbf{Name} \rightarrow \mathbf{Sex}$

$\mathbf{Name} \rightarrow \mathbf{ZIP}$

$\mathbf{ZIP} \rightarrow \mathbf{City}$

- R_1

Name	**Birth**	**Sex**	**ZIP**	**City**
An	1964	F	7000	Mons

$\rightsquigarrow$ Ed is left out...

- R_2

Name	**Birth**	**Sex**	**ZIP**	**City**
Ed	1962	M	7000	Bergen

$\rightsquigarrow$ An is left out...

Prelim. II: Consistent Query Answering

- Let $\{R_1, R_2, \ldots, R_n\}$ be the set of all repairs of I w.r.t. Σ. The consistent answer to a query q is defined by:

$$q_\Sigma(I) := \bigcap_{i=1}^{n} q(R_i)$$

- For example, "`SELECT Name FROM I`" yields $\{\}$.

Prelim. II: Consistent Query Answering

- Let $\{R_1, R_2, \ldots, R_n\}$ be the set of all repairs of I w.r.t. Σ. The consistent answer to a query q is defined by:

$$q_\Sigma(I) := \bigcap_{i=1}^{n} q(R_i)$$

- For example, "`SELECT Name FROM I`" yields $\{\}$.
- This semantics is old (at least 1984):

  $q_\Sigma(I)$ is the certain answer to q on the incomplete relation $\{R_1, R_2, \ldots, R_n\}$.

Prelim. II: Consistent Query Answering

- Let $\{R_1, R_2, \ldots, R_n\}$ be the set of all repairs of I w.r.t. Σ. The consistent answer to a query q is defined by:

$$q_\Sigma(I) := \bigcap_{i=1}^{n} q(R_i)$$

- For example, "`SELECT Name FROM I`" yields $\{\}$.
- This semantics is old (at least 1984):

  $q_\Sigma(I)$ is the certain answer to q on the incomplete relation $\{R_1, R_2, \ldots, R_n\}$.

- Consistent query answering is the complexity of the set:

$$\mathsf{CQA}(\Sigma, q) := \{I \mid I \text{ is a relation and } q_\Sigma(I) = \{\}\}$$

Project-Join-Repair (Example)

Before repairing, apply the project-join dependency

$$\sigma = \bowtie [\{\mathbf{Name}, \mathbf{Birth}, \mathbf{Sex}, \mathbf{ZIP}\}, \{\mathbf{ZIP}, \mathbf{City}\}] \quad :$$

I

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Mons
Ed	1962	M	7000	Bergen

Name	Birth	Sex	ZIP
An	1964	F	7000
Ed	1962	M	7000

ZIP	City
7000	Mons
7000	Bergen

$\sigma(I)$

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Mons
Ed	1962	M	7000	Bergen
An	1964	F	7000	Bergen
Ed	1962	M	7000	Mons

Project-Join-Repair (Ex. Contd.)

$\sigma(I)$

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Mons
Ed	1962	M	7000	Bergen
An	1964	F	7000	Bergen
Ed	1962	M	7000	Mons

I

J_1

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Mons
Ed	1962	M	7000	Mons

has two repairs:

J_2

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Bergen
Ed	1962	M	7000	Bergen

The effect: repairing by value modification!

Project-Join-Repair

- $\Sigma =$ set of fd's $\sigma =$ join dependency (jd)
  $I =$ possibly inconsistent relation
- We repair $\sigma(I)$—rather than I. Note that $\sigma(I) \supseteq I$.
- We require $\Sigma \models \sigma$.

  Therefore, if I satisfies Σ, then $\sigma(I) = I$.

Project-Join-Repair

- $\Sigma =$ set of fd's $\sigma =$ join dependency (jd)
  $I =$ possibly inconsistent relation
- We repair $\sigma(I)$—rather than I. Note that $\sigma(I) \supseteq I$.
- We require $\Sigma \models \sigma$.

  Therefore, if I satisfies Σ, then $\sigma(I) = I$.
- Consistent query answering becomes the complexity of:

  $\mathsf{CQAJD}(\Sigma, \sigma, q) := \{I \mid I \text{ is a relation and } q_\Sigma(\sigma(I)) = \{\}\}$

  Determine tractable cases of this problem!
- If σ is the identity jd, then $\mathsf{CQAJD}(\Sigma, \sigma, q) = \mathsf{CQA}(\Sigma, q)$.

Tractable Cases

- Let fd's in Σ be of the form $X \rightarrow A$ with $A \notin X$.
- Main theorem:

  $\mathsf{CQAJD}(\Sigma, \sigma, q)$ is in **P** under the following condition:

  - Σ is acyclic (and hence has unique key K);
  - no two fd's of Σ have the same right-hand side;
  - for each $X \rightarrow A$ of Σ, $K \subseteq X$ or $\sigma \models X \rightarrow\!\!\!\!\rightarrow A$; and
  - q is a rooted, typed conjunctive query.

Tractable Cases

- Let fd's in Σ be of the form $X \to A$ with $A \notin X$.
- Main theorem:

  $\mathsf{CQAJD}(\Sigma, \sigma, q)$ is in **P** under the following condition:

  - Σ is acyclic (and hence has unique key K);
  - no two fd's of Σ have the same right-hand side;
  - for each $X \to A$ of Σ, $K \subseteq X$ or $\sigma \models X \twoheadrightarrow A$; and
  - q is a rooted, typed conjunctive query.

- There are cases of $\mathsf{CQA}(\Sigma, q)$ that are **NP**-complete under these conditions!
- To conclude, repairing $\sigma(I)$—rather than I—can be more natural and more efficient at the same time.

Preservation of Key Values

- Acyclic $\Leftrightarrow$ attributes can be ordered such that if $XA \rightarrow B$ is in a minimal cover of Σ, then $A < B$.
- Every repair preserves all key values.

$I \not\models X \rightarrow A$

$\underline{K}$	X	A	Y
k	**x**	**a**	**y**
l	**x**	**b**	**y′**

$\Downarrow$

R_1

$\underline{K}$	X	A	Y
k	**x**	**a**	**y**

$\sigma(I) \models X \twoheadrightarrow A$

$\underline{K}$	X	A	Y
k	**x**	**a**	**y**
l	**x**	**b**	**y′**
l	**x**	**a**	**y′**
	...		

J_1

$\underline{K}$	X	A	Y
k	**x**	**a**	**y**
l	**x**	**a**	**y′**

Rooted, Typed Conjunctive Queries

Key positions are underlined in:

$$Answer(\vec{z}) \leftarrow R(\underline{\vec{x}_1}, \vec{y}_1), \ldots, R(\underline{\vec{x}_n}, \vec{y}_n)$$

Typed No variable occurs at distinct positions in R-atoms.

Rooted If $R(\underline{\vec{x}_i}, \vec{y}_i)$ and $R(\underline{\vec{x}_j}, \vec{y}_j)$ with $i \neq j$ share a common variable w, then w occurs in $\vec{z}$ (i.e. w is free)
or all variables of $\vec{x}_i, \vec{x}_j$ occur in $\vec{z}$.

Rooted, Typed Conjunctive Queries

Key positions are underlined in:

$$Answer(\vec{z}) \leftarrow R(\underline{\vec{x}_1}, \vec{y}_1), \ldots, R(\underline{\vec{x}_n}, \vec{y}_n)$$

Typed No variable occurs at distinct positions in R-atoms.

Rooted If $R(\underline{\vec{x}_i}, \vec{y}_i)$ and $R(\underline{\vec{x}_j}, \vec{y}_j)$ with $i \neq j$ share a common variable w, then w occurs in $\vec{z}$ (i.e. w is free)
or all variables of $\vec{x}_i, \vec{x}_j$ occur in $\vec{z}$.

- Get names of male persons living in An's city.

$$\langle x_n \rangle \leftarrow \langle \underline{x_n}, x_b, \mathsf{M}, x_z, x_c \rangle, \langle \underline{\mathsf{An}}, y_b, y_s, y_z, x_c \rangle$$

- Man–Woman pairs with same birth and city.

$$\langle x_n, y_n \rangle \leftarrow \langle \underline{x_n}, x_b, \mathsf{M}, x_z, x_c \rangle, \langle \underline{y_n}, x_b, \mathsf{F}, y_z, x_c \rangle$$

Computing the Consistent Answers

$\sigma(I)$

Name	Birth	Sex	ZIP	City
An	1964	F	7000	Mons
Ed	1962	M	7000	Bergen
An	1964	F	7000	Bergen
Ed	1962	M	7000	Mons

$\Downarrow$

N

Name	Birth	Sex	ZIP	City
An	1964	F	7000	w
Ed	1962	M	7000	w

Under the conditions of the main theorem:

For any rooted, typed conjunctive query q,
$q_\Sigma(\sigma(I)) =$ the ground tuples in $q(N)$.

Conclusion

- Database repairing and consistent query answering under fd's.
- Applying a join dependency prior to repairing by tuple deletion:
  - is semantically meaningful;

    the effect is a sort of repairing by value modification.
  - may give you tractability.