Data Exchange and Metadata Management

Marcelo Arenas

PUC Chile & U. of Oxford

Dagstuhl, November 2012
Outline of the talk

- Data exchange: An overview of the relational case
- Metadata management
- Composition operator
- Concluding remarks
Outline of the talk

- Data exchange: An overview of the relational case
- Metadata management
- Composition operator
- Concluding remarks
The problem of data exchange

Given: A source schema $S$, a target schema $T$ and a specification $\Sigma_{ST}$ of the relationship between these schemas

Data exchange: Problem of materializing an instance of $T$ given an instance of $S$

- Target instance should reflect the source data as accurately as possible, given the constraints imposed by $\Sigma_{ST}$ and $T$
- It should be efficiently computable
- It should allow one to evaluate queries on the target in a way that is semantically consistent with the source data
Data exchange in a picture

\[ \Sigma_{ST} \]

Schema S \quad \text{Schema T}
Data exchange in a picture

\[ \Sigma_{ST} \]

Schema S \rightarrow Schema T
Data exchange in a picture

Schema $S$ \hspace{1cm} \Sigma_{ST} \hspace{1cm} Schema $T$
Data exchange in a picture

Schema S

\[ \Sigma_{ST} \]

Schema T
Data exchange in a picture

\[ \Sigma_{ST} \]

Schema \( S \) \hspace{2cm} \text{Query} \ Q \hspace{2cm} \text{Schema} \ T
Data exchange: Some fundamental questions

Why is data exchange an interesting problem?

- Is it a difficult problem?

What are the challenges in the area?

- What is a good language for specifying the relationship between source and target data?
- What is a good instance to materialize? Why is it good?
- What does it mean to answer a queries over target data?
- How do we answer queries over target data? Can we do this efficiently?
Data exchange in relational databases

It has been extensively studied in the relational world.
- It has also been implemented: IBM Clio

Relational data exchange setting:
- Source and target schemas: Relational schemas
- Relationship between source and target schemas: Source-to-target tuple-generating dependencies (st-tgds)

Semantics of data exchange has been precisely defined.
- Efficient algorithms for materializing target instances and for answering queries over the target schema have been developed
Schema mapping: The key component in relational data exchange

Schema mapping: $\mathcal{M} = (S, T, \Sigma_{ST})$

- $S$ and $T$ are disjoint relational schemas
- $\Sigma_{ST}$ is a finite set of st-tgds:

  $$\forall \bar{x} \forall \bar{y} (\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z}))$$

  - $\varphi(\bar{x}, \bar{y})$: conjunction of relational atomic formulas over $S$
  - $\psi(\bar{x}, \bar{z})$: conjunction of relational atomic formulas over $T$
Relational data exchange problem

Fixed: $\mathcal{M} = (S, T, \Sigma_{ST})$

Problem: Given instance $I$ of $S$, find an instance $J$ of $T$ such that $(I, J)$ satisfies $\Sigma_{ST}$

- $(I, J)$ satisfies $\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})$ if whenever $I$ satisfies $\varphi(\bar{a}, \bar{b})$, there is a tuple $\bar{c}$ such that $J$ satisfies $\psi(\bar{a}, \bar{c})$
Relational data exchange problem

Fixed: \( \mathcal{M} = (S, T, \Sigma_{ST}) \)

Problem: Given instance \( I \) of \( S \), find an instance \( J \) of \( T \) such that \( (I, J) \) satisfies \( \Sigma_{ST} \)

\( (I, J) \) satisfies \( \varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \, \psi(\bar{x}, \bar{z}) \) if whenever \( I \) satisfies \( \varphi(\bar{a}, \bar{b}) \), there is a tuple \( \bar{c} \) such that \( J \) satisfies \( \psi(\bar{a}, \bar{c}) \)

**Notation**

\( J \) is a **solution** for \( I \) under \( \mathcal{M} \)

\( \text{Sol}_\mathcal{M}(I) \): Set of solutions for \( I \) under \( \mathcal{M} \)
Relational data exchange: An example

Example

- **S**: `employee(name)`
- **T**: `dept(name, number)`
- **Σ_{ST}**: `employee(x) → ∃y dept(x, y)`

Solutions for `I = {employee(Peter)}`:
Relational data exchange: An example

Example

- **S**: \( \text{employee}(\text{name}) \)
- **T**: \( \text{dept}(\text{name}, \text{number}) \)
- \( \Sigma_{ST} \): \( \text{employee}(x) \rightarrow \exists y \text{dept}(x, y) \)

Solutions for \( l = \{\text{employee}(Peter)\} \):

- \( J_1 \): \( \text{dept}(Peter, 1) \)
Example

- S: \texttt{employee(name)}
- T: \texttt{dept(name, number)}
- \Sigma_{ST}: \texttt{employee}(x) \rightarrow \exists y \texttt{dept}(x, y)

Solutions for \( I = \{ \texttt{employee(Peter)} \} \):

\( J_1 \): \texttt{dept(Peter,1)}

\( J_2 \): \texttt{dept(Peter,1)}, \texttt{dept(Peter,2)}
Relational data exchange: An example

Example

- **S**: `employee(name)`
- **T**: `dept(name, number)`
- **Σ_{ST}**: `employee(x) → ∃y dept(x, y)`

Solutions for `I = \{employee(Peter)\}`:

- **J_1**: `dept(Peter,1)`
- **J_2**: `dept(Peter,1), dept(Peter,2)`
- **J_3**: `dept(Peter,1), dept(John,1)`
Relational data exchange: An example

Example

- **S**: `employee(name)`
- **T**: `dept(name, number)`
- **Σ**:

\[
\Sigma_{ST} : \text{employee}(x) \rightarrow \exists y \text{ dept}(x, y)
\]

Solutions for \( I = \{\text{employee}(Peter)\} \):

- \( J_1 \): `dept(Peter,1)`
- \( J_2 \): `dept(Peter,1), dept(Peter,2)`
- \( J_3 \): `dept(Peter,1), dept(John,1)`
- \( J_4 \): `dept(Peter,n_1)`
Relational data exchange: An example

Example

- \( S: \) employee\((name)\)
- \( T: \) dept\((name, number)\)
- \( \Sigma_{ST}: \) employee\((x) \rightarrow \exists y \) dept\((x, y)\)

Solutions for \( I = \{\text{employee}(Peter)\}\):

- \( J_1: \) dept\((Peter,1)\)
- \( J_2: \) dept\((Peter,1)\), dept\((Peter,2)\)
- \( J_3: \) dept\((Peter,1)\), dept\((John,1)\)
- \( J_4: \) dept\((Peter, n_1)\)
- \( J_5: \) dept\((Peter, n_1)\), dept\((Peter, n_2)\)
Question
What is a good instance to materialize?
Canonical universal solution

Question
What is a good instance to materialize?

Algorithm (chase)

Input : \((S, T, \Sigma_{ST})\) and an instance \(I\) of \(S\)
Output : Canonical universal solution \(J^*\) for \(I\) under \(\mathcal{M}\)

let \(J^* :=\) empty instance of \(T\)
for every \(\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{z})\) in \(\Sigma_{ST}\) do
  for every \(\bar{a}, \bar{b}\) such that \(I\) satisfies \(\varphi(\bar{a}, \bar{b})\) do
    create a fresh tuple \(\bar{n}\) of pairwise distinct null values
    insert \(\psi(\bar{a}, \bar{n})\) into \(J^*\)
Example

Consider mapping $\mathcal{M}$ specified by dependency:

$$employee(x) \rightarrow \exists y \, dept(x, y)$$

Canonical universal solution for $I = \{employee(Peter), employee(John)\}$:

$$J^* = \{dept(Peter, n_1), \, dept(John, n_2)\}$$
Given: Mapping $\mathcal{M}$, source instance $I$ and query $Q$ over the target schema

- What does it mean to answer $Q$?
Query answering in data exchange

Given: Mapping $\mathcal{M}$, source instance $I$ and query $Q$ over the target schema

- What does it mean to answer $Q$?

**Definition (Certain answers)**

$$\text{certain}_\mathcal{M}(Q, I) = \bigcap_{J \text{ is a solution for } I \text{ under } \mathcal{M}} Q(J)$$
Certain answers: Example

Example

Consider mapping $\mathcal{M}$ specified by:

$$\text{employee}(x) \rightarrow \exists y \text{ dept}(x, y)$$

Given instance $I = \{ \text{employee}(Peter) \}$:

$$\text{certain}_{\mathcal{M}}(\exists y \text{ dept}(x, y), I) = \{Peter\}$$

$$\text{certain}_{\mathcal{M}}(\text{dept}(x, y), I) = \emptyset$$
How can we compute certain answers?

- Naïve algorithm does not work: infinitely many solutions
Query rewriting: An approach for answering queries

How can we compute certain answers?

- Naïve algorithm does not work: infinitely many solutions

Approach proposed in [FKMP03]: **Query Rewriting**

Given a mapping $\mathcal{M}$ and a target query $Q$, compute a query $Q^*$ such that for every source instance $I$ with canonical universal solution $J^*$:

$$\text{certain}_{\mathcal{M}}(Q, I) = Q^*(J^*)$$
Query rewriting over the canonical universal solution

Theorem (FKMP03)

Given a mapping $\mathcal{M}$ specified by st-tgds and a union of conjunctive queries $Q$, there exists a query $Q^*$ such that for every source instance $I$ with canonical universal solution $J^*$:

$$\text{certain}_{\mathcal{M}}(Q, I) = Q^*(J^*)$$
Theorem (FKMP03)

Given a mapping $\mathcal{M}$ specified by st-tgds and a union of conjunctive queries $Q$, there exists a query $Q^*$ such that for every source instance $I$ with canonical universal solution $J^*$:

$$\text{certain}_{\mathcal{M}}(Q, I) = Q^*(J^*)$$

**Proof idea:** Assume that $C(a)$ holds whenever $a$ is a constant.

Then:

$$Q^*(x_1, \ldots, x_m) = C(x_1) \land \cdots \land C(x_m) \land Q(x_1, \ldots, x_m)$$
Computing certain answers: Complexity

Data complexity: Data exchange setting and query are considered to be fixed.

Corollary (FKMP03)

For mappings given by st-tgds, certain answers for \textbf{UCQ} can be computed in polynomial time (data complexity)
Key steps in the development of the area:

- Definition of schema mappings: Precise syntax and semantics
  - Definition of the notion of solution
- Identification of good solutions
- Polynomial time algorithms for materializing good solutions
- Definition of target queries: Precise semantics
- Polynomial time algorithms for computing certain answers for UCQ
Relational data exchange: Some lessons learned

Key steps in the development of the area:

- Definition of schema mappings: Precise syntax and semantics
  - Definition of the notion of solution
- Identification of good solutions
- Polynomial time algorithms for materializing good solutions
- Definition of target queries: Precise semantics
- Polynomial time algorithms for computing certain answers for UCQ

Creating schema mappings is a time consuming and expensive process

- Manual or semi-automatic process in general
Outline of the talk

- Data exchange: An overview of the relational case
- Metadata management
- Composition operator
- Concluding remarks
Can we reuse schema mappings?
Can we reuse schema mappings?

\[ S \xrightarrow{\Sigma_{ST}} T \xrightarrow{\Sigma_{TU}} U \]
Can we reuse schema mappings?

\[ \Sigma_{ST} \quad \Sigma_{TU} \]

\[ \Sigma_{SU}? \]
Can we reuse schema mappings?

We need some operators for schema mappings
Can we reuse schema mappings?

\[ \Sigma_{SU} = \Sigma_{ST} \circ \Sigma_{TU} \]

We need some operators for schema mappings

- **Composition** in the above case
Contributions mentioned in the previous slides are just a first step towards the development of a general framework for data exchange.

In fact, as pointed in [B03],

many information system problems involve not only the design and integration of complex application artifacts, but also their subsequent manipulation.
Metadata management

This has motivated the need for the development of a general infrastructure for managing schema mappings.

The problem of managing schema mappings is called metadata management.

High-level algebraic operators, such as compose, are used to manipulate mappings.

- What other operators are needed?
More operators are needed

\[ \Sigma_{ST} \]
More operators are needed
More operators are needed

\[ \Sigma_{ST} \]
More operators are needed

\[ \Sigma_{VS} = \Sigma_{SV}^{-1} \]

An \textbf{inverse} operator is needed in this case
More operators are needed

An **inverse** operator is needed in this case

- Combined with the composition operator
More operators are needed

\[ \Sigma_{VS} = \Sigma_{SV}^{-1} \]

\[ \Sigma_{SV} \circ \Sigma_{ST} = (\Sigma_{SV}^{-1} \circ \Sigma_{ST}) \circ \Sigma_{TU} \]

An **inverse** operator is needed in this case

- Combined with the composition operator
Outline of the talk

- Data exchange: An overview of the relational case
- Metadata management
- Composition operator
- Concluding remarks
The composition operator

Question

What is the semantics of the composition operator?
The composition operator

**Question**
What is the semantics of the composition operator?

**Notation**
We can view a mapping $\mathcal{M}$ as a set of pairs:

$$(I, J) \in \mathcal{M} \iff J \in \text{Sol}_{\mathcal{M}}(I)$$
# The composition operator

## Question

What is the semantics of the composition operator?

## Notation

We can view a mapping $\mathcal{M}$ as a set of pairs:

$$(I, J) \in \mathcal{M} \iff J \in \text{Sol}_{\mathcal{M}}(I)$$

## Definition (FKPT04)

Let $\mathcal{M}_{12}$ be a mapping from $S_1$ to $S_2$, and $\mathcal{M}_{23}$ a mapping from $S_2$ to $S_3$:

$$\mathcal{M}_{12} \circ \mathcal{M}_{23} = \{(l_1, l_3) \mid \exists l_2 : (l_1, l_2) \in \mathcal{M}_{12} \text{ and } (l_2, l_3) \in \mathcal{M}_{23}\}$$
Expressing the composition of mappings

**Question**

What is the right language for expressing the composition?

- st-tgds?
Expressing the composition of mappings

**Question**

What is the right language for expressing the composition?

- st-tgds?

**Example (FKPT04)**

Consider mappings:

\[
M_{12} : \quad \text{takes}(n, c) \rightarrow \text{takes}_1(n, c)
\]

\[
\text{takes}(n, c) \rightarrow \exists s \ \text{student}(n, s)
\]

\[
M_{23} : \quad \text{student}(n, s) \land \text{takes}_1(n, c) \rightarrow \text{enrolled}(s, c)
\]
Expressing the composition of mappings

Question
What is the right language for expressing the composition?
- st-tgds?

Example (FKPT04)
Consider mappings:

\[ M_{12} : \quad \text{takes}(n, c) \rightarrow \text{takes}_1(n, c) \]
\[ \text{takes}(n, c) \rightarrow \exists s \, \text{student}(n, s) \]
\[ M_{23} : \quad \text{student}(n, s) \land \text{takes}_1(n, c) \rightarrow \text{enrolled}(s, c) \]

Does the following st-tgd express the composition?
\[ \text{takes}(n, c) \rightarrow \exists y \, \text{enrolled}(y, c) \]
Example (Cont’d)

This is the right dependency:

\[ \forall n \exists y \forall c \ (\text{takes}(n, c) \rightarrow \text{enrolled}(y, c)) \]
Example (Cont’d)

This is the right dependency:

$$\forall n \exists y \forall c \ (\text{takes}(n, c) \rightarrow \text{enrolled}(y, c))$$

Is first-order logic enough?

- Complexity theory can help us to answer this question
Expressing the composition of mappings: A complexity argument

How difficult is the composition problem?

- Fix mappings $M_{12}$ and $M_{23}$
- Problem: Decide whether $(l_1, l_3) \in M_{12} \circ M_{23}$

If $M_{12} \circ M_{23}$ is defined by a set of first-order sentences, then the composition problem can be solved efficiently: It is in $AC^0$

- $AC^0 \subsetneq PTIME$
Expressing the composition of mappings: A complexity argument

How difficult is the composition problem?

- Fix mappings $\mathcal{M}_{12}$ and $\mathcal{M}_{23}$
- Problem: Decide whether $(I_1, I_3) \in \mathcal{M}_{12} \circ \mathcal{M}_{23}$

If $\mathcal{M}_{12} \circ \mathcal{M}_{23}$ is defined by a set of first-order sentences, then the composition problem can be solved efficiently: It is in $\text{AC}^0$

- $\text{AC}^0 \subset \text{PTIME}$

But the composition problem is not easy: It can be NP-hard

- $\text{AC}^0 \subset \text{PTIME} \subset \text{NP}$
Expressing the composition of mappings: A complexity argument

Let see a difficult case taken from [FKPT04].

\( \mathcal{M}_{12} \) is specified by:

\[
\begin{align*}
node(x) & \quad \rightarrow \quad \exists y \ coloring(x, y) \\
edge(x, y) & \quad \rightarrow \quad edge'(x, y)
\end{align*}
\]

\( \mathcal{M}_{23} \) is specified by:

\[
\begin{align*}
edge'(x, y) \land coloring(x, u) \land coloring(y, u) & \quad \rightarrow \quad error(x, y) \\
coloring(x, y) & \quad \rightarrow \quad color(y)
\end{align*}
\]
Expressing the composition of mappings: A complexity argument

What is the complexity of verifying whether \((l_1, l_3) \in M_{12} \circ M_{23}\)?
Expressing the composition of mappings: A complexity argument

What is the complexity of verifying whether \((l_1, l_3) \in M_{12} \circ M_{23}\)?

Given a graph \(G = (N, E)\), consider instances \(l_1, l_3\):

\[
\begin{align*}
\text{node in } l_1 & : N \\
\text{edge in } l_1 & : E \\
\text{color in } l_3 & : \{1, 2, 3\} \\
\text{error in } l_3 & : \emptyset
\end{align*}
\]

Then: \(G\) is 3-colorable iff \((l_1, l_3) \in M_{12} \circ M_{23}\)
Expressing the composition of mappings

Back to our initial question:

What is the right language for expressing the composition?
Expressing the composition of mappings

Back to our initial question:

What is the right language for expressing the composition?

Complexity theory can help us again:

- NP-hardness and Fagin’s theorem: We need at least existential second-order logic
Expressing the composition of mappings

Back to our initial question:

What is the right language for expressing the composition?

Complexity theory can help us again:

- NP-hardness and Fagin’s theorem: We need at least existential second-order logic
- Good news: There is a nice second-order language for expressing the composition
SO tgds: The right language for expressing the composition of mappings

Example

Consider again the mappings:

\[ \mathcal{M}_{12} : \text{takes}(n, c) \rightarrow \text{takes}_1(n, c) \]
\[ \text{takes}(n, c) \rightarrow \exists s \text{ student}(n, s) \]
\[ \mathcal{M}_{23} : \text{student}(n, s) \land \text{takes}_1(n, c) \rightarrow \text{enrolled}(s, c) \]

The following SO tgd defines the composition:

\[ \exists f \forall n \forall c \left( \text{takes}(n, c) \rightarrow \text{enrolled}(f(n), c) \right) \]
SO tgds: The right language for expressing the composition of mappings

Example

Consider again mappings $\mathcal{M}_{12}$:

- $\text{node}(x) \rightarrow \exists y \ \text{coloring}(x, y)$
- $\text{edge}(x, y) \rightarrow \text{edge}'(x, y)$

and $\mathcal{M}_{23}$:

- $\text{edge}'(x, y) \land \text{coloring}(x, u) \land \text{coloring}(y, u) \rightarrow \text{error}(x, y)$
- $\text{coloring}(x, y) \rightarrow \text{color}(y)$
SO tgds: The right language for expressing the composition of mappings

Example (Cont’d)

The following SO tgd defines the composition:

\[
\exists f \left[ \forall x (\text{node}(x) \rightarrow \text{color}(f(x))) \land \\
\forall x \forall y (\text{edge}(x, y) \land f(x) = f(y) \rightarrow \text{error}(x, y)) \right]
\]
SO tgds: The right language for expressing the composition of mappings

Example (Cont’d)

The following SO tgd defines the composition:

\[ \exists f \left[ \forall x (\text{node}(x) \rightarrow \text{color}(f(x))) \land \right. \]

\[ \forall x \forall y (\text{edge}(x, y) \land f(x) = f(y) \rightarrow \text{error}(x, y)) \] \]

This example shows the main ingredients of SO tgds:

- Predicates including terms: \( \text{color}(f(x)) \)
- Equality between terms in the premises: \( f(x) = f(y) \)
SO tgds: The right language for expressing the composition of mappings

SO tgds were introduced in [FKPT04]

- They have good properties regarding data exchange and composition
  - Canonical universal solution and certain answers to UCQ can be computed in polynomial time (data complexity)
SO tgds: The right language for expressing the composition of mappings

SO tgds were introduced in [FKPT04]

- They have good properties regarding data exchange and composition
  - Canonical universal solution and certain answers to UCQ can be computed in polynomial time (data complexity)

**Theorem (FKPT04)**

*If \( M_{12} \) and \( M_{23} \) are specified by SO tgds, then \( M_{12} \circ M_{23} \) can be specified by an SO tgd*
SO tgds: The right language for expressing the composition of mappings

SO tgds were introduced in [FKPT04]

- They have good properties regarding data exchange and composition
- Canonical universal solution and certain answers to UCQ can be computed in polynomial time (data complexity)

**Theorem (FKPT04)**

*If $M_{12}$ and $M_{23}$ are specified by SO tgds, then $M_{12} \circ M_{23}$ can be specified by an SO tgd*

- There exists an exponential time algorithm that computes such SO tgds
SO tgds: The right language for expressing the composition of mappings

Corollary (FKPT04)

The composition of a finite number of mappings, each defined by a finite set of st-tgds, is defined by an SO tgd
SO tgds: The right language for expressing the composition of mappings

Corollary (FKPT04)

The composition of a finite number of mappings, each defined by a finite set of st-tgds, is defined by an SO tgd

But not only that, SO tgds are exactly the right language:

Theorem (FKPT05)

Every SO tgd defines the composition of a finite number of mappings, each defined by a finite set of st-tgds.
Outline of the talk

- Data exchange: An overview of the relational case
- Metadata management
- Composition operator
- Concluding remarks
Concluding remarks

- Composition and inverse operators are fundamental in metadata management.
- The problem of composing schema mappings given by st-tgds is solved.
- Considerable progress has been made on the problem of inverting schema mappings.
- Combining these operators is an open issue.
  - Some progress has been made.
  - But we do not know whether there is a good language for both operators. Is there a reasonable language that is closed under both operators?
Thank you!
Backup slides

- Inverse operator

- Combination of both operators
  - Key ingredient: Conditional tables
Backup slides

- Inverse operator

- Combination of both operators
  - Key ingredient: Conditional tables
The inverse operator

\[ \Sigma_{ST} \]

Schema S \rightarrow \Sigma_{ST} \rightarrow \text{Schema T}
The inverse operator

\[ \Sigma_{ST} \]

Schema S \quad \rightarrow \quad \Sigma_{ST} \quad \rightarrow \quad \text{Schema T}
The inverse operator

Question
What is the semantics of the inverse operator?

This turns out to be a very difficult question.

We consider three notions of inverse here:

- Fagin-inverse
- Quasi-inverse
- Maximum recovery
The notion of Fagin-inverse

Intuition: A mapping composed with its inverse should be equal to the identity mapping
The notion of Fagin-inverse

Intuition: A mapping composed with its inverse should be equal to the identity mapping

What is the identity mapping?

\[ \text{Id}_S = \{(I, I) \mid I \text{ is an instance of } S\}? \]
The notion of Fagin-inverse

Intuition: A mapping composed with its inverse should be equal to the identity mapping.

What is the identity mapping?

\[ \text{Id}_S = \{ (I, I) \mid I \text{ is an instance of } S \} \]

For mapping specified by st-tgds, \( \text{Id}_S \) is not the right notion.

\[ \text{Id}_S = \{ (I_1, I_2) \mid I_1, I_2 \text{ are instances of } S \text{ and } I_1 \subseteq I_2 \} \]
Definition (F06)

Let $\mathcal{M}$ be a mapping from $S_1$ to $S_2$, and $\mathcal{M}^*$ a mapping from $S_2$ to $S_1$. Then $\mathcal{M}^*$ is a Fagin-inverse of $\mathcal{M}$ if:

$$\mathcal{M} \circ \mathcal{M}^* = \overline{\text{Id}}_{S_1}$$
The notion of Fagin-inverse: Formal definition

**Definition (F06)**

Let \( \mathcal{M} \) be a mapping from \( S_1 \) to \( S_2 \), and \( \mathcal{M}^* \) a mapping from \( S_2 \) to \( S_1 \). Then \( \mathcal{M}^* \) is a Fagin-inverse of \( \mathcal{M} \) if:

\[
\mathcal{M} \circ \mathcal{M}^* = \overline{\text{Id}}_{S_1}
\]

**Example**

Consider mapping \( \mathcal{M} \) specified by:

\[
A(x) \rightarrow R(x) \land \exists y S(x, y)
\]

Then the following are Fagin-inverses of \( \mathcal{M} \):

\[
\begin{align*}
\mathcal{M}_1^* & : \ R(x) \rightarrow A(x) \\
\mathcal{M}_2^* & : \ S(x, y) \rightarrow A(x)
\end{align*}
\]
Is Fagin-inverse the right notion of inverse for mappings?

On the positive side: It is a natural notion
  ▶ With good computational properties

On the negative side: A mapping specified by st-tgds is not guaranteed to admit a Fagin-inverse
  ▶ For example: Mapping specified by \( A(x, y) \rightarrow R(x) \) does not admit a Fagin-inverse

In fact: This notion turns out to be rather restrictive, as it is rare that a schema mapping possesses a Fagin-inverse.
Is Fagin-inverse the right notion of inverse for mappings?

The notion of quasi-inverse was introduced in [FKPT07] to overcome this limitation.

- The idea is to relax the notion of Fagin-inverse by not differentiating between source instances that are equivalent for data exchange purposes.

Numerous non-Fagin-invertible mappings possess natural and useful quasi-inverses.

- But there are still simple mappings specified by st-tgds that have no quasi-inverse.

The notion of maximum recovery overcome this limitation.
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
  - $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$
Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$

  - $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

**Example**

Consider a mapping $\mathcal{M}$ specified by:

$$emp(x, y, z) \land y \neq z \rightarrow shuttle(x, z)$$

What mappings are recoveries of $\mathcal{M}$?
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$
- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
- $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

Example

Consider a mapping $\mathcal{M}$ specified by:

$$emp(x, y, z) \land y \neq z \rightarrow shuttle(x, z)$$

What mappings are recoveries of $\mathcal{M}$?

$\mathcal{M}_1^*$: $shuttle(x, z) \rightarrow \exists u \exists v emp(x, u, v)$
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
- $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

Example

Consider a mapping $\mathcal{M}$ specified by:

$emp(x, y, z) \land y \neq z \rightarrow shuttle(x, z)$

What mappings are recoveries of $\mathcal{M}$?

$\mathcal{M}_1^*: shuttle(x, z) \rightarrow \exists u \exists v \ emp(x, u, v)$ ✔
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
- $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

Example

Consider a mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

What mappings are recoveries of $\mathcal{M}$?

- $\mathcal{M}_1^*$: $\text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v)$ ✓
- $\mathcal{M}_2^*$: $\text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z)$
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
  - $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

**Example**

Consider a mapping $\mathcal{M}$ specified by:

$$emp(x, y, z) \land y \neq z \rightarrow shuttle(x, z)$$

What mappings are recoveries of $\mathcal{M}$?

- $\mathcal{M}_1^*$: $shuttle(x, z) \rightarrow \exists u \exists v \ emp(x, u, v)$ ✓
- $\mathcal{M}_2^*$: $shuttle(x, z) \rightarrow \exists u \ emp(x, u, z)$ ✓
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
  - $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

Example

Consider a mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

What mappings are recoveries of $\mathcal{M}$?

- $\mathcal{M}_1^*$: $\text{shuttle}(x, z) \rightarrow \exists u \exists v \ \text{emp}(x, u, v)$
  - $\checkmark$
- $\mathcal{M}_2^*$: $\text{shuttle}(x, z) \rightarrow \exists u \ \text{emp}(x, u, z)$
  - $\checkmark$
- $\mathcal{M}_3^*$: $\text{shuttle}(x, z) \rightarrow \exists u \ \text{emp}(x, z, u)$
Recovery: specifies how to recover sound information

Data may be lost in the exchange through a mapping $\mathcal{M}$

- We would like to find a mapping $\mathcal{M}^*$ that at least recovers sound data w.r.t. $\mathcal{M}$
- $\mathcal{M}^*$ is called a recovery of $\mathcal{M}$

**Example**

Consider a mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

What mappings are recoveries of $\mathcal{M}$?

- $\mathcal{M}_1^*$: $\text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v)$ ✓
- $\mathcal{M}_2^*$: $\text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z)$ ✓
- $\mathcal{M}_3^*$: $\text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, z, u)$ ×
Maximum recovery: The most informative recovery

Example

Consider again mapping $\mathcal{M}$ specified by:

$$emp(x, y, z) \land y \neq z \rightarrow shuttle(x, z)$$

These mappings are recoveries of $\mathcal{M}$:

- $\mathcal{M}^*_1$: $shuttle(x, z) \rightarrow \exists u \exists v \ emp(x, u, v)$
- $\mathcal{M}^*_2$: $shuttle(x, z) \rightarrow \exists u \ emp(x, u, z)$
Example

Consider again mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

These mappings are recoveries of $\mathcal{M}$:

- $\mathcal{M}_1^*: \text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v)$

- $\mathcal{M}_2^*: \text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z)$

Intuitively: $\mathcal{M}_2^*$ is better than $\mathcal{M}_1^*$
Maximum recovery: The *most informative* recovery

Example

Consider again mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

These mappings are recoveries of $\mathcal{M}$:

- $\mathcal{M}_{1}^*$: $\text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v)$
- $\mathcal{M}_{2}^*$: $\text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z)$
- $\mathcal{M}_{4}^*$: $\text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z) \land u \neq z$

Intuitively: $\mathcal{M}_{2}^*$ is better than $\mathcal{M}_{1}^*$
Example

Consider again mapping \( \mathcal{M} \) specified by:

\[
emp(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)
\]

These mappings are recoveries of \( \mathcal{M} \):

- \( \mathcal{M}_1^*: \text{shuttle}(x, z) \rightarrow \exists u \exists v \ emp(x, u, v) \)
- \( \mathcal{M}_2^*: \text{shuttle}(x, z) \rightarrow \exists u \ emp(x, u, z) \)
- \( \mathcal{M}_4^*: \text{shuttle}(x, z) \rightarrow \exists u \ emp(x, u, z) \land u \neq z \)

Intuitively: \( \mathcal{M}_2^* \) is better than \( \mathcal{M}_1^* \)

\( \mathcal{M}_4^* \) is better than \( \mathcal{M}_2^* \) and \( \mathcal{M}_1^* \)
Maximum recovery: The *most informative* recovery

**Example**

Consider again mapping $\mathcal{M}$ specified by:

$$\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)$$

These mappings are recoveries of $\mathcal{M}$:

- $\mathcal{M}_1^*: \text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v)$
- $\mathcal{M}_2^*: \text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z)$
- $\mathcal{M}_4^*: \text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z) \land u \neq z$

Intuitively: $\mathcal{M}_2^*$ is better than $\mathcal{M}_1^*$

$\mathcal{M}_4^*$ is better than $\mathcal{M}_2^*$ and $\mathcal{M}_1^*$

We would like to find a recovery of $\mathcal{M}$ that is better than any other recovery: Maximum recovery
The notion of recovery: Formalization

Definition (APR08)

Let $\mathcal{M}$ be a mapping from $S_1$ to $S_2$ and $\mathcal{M}^*$ a mapping from $S_2$ to $S_1$. Then $\mathcal{M}^*$ is a recovery of $\mathcal{M}$ if:

for every instance $I$ of $S_1$: $(I, I) \in \mathcal{M} \circ \mathcal{M}^*$
The notion of recovery: Formalization

**Definition (APR08)**

Let \( \mathcal{M} \) be a mapping from \( S_1 \) to \( S_2 \) and \( \mathcal{M}^* \) a mapping from \( S_2 \) to \( S_1 \). Then \( \mathcal{M}^* \) is a recovery of \( \mathcal{M} \) if:

\[
\text{for every instance } I \text{ of } S_1: (I, I) \in \mathcal{M} \circ \mathcal{M}^*
\]

**Example**

Consider again mapping \( \mathcal{M} \) specified by:

\[
\text{emp}(x, y, z) \land y \neq z \rightarrow \text{shuttle}(x, z)
\]

This mapping is not a recovery of \( \mathcal{M} \):

\[
\mathcal{M}^*_3: \text{shuttle}(x, z) \rightarrow \exists u \text{ emp}(x, z, u)
\]
The notion of recovery: Formalization

Example (Cont’d)

On the other hand, these mappings are recoveries of $\mathcal{M}$:

\begin{align*}
\mathcal{M}^*_1 & : \text{shuttle}(x, z) \rightarrow \exists u \exists v \text{emp}(x, u, v) \\
\mathcal{M}^*_2 & : \text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z) \\
\mathcal{M}^*_4 & : \text{shuttle}(x, z) \rightarrow \exists u \text{emp}(x, u, z) \land u \neq z
\end{align*}
The notion of maximum recovery
The notion of maximum recovery

\[ M \xrightarrow{M^*_1} M \]

\[ M \xleftarrow{M^*_1} M \]

\[ M \xrightarrow{M^*_1} M \]

\[ M \xleftarrow{M^*_1} M \]
The notion of maximum recovery
The notion of maximum recovery
The notion of maximum recovery

Definition (APR08)

\( M^* \) is a maximum recovery of \( M \) if:

- \( M^* \) is a recovery of \( M \)
- for every recovery \( M' \) of \( M \): \( M \circ M^* \subseteq M \circ M' \)
A basic property of (maximum) recoveries

We have seen three notions of inversion for mappings.

- How can we show that a notion of inverse is appropriate?
A basic property of (maximum) recoveries

We have seen three notions of inversion for mappings.

- How can we show that a notion of inverse is appropriate?

A criterion: How much of the initial information is recovered?
A basic property of (maximum) recoveries

We have seen three notions of inversion for mappings.

- How can we show that a notion of inverse is appropriate?

A criterion: How much of the initial information is recovered?

- How close is a space of solution to a particular solution?
We have seen three notions of inversion for mappings.

- How can we show that a notion of inverse is appropriate?

A criterion: How much of the initial information is recovered?

- How close is a space of solution to a particular solution? How close is $\text{Sol}_{M\circ M^*}(I)$ to $I$?
A basic property of (maximum) recoveries

We have seen three notions of inversion for mappings.

- How can we show that a notion of inverse is appropriate?

A criterion: How much of the initial information is recovered?

- How close is a space of solution to a particular solution? How close is $\text{Sol}_{M \circ M^*}(I)$ to $I$?

Simple approach: Compare the information that can be retrieved from $I$ and $\text{Sol}_{M \circ M^*}(I)$
A basic property of (maximum) recoveries

To compare the information that can be retrieved from $I$ and $\text{Sol}_{M \circ M^*}(I)$: Compare $Q(I)$ to certain $M \circ M^*(Q, I)$
A basic property of (maximum) recoveries

To compare the information that can be retrieved from $I$ and $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(I)$: Compare $Q(I)$ to $\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I)$

**Observation**

Let $\mathcal{M}$ be a mapping from $S$ to $T$, $I$ an instance of $S$, $Q$ a query over $S$ and $\mathcal{M}^*$ a recovery of $\mathcal{M}$:

$$\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I) \subseteq Q(I)$$
A basic property of (maximum) recoveries

To compare the information that can be retrieved from $I$ and $\text{Sol}_{M \circ M^*}(I)$: Compare $Q(I)$ to certain $M \circ M^*(Q, I)$

**Observation**

Let $M$ be a mapping from $S$ to $T$, $I$ an instance of $S$, $Q$ a query over $S$ and $M^*$ a recovery of $M$:

$$\text{certain}_{M \circ M^*}(Q, I) \subseteq Q(I)$$

Information retrieved from $\text{Sol}_{M \circ M^*}(I)$ is sound w.r.t. $I$. 
A basic property of (maximum) recoveries

To compare the information that can be retrieved from $I$ and $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(I)$: Compare $Q(I)$ to $\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I)$

Observation

Let $\mathcal{M}$ be a mapping from $S$ to $T$, $I$ an instance of $S$, $Q$ a query over $S$ and $\mathcal{M}^*$ a recovery of $\mathcal{M}$:

$$\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I) \subseteq Q(I)$$

Information retrieved from $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(I)$ is sound w.r.t. $I$.

- Is $\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I) = Q(I)$?
A basic property of (maximum) recoveries

To compare the information that can be retrieved from $I$ and $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(I)$: Compare $Q(I)$ to certain $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I)$

**Observation**

Let $\mathcal{M}$ be a mapping from $S$ to $T$, $I$ an instance of $S$, $Q$ a query over $S$ and $\mathcal{M}^*$ a recovery of $\mathcal{M}$:

$$\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I) \subseteq Q(I)$$

Information retrieved from $\text{Sol}_{\mathcal{M} \circ \mathcal{M}^*}(I)$ is sound w.r.t. $I$.

- Is $\text{certain}_{\mathcal{M} \circ \mathcal{M}^*}(Q, I) = Q(I)$?
- Not always possible: $P(x, y) \rightarrow R(x)$ and $Q(x, y) = P(x, y)$
A fundamental property of maximum recoveries

Definition

- \( M' \) recovers \( Q \) under \( M \) if for every source instance \( I \):

\[
Q(I) = \text{certain}_{M \circ M'}(Q, I)
\]

- \( Q \) can be recovered under \( M \) if the above mapping \( M' \) exists
A fundamental property of maximum recoveries

Definition

- $M'$ recovers $Q$ under $M$ if for every source instance $I$:

  $$Q(I) = \text{certain}_{M,M'}(Q, I)$$

- $Q$ can be recovered under $M$ if the above mapping $M'$ exists

Theorem (APRR09)

Let $M^*$ be a maximum recovery of a mapping $M$. If $Q$ can be recovered under $M$, then $M^*$ recovers $Q$ under $M$. 
On the existence of maximum recoveries

Maximum recoveries overcome one of the limitations of Fagin-inverses and quasi-inverses.
On the existence of maximum recoveries

Maximum recoveries overcome one of the limitations of Fagin-inverses and quasi-inverses.

Theorem (APR08)

*Every mapping specified by st-tgds has a maximum recovery.*
On the existence of maximum recoveries

Maximum recoveries overcome one of the limitations of Fagin-inverses and quasi-inverses.

**Theorem (APR08)**

*Every mapping specified by st-tgds has a maximum recovery.*

**Example**

Consider a mapping \( \mathcal{M} \) specified by:

\[
P(x, y) \land P(y, z) \rightarrow R(x, z) \land T(y)
\]

\( \mathcal{M} \) has neither an inverse nor a quasi-inverse [FKPT07]. A maximum recovery of \( \mathcal{M} \) is specified by:

\[
R(x, z) \rightarrow \exists y \ P(x, y) \land P(y, z)
\]

\[
T(y) \rightarrow \exists x \exists z \ P(x, y) \land P(y, z)
\]
Maximum recoveries strictly generalize Fagin-inverses

\( \mathcal{M} \) is closed-down on the left if it satisfies the following condition:

If \( J \) is a solution for \( I_2 \) and \( I_1 \subseteq I_2 \), then \( J \) is a solution for \( I_1 \).

The notion of Fagin-inverse is defined in [F06] focusing on these mappings.
Maximum recoveries strictly generalize Fagin-inverses

\( M \) is closed-down on the left if it satisfies the following condition:

If \( J \) is a solution for \( I_2 \) and \( I_1 \subseteq I_2 \), then \( J \) is a solution for \( I_1 \)

The notion of Fagin-inverse is defined in [F06] focusing on these mappings.

**Theorem (APR08)**

*If \( M \) is closed-down on the left and Fagin-invertible: \( M^* \) is an inverse of \( M \) iff \( M^* \) is a maximum recovery of \( M \).*
Maximum recoveries strictly generalize Fagin-inverses

\( \mathcal{M} \) is closed-down on the left if it satisfies the following condition:

If \( J \) is a solution for \( I_2 \) and \( I_1 \subseteq I_2 \), then \( J \) is a solution for \( I_1 \)

The notion of Fagin-inverse is defined in [F06] focusing on these mappings.

**Theorem (APR08)**

*If \( \mathcal{M} \) is closed-down on the left and Fagin-invertible: \( \mathcal{M}^* \) is an inverse of \( \mathcal{M} \) iff \( \mathcal{M}^* \) is a maximum recovery of \( \mathcal{M} \).*

A similar theorem can be proved for the notion of quasi-inverse.
Computing maximum recoveries

The simple process of “reversing the arrows” of st-tgds does not work properly

- For example, consider mapping specified by st-tgds $A(x) \rightarrow T(x)$ and $B(x) \rightarrow T(x)$
Computing maximum recoveries

The simple process of “reversing the arrows” of st-tgds does not work properly

- For example, consider mapping specified by st-tgds
  \[ A(x) \rightarrow T(x) \] and \[ B(x) \rightarrow T(x) \]

We present an algorithm that is based on query rewriting.

- We can reuse the large body of work on query rewriting

**Definition**

Given a mapping \( \mathcal{M} \) and a target query \( Q \): Query \( Q' \) is a rewriting over the source of \( Q \) if for every source instance \( I \):

\[
\text{certain}_\mathcal{M}(Q, I) = Q'(I)
\]
Computing maximum recoveries

### Algorithm

**Input**: A mapping \( \mathcal{M} = (S, T, \Sigma) \), where \( \Sigma \) is a set of st-tgds

**Output**: A mapping \( \mathcal{M}^* = (T, S, \Sigma^*) \) that is a maximum recovery of \( \mathcal{M} \)

1. let \( \Sigma^* := \emptyset \)
2. for every \( \varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z} \psi(\bar{x}, \bar{y}) \) in \( \Sigma \) do
   - compute a first-order logic formula \( \alpha(\bar{x}) \) that is a source rewriting of \( \exists \bar{z} \psi(\bar{x}, \bar{z}) \) under \( \mathcal{M} \)
   - add dependency \( \psi(\bar{x}, \bar{z}) \land C(\bar{x}) \rightarrow \alpha(\bar{x}) \) to \( \Sigma^* \)
Theorem (APR08, APR09)

There is an exponential time algorithm that, given a mapping $M$ specified by st-tgds, computes a maximum recovery of $M$. 
Complexity of the algorithm

Theorem (APR08, APR09)

There is an exponential time algorithm that, given a mapping $\mathcal{M}$ specified by st-tgds, computes a maximum recovery of $\mathcal{M}$.

A few words about the language needed to express the maximum recovery:

- Output of the algorithm: $\text{CQ}^{\mathcal{C}(\cdot)-\text{to-UCQ}}_\approx$ dependencies
- Predicate $\mathcal{C}(\cdot)$, disjunction and equality are needed
Backup slides

- Inverse operator

- Combination of both operators
  - Key ingredient: Conditional tables
We need to combine the operators

Can we combine the composition and inverse operators?

- Is there a good language for both operators?
We need to combine the operators

Can we combine the composition and inverse operators?
  ▶ Is there a good language for both operators?

Some bad news:

**Theorem (APR11)**

*There exists a mapping specified by an SO tgd that has neither a Fagin-inverse nor a quasi-inverse nor a maximum recovery.*
We need to combine the operators

Can we combine the composition and inverse operators?

- Is there a good language for both operators?

Some bad news:

**Theorem (APR11)**

*There exists a mapping specified by an SO tgd that has neither a Fagin-inverse nor a quasi-inverse nor a maximum recovery.*

Do we need yet another notion of inverse?
We need to combine the operators

Can we combine the composition and inverse operators?
  ▶ Is there a good language for both operators?

Some bad news:

Theorem (APR11)

There exists a mapping specified by an SO tgd that has neither a Fagin-inverse nor a quasi-inverse nor a maximum recovery.

Do we need yet another notion of inverse?
  ▶ No, we need to revisit the semantics of mappings
What went wrong?

Key observation: A target instance of a mapping can be the source instance of another mapping.

- Sources instances may contain null values
What went wrong?

Key observation: A target instance of a mapping can be the source instance of another mapping.

- Sources instances may contain null values

Example

Consider a mapping $\mathcal{M}$ specified by:

\[
P(x, y) \rightarrow R(x, y) \\
P(x, x) \rightarrow T(x)
\]

The canonical universal solution for $I = \{P(n, a)\}$ under $\mathcal{M}$:

\[
J^* = \{R(n, a)\}
\]

But $J^*$ is not a good solution for $I$.

- It cannot represent the fact that if $n$ is given value $a$, then $T(a)$ should hold in the target.
A solution to the problem

We use conditional tables instead of the usual instances.

▷ What about complexity?
A solution to the problem

We use conditional tables instead of the usual instances.

▶ What about complexity?

Example

Consider again mapping $\mathcal{M}$ specified by:

\[
\begin{align*}
P(x, y) & \rightarrow R(x, y) \\
P(x, x) & \rightarrow T(x)
\end{align*}
\]

The following conditional table is a good solution for $I = \{P(n, a)\}$:

<table>
<thead>
<tr>
<th>$R(n, a)$</th>
<th>true</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T(n)$</td>
<td>$n = a$</td>
</tr>
</tbody>
</table>
Can conditional tables be used in real data exchange systems?

Good news: We just need positive conditions

- Good solutions can be computed in polynomial time (data complexity)
- Certain answers for UCQ can be computed in polynomial time (data complexity)
Can conditional tables be used in real data exchange systems?

Good news: We just need positive conditions

- Good solutions can be computed in polynomial time (data complexity)
- Certain answers for UCQ can be computed in polynomial time (data complexity)

**Theorem (APR11)**

*If instances are replaced by positive conditional tables:*

- *SO tgds are still the right language for the composition of mappings given by st-tgds*
- *Every mapping specified by an SO tgd admits a maximum recovery*


