Locally Consistent Transformations and Query Answering in Data Exchange

Marcelo Arenas	Pablo Barceló	Ronald Fagin	Leonid Libkin
U. of Toronto	U. of Toronto	IBM Almaden	U. of Toronto

Data Exchange Setting

- Data Exchange Setting: $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$

  $\mathbf{S}$: Source schema.

  $\mathbf{T}$: Target schema.

  Σ_{st}: Set of source-to-target dependencies.

  - Source-to-target dependency: FO sentence of the form

$$\forall \bar{x} \, (\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists \bar{y} \, \psi_{\mathbf{T}}(\bar{x}, \bar{y})).$$

  - $\varphi_{\mathbf{S}}(\bar{x})$: FO formula over $\mathbf{S}$.
  - $\psi_{\mathbf{T}}(\bar{x}, \bar{y})$: conjunction of FO atomic formulas over $\mathbf{T}$.

Example: Data Exchange Setting

- $\mathbf{S} = \langle Employee(\cdot) \rangle$
- $\mathbf{T} = \langle Dept(\cdot, \cdot) \rangle$
- $\Sigma_{st} = \{\forall x \; Employee(x) \rightarrow \exists y \; Dept(x, y)\}.$

LAV & GAV

- **LAV setting:** each dependency in Σ_{st} is of the form

$$S(\bar{x}) \rightarrow \exists \bar{y}\, \psi_{\mathbf{T}}(\bar{x}, \bar{y})$$

where S is a relation symbol in $\mathbf{S}$.

- **GAV setting:** each dependency in Σ_{st} is of the form

$$\varphi_{\mathbf{S}}(\bar{x}) \rightarrow T(\bar{x})$$

where T is a relation symbol in $\mathbf{T}$.

Data Exchange Problem

- Given a source instance I, find a target instance J such that (I, J) satisfies Σ_{st}.
  - J is called a solution for I.

- Previous example: Possible solutions for $I = \{Employee(peter)\}$:
  - $J_1 = \{Dept(peter, 1)\}$.
  - $J_2 = \{Dept(peter, 1), Dept(peter, 2)\}$.
  - $J_3 = \{Dept(peter, 1), Dept(john, 1)\}$.
  - $J_4 = \{Dept(peter, n_1)\}$.
  - $J_5 = \{Dept(peter, n_1), Dept(peter, n_2)\}$.

Query Answering

- Q is a query over target schema.

  What does it mean to answer Q?

$$\underline{\mathit{certain}}(Q, I) \;=\; \bigcap_{J \text{ is a solution for } I} Q(J)$$

- Previous example:

  - $\underline{\mathit{certain}}(\exists y\, \mathit{Dept}(x, y),\ I) = \{\mathit{peter}\}$.
  - $\underline{\mathit{certain}}(\mathit{Dept}(x, y),\ I) = \emptyset$.
  - $\underline{\mathit{certain}}(\exists x \exists y_1 \exists y_2\, \mathit{Dept}(x, y_1) \wedge \mathit{Dept}(x, y_2) \wedge y_1 \neq y_2,\ I) = \mathit{false}$.

Query Rewriting

- How can we compute <u>*certain*</u>(Q, I)?
  - Naïve algorithm does not work: infinitely many solutions.

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

  Look for some specific $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$, and find conditions under which <u>*certain*</u>$(Q, I) = Q'(\mathcal{F}(I))$ for every source instance I.

- What is a good alternative for $\mathcal{F}$?

Outline

- Universal solutions.
  - Canonical universal solution.
- Query rewriting over the canonical universal solution.
- Locality in data exchange.
  - Proving inexpressibility results.
- Expressibility: canonical universal solution versus core.
- Query rewriting under the universal solutions semantics.
- Final comments.

Universal Solutions

- Notation:

  Const: infinite set of constants.

  Var: infinite set of null values, disjoint from Const.

  Const(J): constants in J.

  Var(J): null values in J.

  Homomorphism $h : J \to J'$: mapping from adom(J) to adom(J') such that $h(c) = c$ for all $c \in \mathsf{Const}(J)$, and $\bar{t} \in J(R)$ implies $h(\bar{t}) \in J'(R)$.

- A universal solution for I is a solution J such that for every solution J' for I, there exists a homomorphism $h : J \to J'$.

Universal Solutions

- Possible solutions for $I = \{Employee(peter)\}$:
  - $J_1 = \{Dept(peter, 1)\}$.
  - $J_4 = \{Dept(peter, n_1)\}$.
  - $J_5 = \{Dept(peter, n_1), Dept(peter, n_2)\}$.

- J_1 is not a universal solution for I.

- J_4 is a universal solution for I:
  - From J_4 to J_1: $h(peter) = peter$ and $h(n_1) = 1$.
  - From J_4 to J_5: $h(peter) = peter$ and $h(n_1) = n_1$.
  - $\ldots$

- J_5 is also a universal solution for I.

Universal Solutions

- A universal solution is more general than an arbitrary solution: it can be homomorphically mapped into that solution.

- All universal solutions are homomorphically equivalent.

- Universal solutions always exist [FKMP03].

- We are interested in a special kind of universal solution: canonical universal solution.

Canonical Universal Solution

Input: $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and a source instance I

Output: canonical universal solution J for I

Algorithm:

> for every $\forall \bar{x}\,(\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists y\, \psi_{\mathbf{T}}(\bar{x}, \bar{y})) \in \Sigma_{st}$ do
>
> > for every $\bar{a}$ such that I satisfies $\varphi_{\mathbf{S}}(\bar{a})$ do
> >
> > > create a fresh tuple of null values $\bar{b}$
> > >
> > > insert $\psi_{\mathbf{T}}(\bar{a}, \bar{b})$ into J

Canonical Universal Solution

- Example: $\Sigma_{st} = \{\forall x\, Employee(x) \rightarrow \exists y\, Dept(x, y)\}$ and $I = \{Employee(peter),\ Employee(john)\}$.
  - For $a = peter$ do
    Create a fresh null value n_1
    Insert $Dept(peter, n_1)$ into J
  - For $a = john$ do
    Create a fresh null value n_2
    Insert $Dept(john, n_2)$ into J

Canonical universal solution:

$$\{Dept(peter, n_1),\ Dept(john, n_2)\}$$

Query Rewriting over the Canonical Universal Solution

- $\mathcal{F}_{\text{univ}}(I)$: canonical universal solution of I.
  - Can be computed in polynomial time.

- Theorem [FKMP03] For every data exchange setting and conjunctive query Q, there exists Q' such that for every source instance I, $\underline{certain}(Q, I) = Q'(\mathcal{F}_{\text{univ}}(I))$.
  - $C(x)$: holds whenever $x \in \mathsf{Const}$.
  - $Q'(x_1, \ldots, x_m) = C(x_1) \wedge \cdots \wedge C(x_m) \wedge Q(x_1, \ldots, x_m)$.

Query Rewriting over the Canonical Universal Solution

- Example: $\Sigma_{st} = \{\forall x\, Employee(x) \to \exists y\, Dept(x, y)\}$,
  $I = \{Employee(peter),\ Employee(john)\}$ and
  $J = \{Dept(peter, n_1),\ Dept(john, n_2)\}$

Query	:	$Q(x, y) = \exists y\, Dept(x, y)$
		$\underline{certain}(Q, I) = \{peter, john\}$
Rewriting	:	$Q'(x, y) = C(x) \wedge \exists y\, Dept(x, y)$
		$Q'(J) = \{peter, john\}$

Query Rewriting over the Canonical Universal Solution

- Can the theorem be extended to other classes of queries?

  Theorem [FKMP03] There exists a data exchange setting and a conjunctive query Q with one inequality such that Q is not FO-rewritable over $\mathcal{F}_{\text{univ}}$.

  - For every FO query Q', there exists an instance I such that <u>*certain*</u>$(Q, I) \neq Q'(\mathcal{F}_{\text{univ}}(I))$.

- How can we prove this theorem?
  - How can we prove inexpressibility results in data exchange?
  - Can we find "simple" proofs?

- This resembles the problem of proving inexpressibility results in relational databases.

Proving Inexpressibility Results: Idea

- Find a nontrivial property $\mathcal{P}$ that every FO-rewritable query over $\mathcal{F}_{\text{univ}}$ satisfies.
  - $\mathcal{P}$ should be as close as possible to the class of FO-rewritable queries.
  - In our scenario: locality.

- If Q does not satisfy $\mathcal{P}$, then Q is not FO-rewritable.

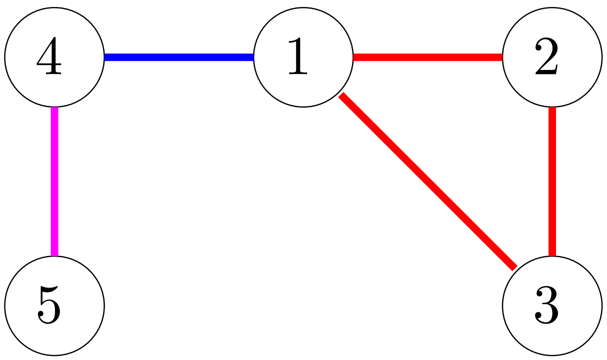
Locality in Data Exchange: Notation

I is an instance of source schema $\mathbf{S}$.

- Gaifman graph $\mathcal{G}(I)$ of an instance I:
  - $\text{adom}(I)$ is the set of nodes of $\mathcal{G}(I)$.
  - There exists an edge between a and b iff a and b belong to the same tuple of a relation in I.

- Example: $I(R) = \{(1,2,3)\}$ and $I(T) = \{(1,4), (4,5)\}$.

$\mathcal{G}(I)$:



Locality in Data Exchange: Notation

- $d_I(a, b)$: distance between a and b in $\mathcal{G}(I)$.
  - Previous example: $d_I(1, 2) = 1$ and $d_I(2, 4) = 2$.

- $d_I(\bar{a}, b)$: minimum value of $d_I(a, b)$, where a is in $\bar{a}$.

- $N_d^I(\bar{a})$: restriction of I to the elements at distance at most d from $\bar{a}$.
  - Example: $\mathrm{adom}(N_2^I(5)) = \{1, 4, 5\}$, $N_2^I(5)(R) = \emptyset$ and $N_2^I(5)(T) = \{(1, 4), (4, 5)\}$.

- $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$: members of $\bar{a}$ and $\bar{b}$ are treated as distinguished elements.
  - $\bar{a} = (a_1, \ldots, a_m)$ and $\bar{b} = (b_1, \ldots, b_m)$.
  - There is an isomorphism $f : N_d^I(\bar{a}) \to N_d^I(\bar{b})$ such that $f(a_i) = b_i$ $(1 \leq i \leq m)$.

Locality in Data Exchange: Definition

Data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, Q is m-ary query over $\mathbf{T}$.

Definition Q is **locally source-dependent** if there is $d \geq 0$ such that for every instance I of $\mathbf{S}$ and m-tuples $\bar{a}, \bar{b}$ in I,

$$N_d^I(\bar{a}) \cong N_d^I(\bar{b}) \quad \Longrightarrow \quad \begin{array}{c} \bar{a} \in \underline{certain}(Q, I) \\ \text{iff} \\ \bar{b} \in \underline{certain}(Q, I) \end{array}$$

Locality in Data Exchange: Main Theorem

Theorem If Q is FO-rewritable over the canonical universal solution, then Q is locally source-dependent.

This theorem can be used to prove inexpressibility results.

- If a query is not locally source-dependent, then it is not FO-rewritable.

Example

Data exchange setting:

$$
\begin{aligned}
\mathbf{S} &= \langle G(\cdot,\cdot),\ R(\cdot),\ S(\cdot)\rangle \\
\mathbf{T} &= \langle G'(\cdot,\cdot),\ R'(\cdot),\ S'(\cdot)\rangle \\
\Sigma_{st} &= \forall x \forall y\, G(x,y) \rightarrow G'(x,y), \\
&\quad\ \forall x\, R(x) \rightarrow R'(x), \\
&\quad\ \forall x\, S(x) \rightarrow S'(x).
\end{aligned}
$$

Query:

$$
Q(x) \quad = \quad R'(x) \quad \vee \quad S'(x) \wedge \exists y \exists z (R'(y) \wedge G'(y,z) \wedge \neg R'(z))
$$

Example: Proving Inexpressibility

- Assume that Q is FO-rewritable over the canonical universal solution.

  Then there exists $d \geq 0$ such that

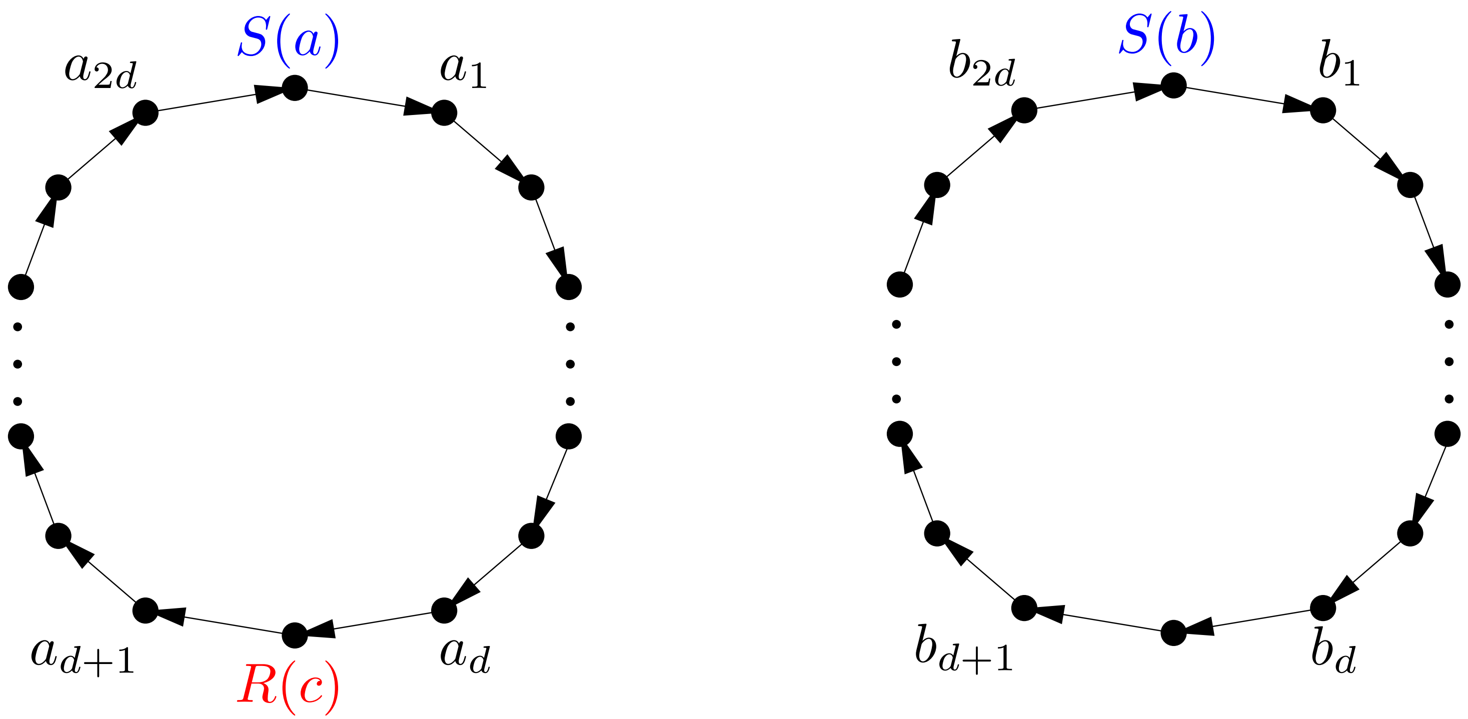
  $N_d^I(a) \cong N_d^I(b) \implies a \in \underline{certain}(Q, I)$ iff $b \in \underline{certain}(Q, I)$.

- Contradiction: find a source instance I such that

  $N_d^I(a) \cong N_d^I(b)$, $a \in \underline{certain}(Q, I)$ and $b \notin \underline{certain}(Q, I)$.

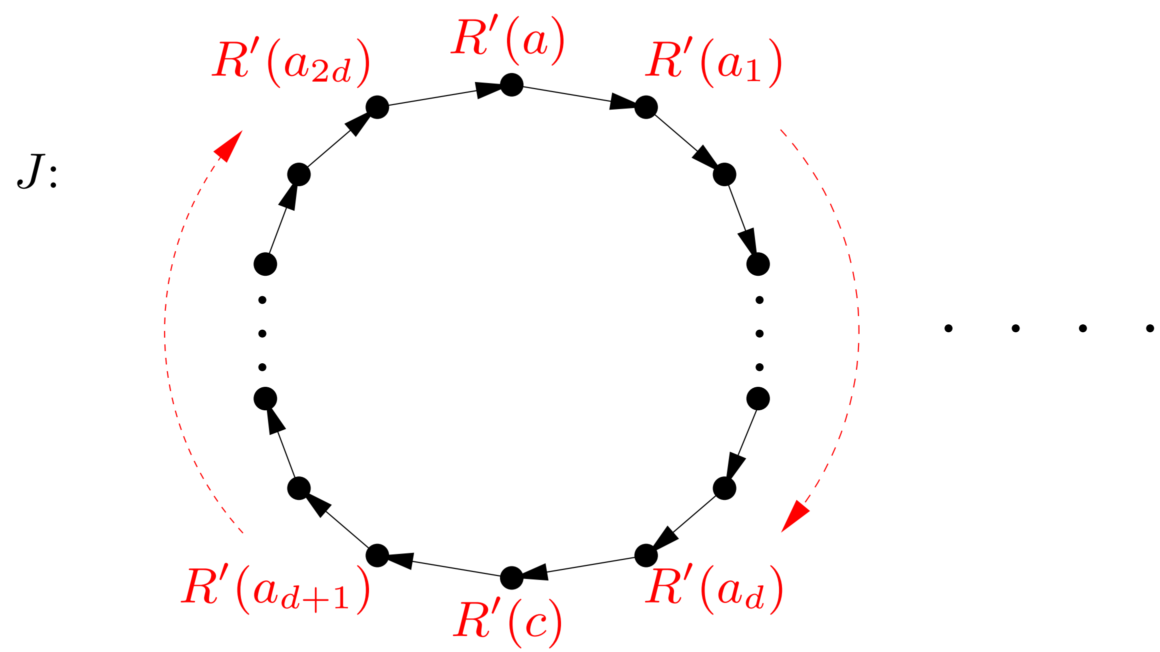
Example: Defining Instance I

I:



Example: $a \in \underline{certain}(Q, I)$

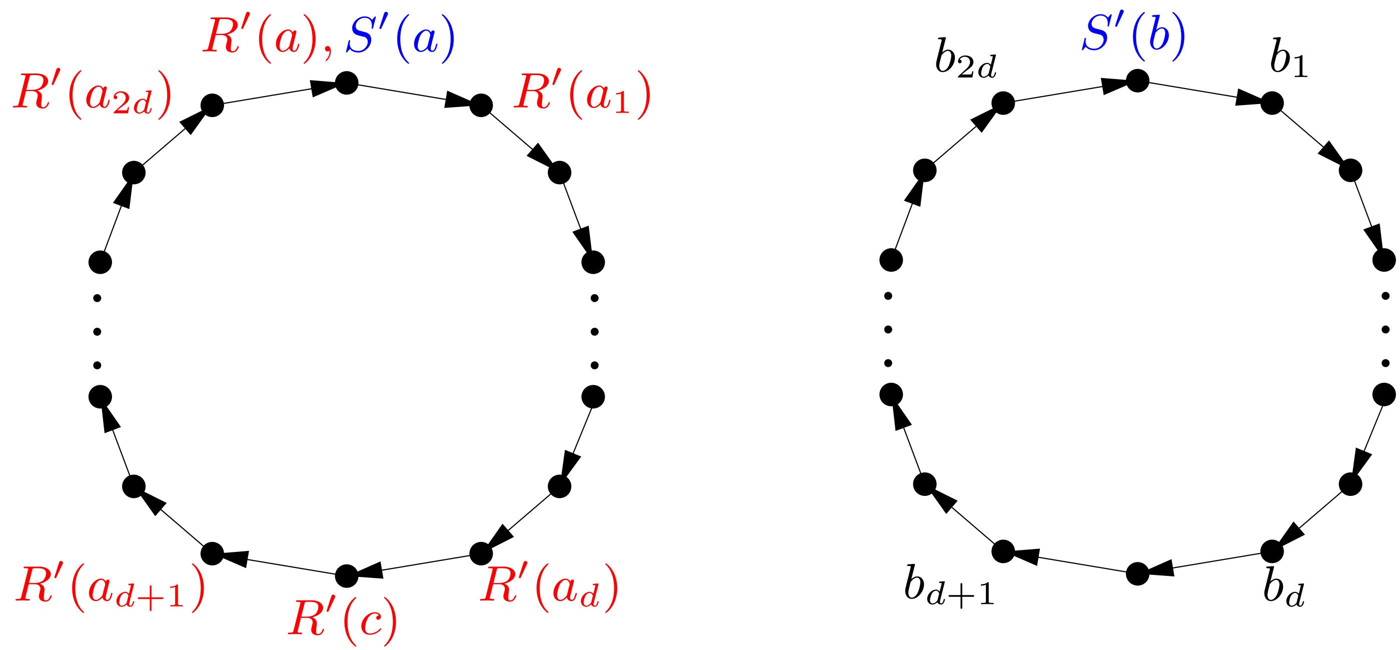
J does not satisfy $S'(a) \wedge \exists y \exists z (R'(y) \wedge G'(y,z) \wedge \neg R'(z))$:



Then: J satisfies $R'(a)$.

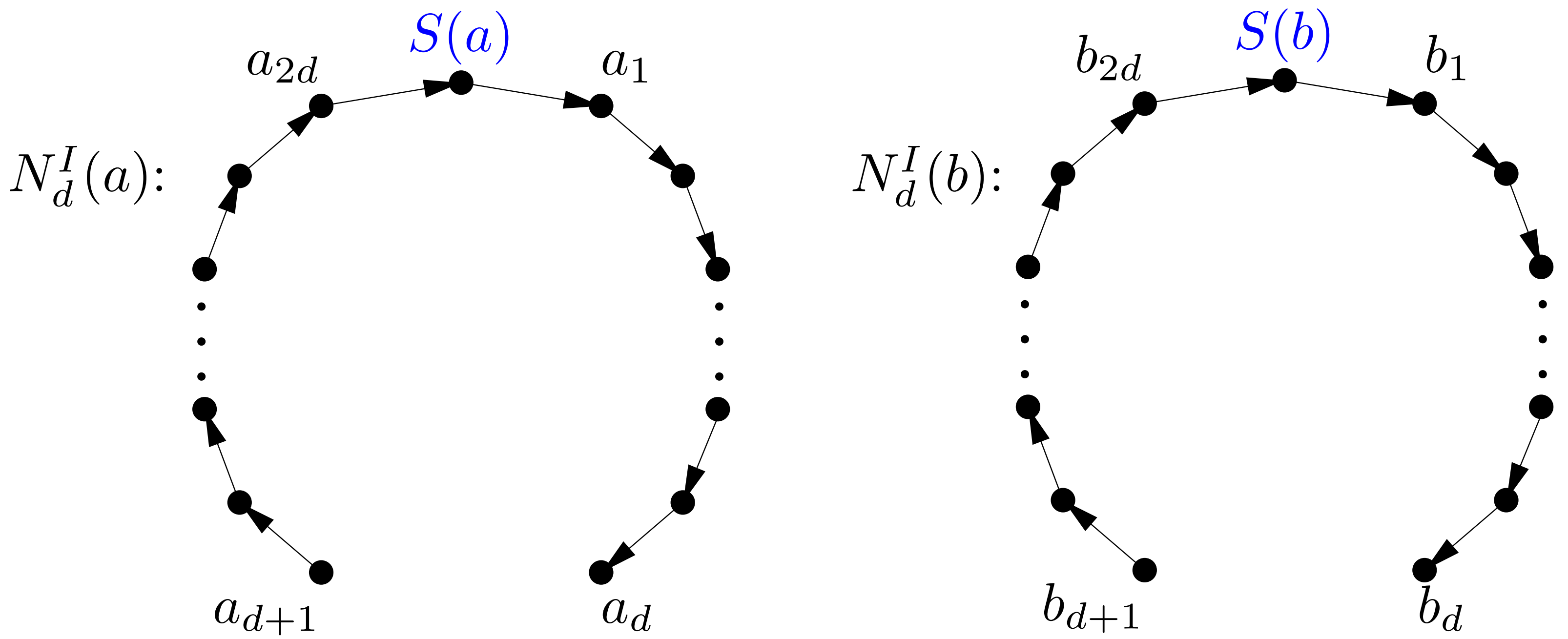
Example: $b \notin \underline{certain}(Q, I)$

J:



J does not satisfy $R'(b) \;\vee\; S'(b) \wedge \exists y \exists z (R'(y) \wedge G'(y, z) \wedge \neg R'(z))$.

Example: Getting a Contradiction



Conclusion: Q is **not** FO-rewritable over the canonical universal solution.

What about other Transformations?

- Universal solutions need not be isomorphic.
  - Decision to choose one is somewhat arbitrary.

- Core of a universal solution J: subinstance J^* of J such that there is a homomorphism from J to J^*, but there is no homomorphism from J to a proper subinstance of J^*.

- Every universal solution has the same core.

- Core is itself a universal solution.
  - It is the smallest universal solution.

- Core can be computed in polynomial time [FKP03].

Example: Core

- Setting: $\mathbf{S} = \langle Employee(\cdot) \rangle$, $\mathbf{T} = \langle Dept(\cdot, \cdot) \rangle$ and $\Sigma_{st} = \{\forall x \; Employee(x) \rightarrow \exists y \; Dept(x, y)\}$.

- Source instance: $I = \{Employee(peter)\}$.

  Universal solutions:

  - $\{Dept(peter, n_1)\}$.
  - $\{Dept(peter, n_1), Dept(peter, n_2)\}$.
  - $\ldots$

- Core: $\{Dept(peter, n_1)\}$.

Query Rewriting over the Core

- $\mathcal{F}_{\text{core}}(I)$: core of the canonical universal solution for I.

- Theorem [FKMP03] For every data exchange setting and conjunctive query Q, there exists Q' such that for every source instance I, $\underline{\textit{certain}}(Q, I) = Q'(\mathcal{F}_{\text{core}}(I))$.
  - Certain answers for conjunctive queries can be computed more efficiently by using the core.

- Rewritability over the core: Can we use locality?

Expressibility: Canonical Universal Solution versus Core

Theorem If Q is FO-rewritable over the core, then Q is also FO-rewritable over the canonical universal solution.

- There is a cubic-time algorithm that, given a rewriting of Q over the core, finds a rewriting of Q over the canonical universal solution.

Corollary If Q is FO-rewritable over the core, then Q is locally source-dependent.

Theorem There exists an FO query that is FO-rewritable over the canonical universal solution, but not FO-rewritable over the core.

What about other Semantics?

- Usual certain answers semantics sometimes exhibit counterintuitive behavior.
  - For every Boolean query Q, either $\underline{\mathit{certain}}(Q, I) = \mathit{false}$ for all instances I, or $\underline{\mathit{certain}}(\neg Q, I) = \mathit{false}$ for all instances I.

- May be more meaningful to consider semantics based on universal solutions:

$$\underline{\mathit{u\text{-}certain}}(Q, I) = \bigcap_{J \text{ is a universal solution for } I} Q(J).$$

Query Rewriting under the Universal Solutions Semantics

- Given query Q, we want to find Q' such that $\underline{\textit{u-certain}}(Q, I) = Q'(\mathcal{F}(I))$ for every source instance I.

- Theorem [FKP03] For every data exchange setting and existential query Q, there exists Q' such that for every source instance I, $\underline{\textit{u-certain}}(Q, I) = Q'(\mathcal{F}_{\text{core}}(I))$.

- **Definition** Q is locally source-dependent under the universal solution semantics if there is $d \geq 0$ such that:

$$N_d^I(\bar{a}) \cong N_d^I(\bar{b}) \quad \Longrightarrow \quad \begin{array}{c} \bar{a} \in \underline{\textit{u-certain}}(Q, I) \\ \text{iff} \\ \bar{b} \in \underline{\textit{u-certain}}(Q, I) \end{array}$$

- **Theorem** All the previous results hold for the universal solution semantics.
  - If Q is FO-rewritable over the canonical universal solution (core) under the universal solutions semantics, then Q is locally source-dependent under the universal solutions semantics.

Final Comments

- Previous results can be extended to data exchange settings where the underlying language for both source-to-target dependencies and queries correspond to SQL `select-from-where-groupby-having` statements.

- Previous results cannot be extended to data exchange settings containing target dependencies.
  - Except for GAV+egd.

Locally Consistent Transformations

- To solve the query rewriting problem we need to understand how neighborhoods are transformed when computing target instances.

- **Theorem** In a LAV setting, for every $m, d \geq 0$ there exists $d' \geq 0$ such that, for every instance I of $\mathbf{S}$ and m-tuples $\bar{a}, \bar{b}$ in I,

$$N_{d'}^{I}(\bar{a}) \cong N_{d'}^{I}(\bar{b}) \implies N_{d}^{\mathcal{F}_{\text{univ}}(I)}(\bar{a}) \cong N_{d}^{\mathcal{F}_{\text{univ}}(I)}(\bar{b}).$$

Locally Consistent Transformations

- **Corollary** In a LAV setting, every query that is FO-rewritable over the canonical universal solution is locally source-dependent.

- This result does not hold for GAV settings.
  - To prove the general theorem we study a notion of locality based on FO-logical equivalence.