Locality of Queries and Transformations

Marcelo Arenas
Pontificia Universidad Católica de Chile
and
Center for Web Research

Joint work with Pablo Barceló, Ron Fagin and Leonid Libkin

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

The Problem of Data Exchange

- Given: A source schema S, a target schema T and a specification Σ of the relationship between these schemas.

- Data exchange: Problem of finding 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 Σ and T.

  - It should be efficiently computable.

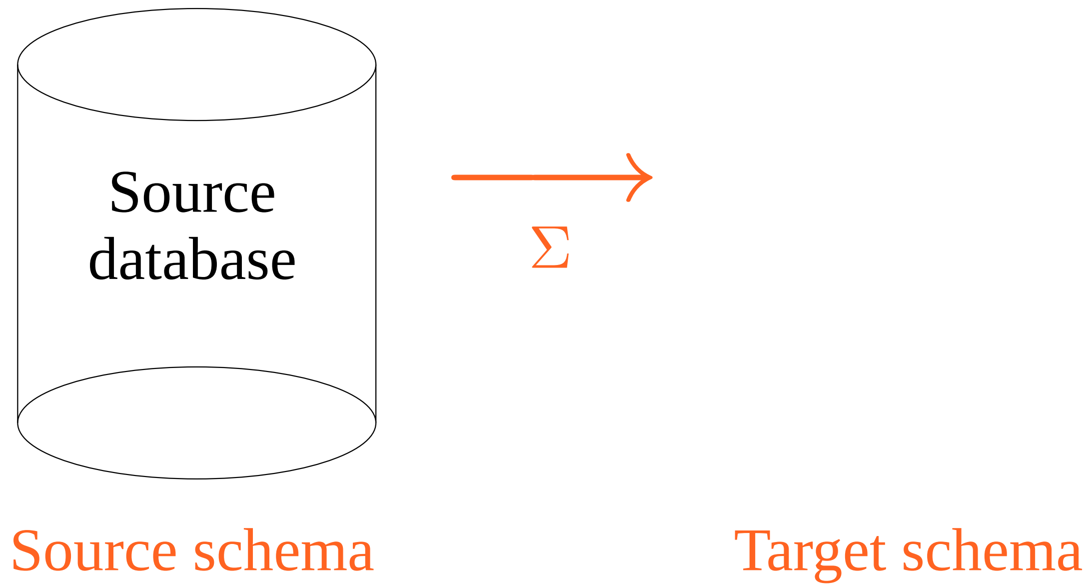
  - It should allow one to evaluate queries on the target in a way that is semantically consistent with the source data.

Data Exchange

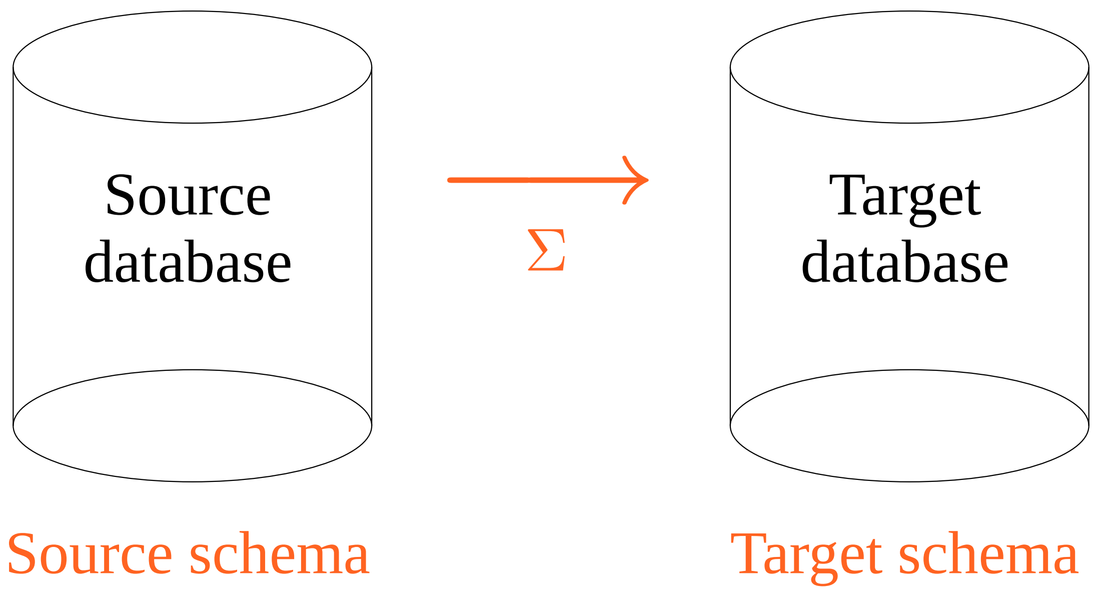
$$\xrightarrow[\Sigma]{}$$

Source schema Target schema

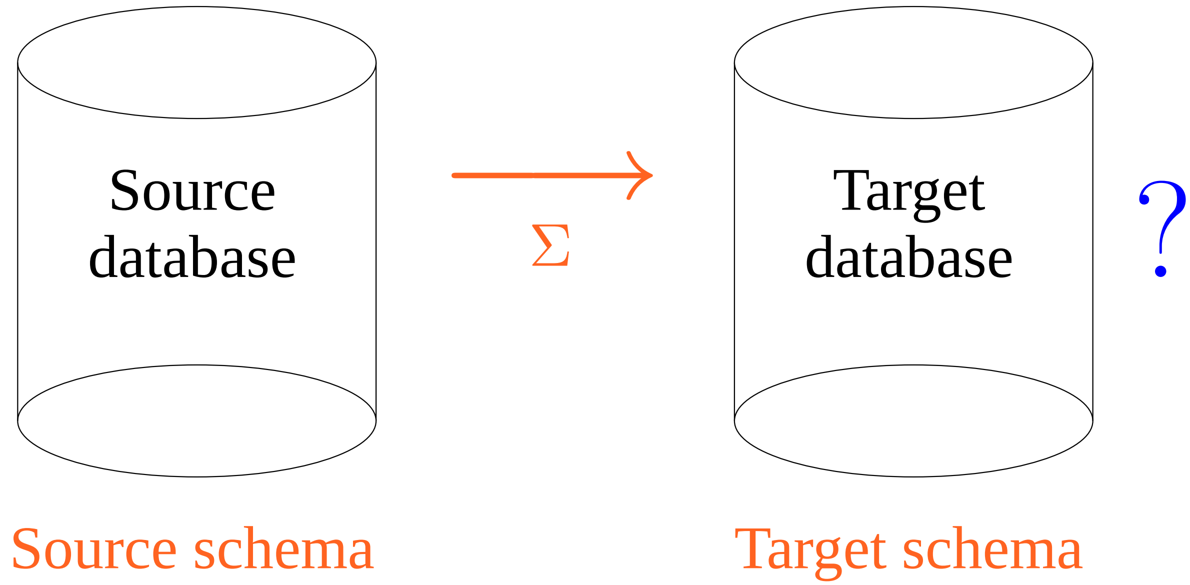
Data Exchange



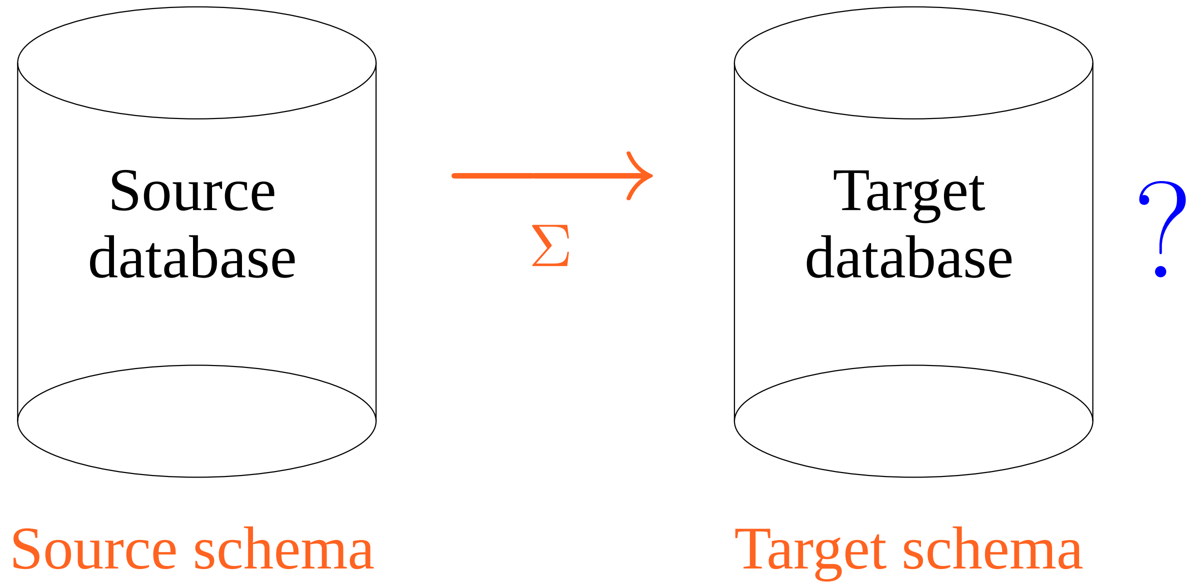
Data Exchange



Data Exchange



Data Exchange



Query over the target: Q

Answer to Q in the target instance should represent the answer to Q in the space of possible translations of the source instance.

Data Exchange in Relational Databases

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

- Relational data exchange settings:
  - Source and target schemas: Relational schemas.
  - Relationship between source and target schemas: Source-to-target dependencies.

- Semantics of data exchange has been precisely defined.
  - Algorithms for materializing target instances and for answering queries over the target have been developed.

Data exchange settings

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 $\mathbf{T}$.

Data exchange settings: Example

$$\mathbf{S} = \langle Employee(\cdot) \rangle$$

$$\mathbf{T} = \langle Dept(\cdot, \cdot) \rangle$$

$$\Sigma_{st} = \{\forall x\, (Employee(x) \to \exists y\, Dept(x, y))\}.$$

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.

Example: Possible solutions for $I = \{Employee(peter)\}$:

- $J_1 = \{Dept(peter, 1)\}$.
- $J_2 = \{Dept(peter, 1), Dept(peter, 2)\}$.
- $J_3 = \{Dept(peter, 1), Dept(john, 1)\}$.
- $J_4 = \{Dept(peter, X)\}$.
- $J_5 = \{Dept(peter, X), Dept(peter, Y)\}$.

Query answering

Q: Query over the target schema.

- What does it mean to answer Q?

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

Example:

- $\underline{\textit{certain}}(\exists y\; Dept(x, y),\; I) = \{peter\}$.
- $\underline{\textit{certain}}(Dept(x, y),\; I) = \emptyset$.

Query rewriting

How can we compute $\underline{\textit{certain}}(Q, I)$?

- Naïve algorithm does not work: infinitely many solutions.

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

Look for some specific $\mathcal{F} : \text{inst}(\mathbf{S}) \to \text{inst}(\mathbf{T})$, and find conditions under which $\underline{\textit{certain}}(Q, I) = Q'(\mathcal{F}(I))$ for every source instance I.

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

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

Canonical solution

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

Output: Canonical solution J for I

Algorithm:

for every $\forall \bar{x}\, (\varphi_{\mathbf{S}}(\bar{x}) \longrightarrow \exists \bar{y}\, \psi_{\mathbf{T}}(\bar{x}, \bar{y})) \in \Sigma_{st}$ do

for every $\bar{a}$ such that I satisfies $\varphi_{\mathbf{S}}(\bar{a})$ do

create a fresh tuple of null values $\overline{Y}$

insert $\psi_{\mathbf{T}}(\bar{a}, \overline{Y})$ into J

Canonical solution: Example

$\Sigma_{st} = \{\forall x\, (Employee(x) \to \exists y\, Dept(x, y))\}$ and
$I = \{Employee(peter), Employee(john)\}$.

- For $a = peter$ do

  Create a fresh null value X

  Insert $Dept(peter, X)$ into J

- For $a = john$ do

  Create a fresh null value Y

  Insert $Dept(john, Y)$ into J

Canonical solution:

$$\{Dept(peter, X),\ Dept(john, Y)\}$$

Query rewriting over the canonical solution

$\mathcal{F}_{\text{can}}(I)$: canonical solution for I.

- Can be computed in polynomial time (data complexity).

Theorem [FKMP03]: For every data exchange setting and union of conjunctive queries Q, there exists Q' such that for every source instance I, *certain*$(Q, I) = Q'(\mathcal{F}_{\text{can}}(I))$.

- $C(x)$: holds whenever x is a constant.
- $Q'(x_1, \ldots, x_m) = C(x_1) \wedge \cdots \wedge C(x_m) \wedge Q(x_1, \ldots, x_m)$.

Query Rewriting over the Canonical Universal Solution

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

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

Query rewriting over the canonical 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 $\mathcal{F}_{\mathrm{can}}$.

- For every FO query Q', there exists an instance I such that $\underline{\textit{certain}}(Q, I) \neq Q'(\mathcal{F}_{\mathrm{can}}(I))$.

We would like to study the query rewriting problem.

- We need some tools: How can we prove that a query is not FO-rewritable?

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

Query rewriting: Some facts

The problem of deciding whether an FO formula is FO-rewritable over $\mathcal{F}_{\mathrm{can}}$ is undecidable.

There exists other classes of queries that are FO-rewritable over the canonical solution.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

Proving Inexpressibility Results in Relational Databases

- Given: Relation schema $S(\cdot, \cdot)$

- Well known result: transitive closure of S is not expressible in relational algebra (FO).

- How do we prove this?

Locality of Queries: Notation

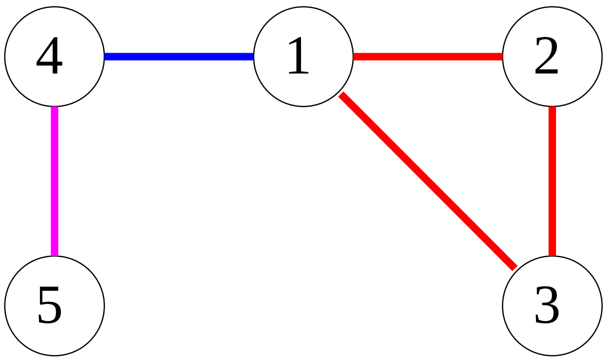
I: source instance.

Gaifman graph $\mathcal{G}(I)$ of I:

- dom(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 of Queries: Notation

$d_I(a, b)$: distance between a and b in $\mathcal{G}(I)$.

$d_I(\bar{a}, b)$: 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: $\text{dom}(N_2^I(5)) = \{1, 4, 5\}$, $N_2^I(5)(R) = \emptyset$ 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 of Queries: Gaifman Theorem

Theorem [G81] For every FO query Q, there exists $d \geq 0$ such that for every instance I and tuples $\bar{a}$, $\bar{b}$ in I,

$$N_d^I(\bar{a}) \cong N_d^I(\bar{b}) \quad \Longrightarrow \quad \bar{a} \in Q(I) \text{ iff } \bar{b} \in Q(I).$$

This theorem can be used to prove inexpressibility results.

- If a query is not "local", then it is not FO-expressible.

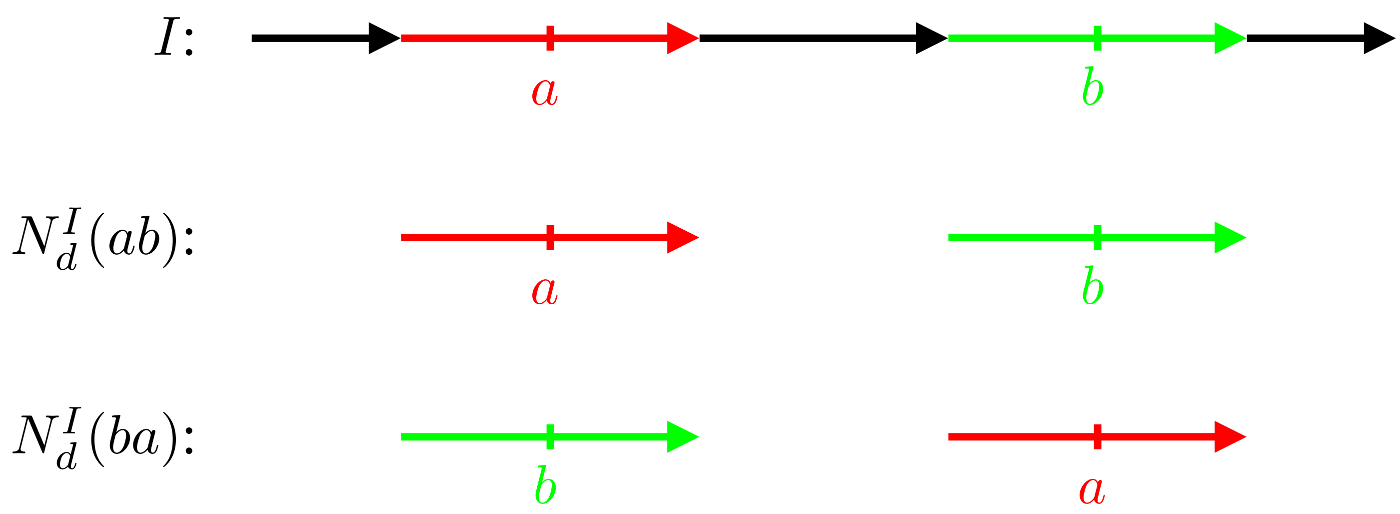
Proving Inexpressibility: Example

Assume the transitive closure of $S(\cdot,\cdot)$ is expressible in FO.

Then there is $d \geq 0$ such that:

$$N_d^I(ab) \cong N_d^I(cd) \quad \Longrightarrow \quad \begin{array}{c} (a,b) \text{ is in the transitive closure of } S \\ \text{iff} \\ (c,d) \text{ is in the transitive closure of } S \end{array}$$

Proving Inexpressibility: Example



Contradiction: by Gaifman's Theorem, (a, b) and (b, a) are in the transitive closure of S.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

Locality in data exchange: Definition

Given: $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and query Q 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 tuples $\bar{a}, \bar{b}$ in I,

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

Locality in data exchange: Main theorem

Theorem: If Q is FO-rewritable over the canonical 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: Proving inexpressibility

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:

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

Example: Proving inexpressibility

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

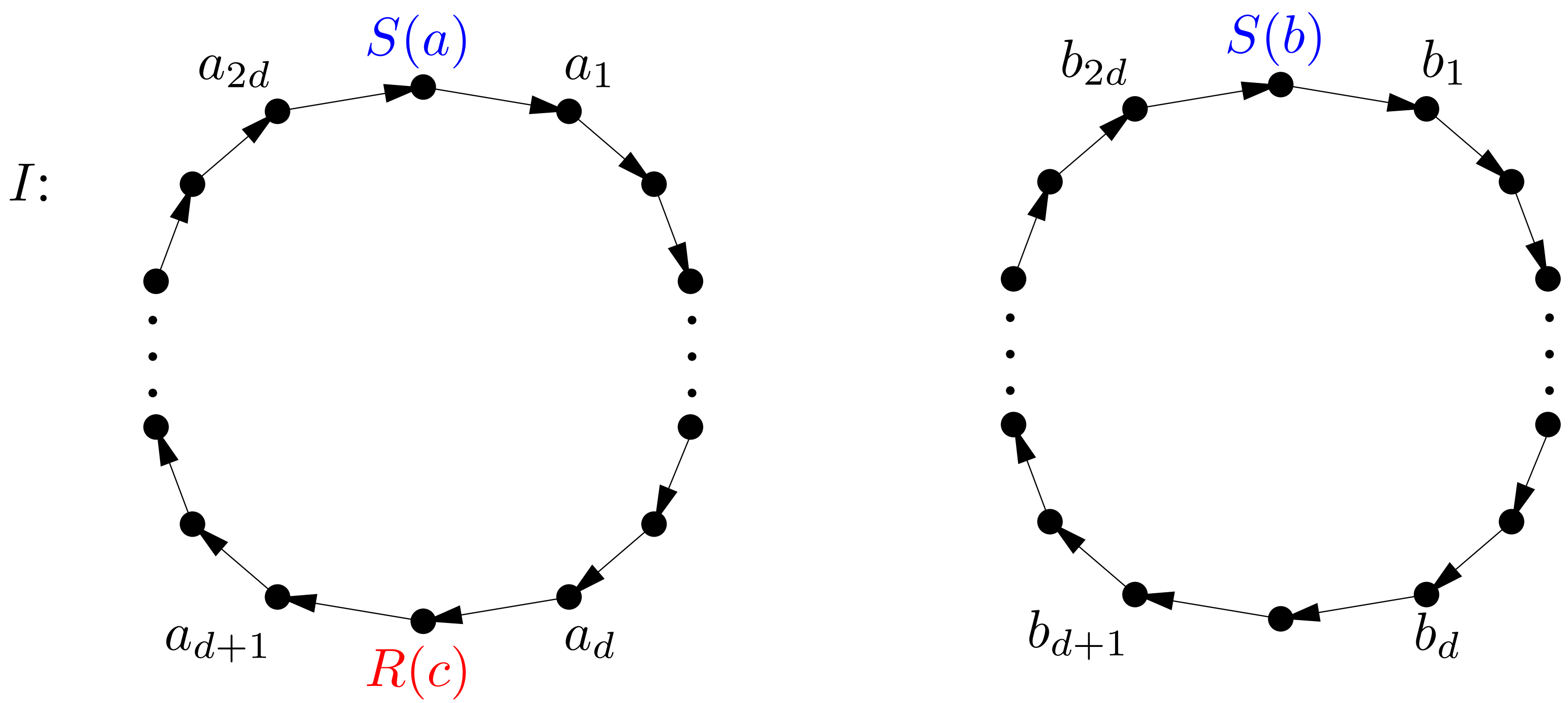
Then there exists $d \geq 0$ such that

$$N_d^I(a) \cong N_d^I(b) \quad \Longrightarrow \quad a \in \underline{\textit{certain}}(Q, I) \text{ iff } b \in \underline{\textit{certain}}(Q, I).$$

Contradiction: Find a source instance I such that

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

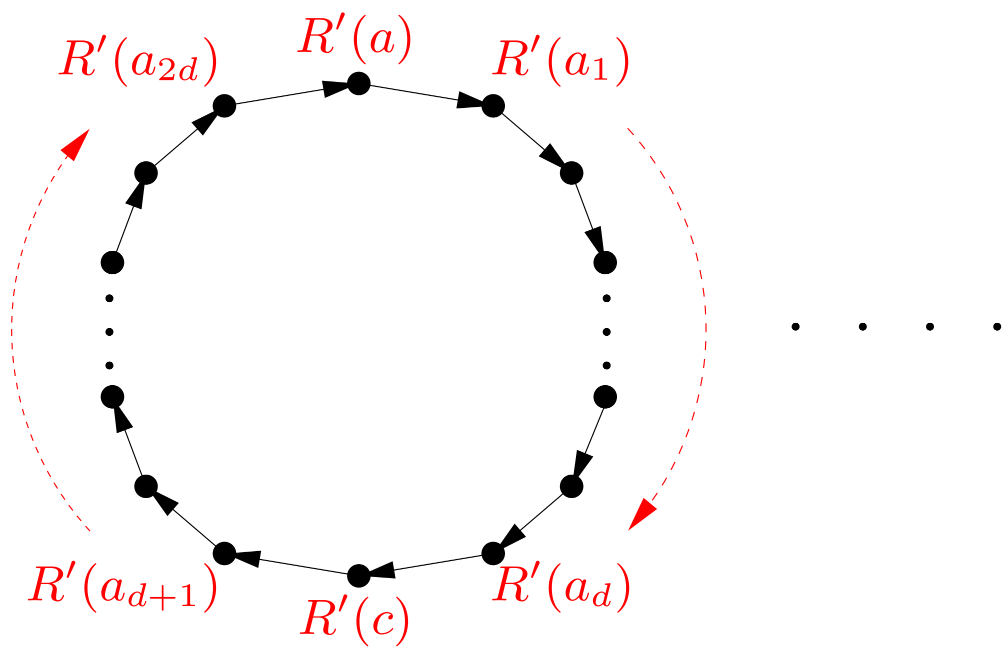
Example: Defining instance I



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

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

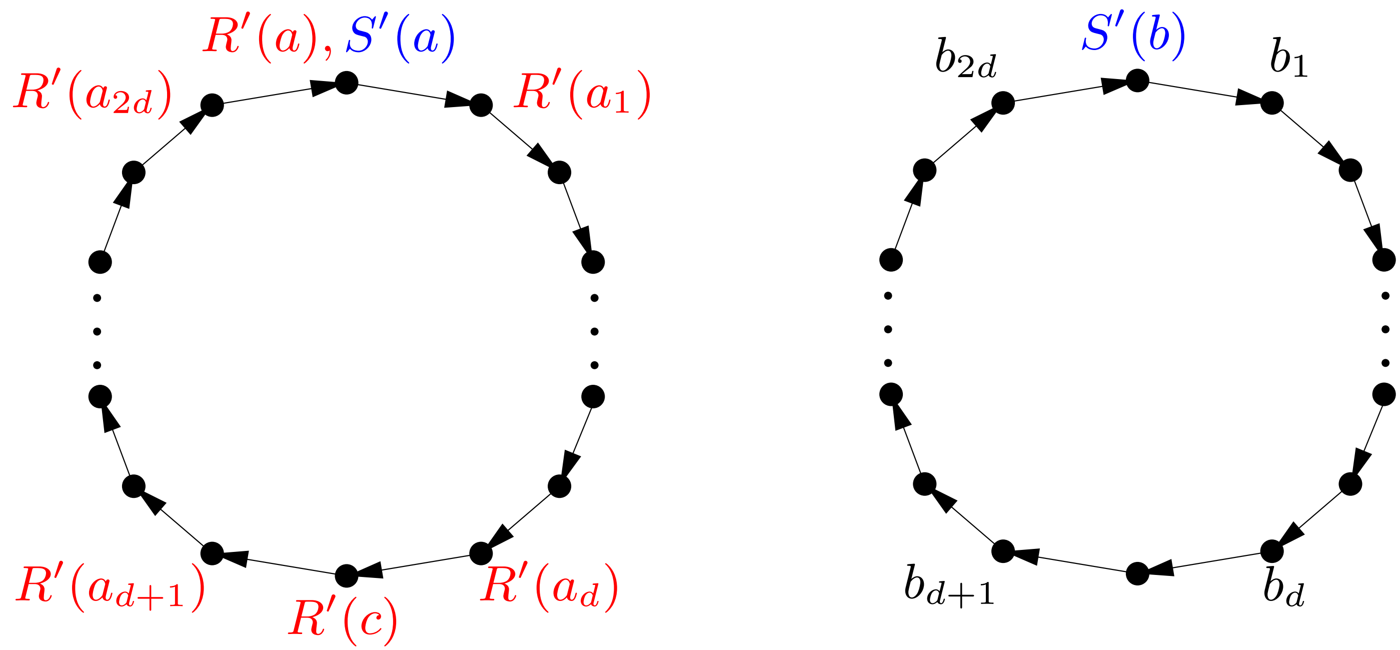
J:



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

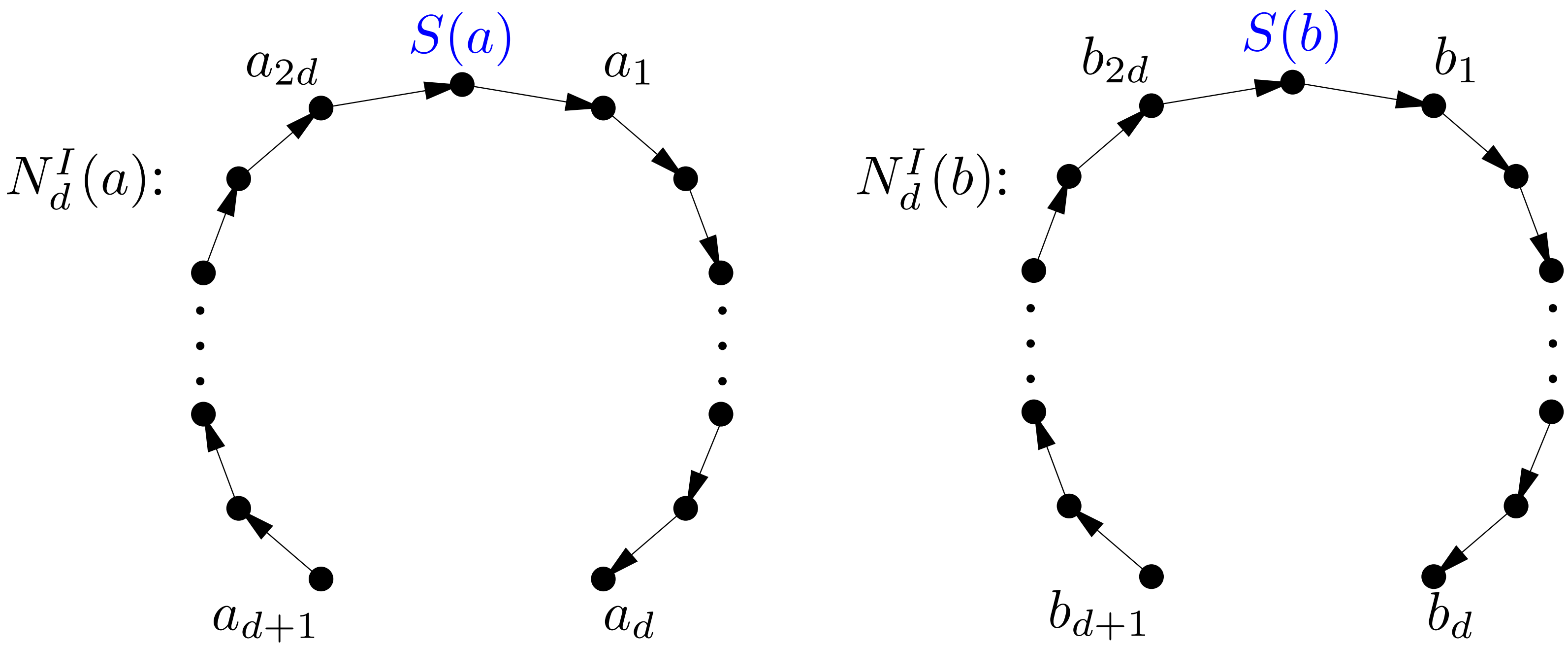
Example: $b \notin \underline{certain}(Q, I)$

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 solution.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

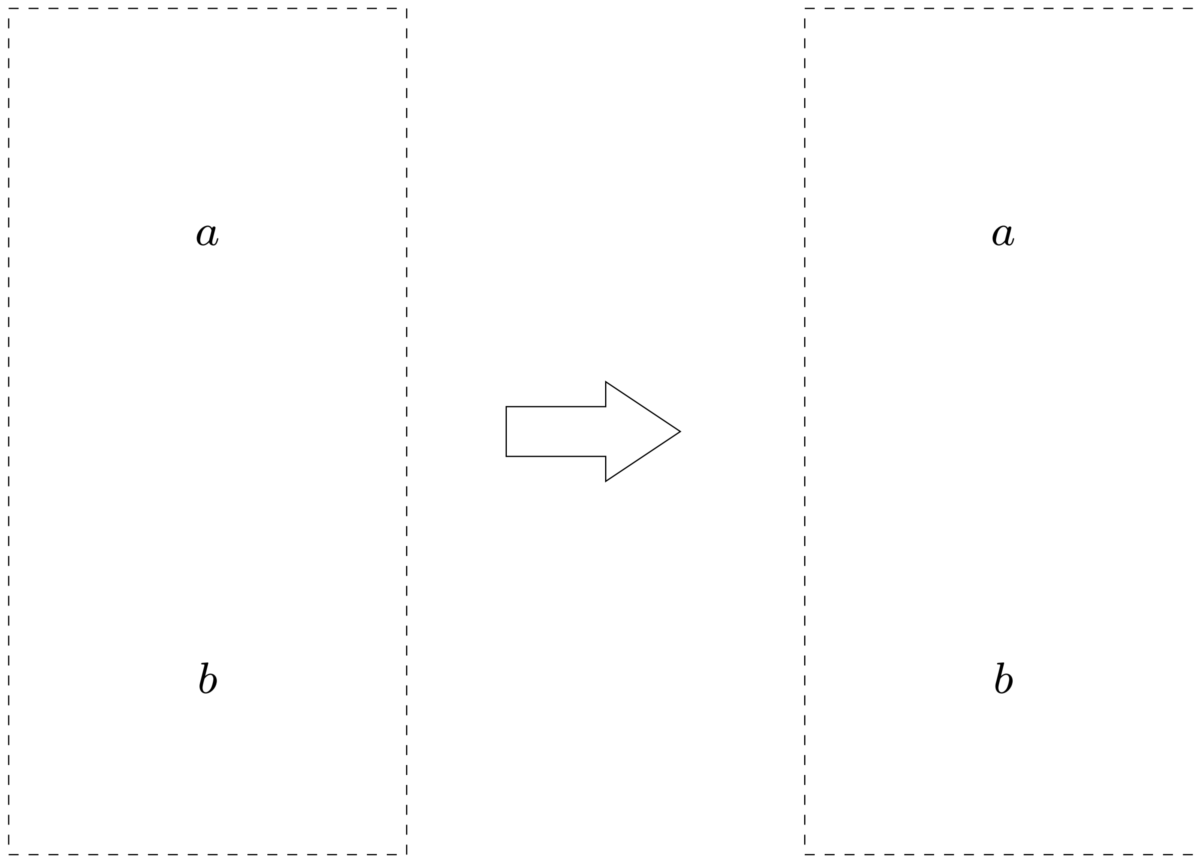
What is new?

Locality in data exchange: Isomorphic neighborhoods in the source and queries over the target.

- We cannot directly apply Gaifman's Theorem.

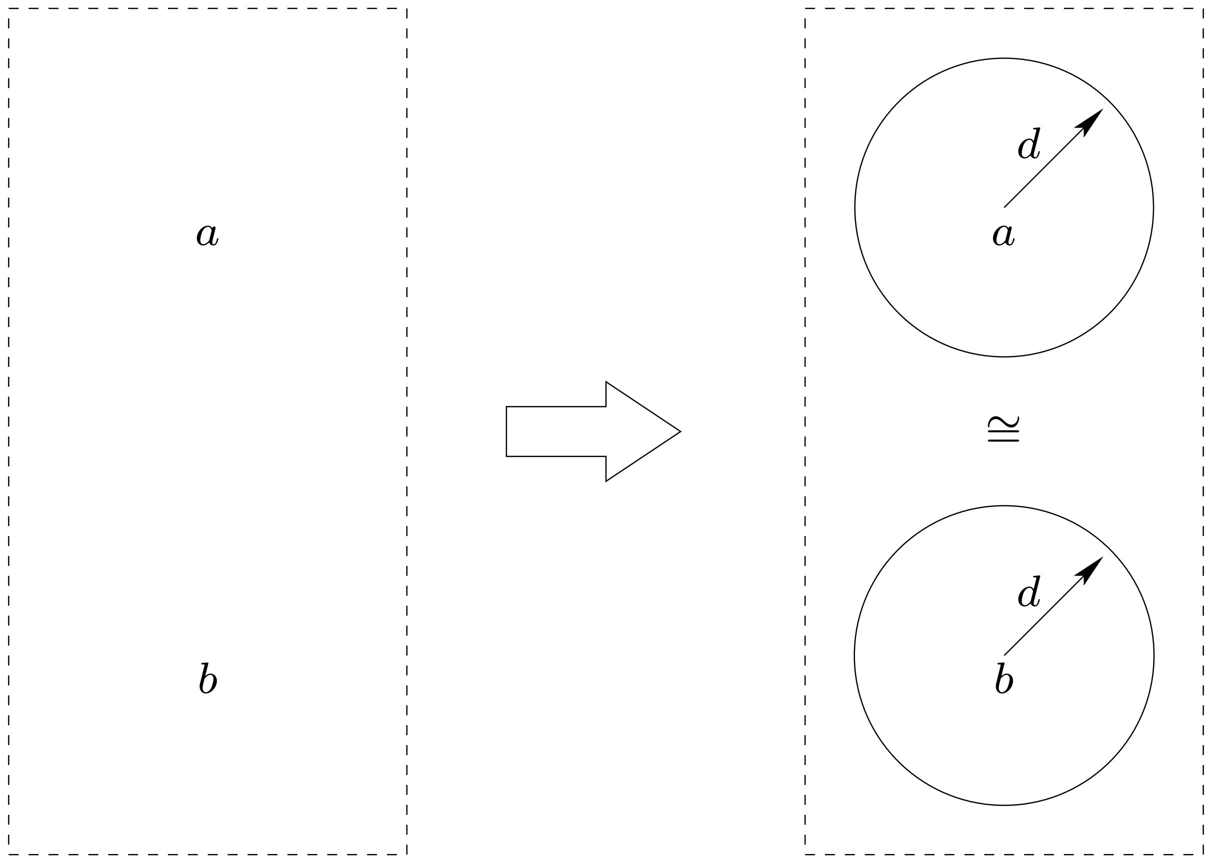
We need to introduce notions of locality for transformations.

Locality of transformations under isomorphism



source canonical solution

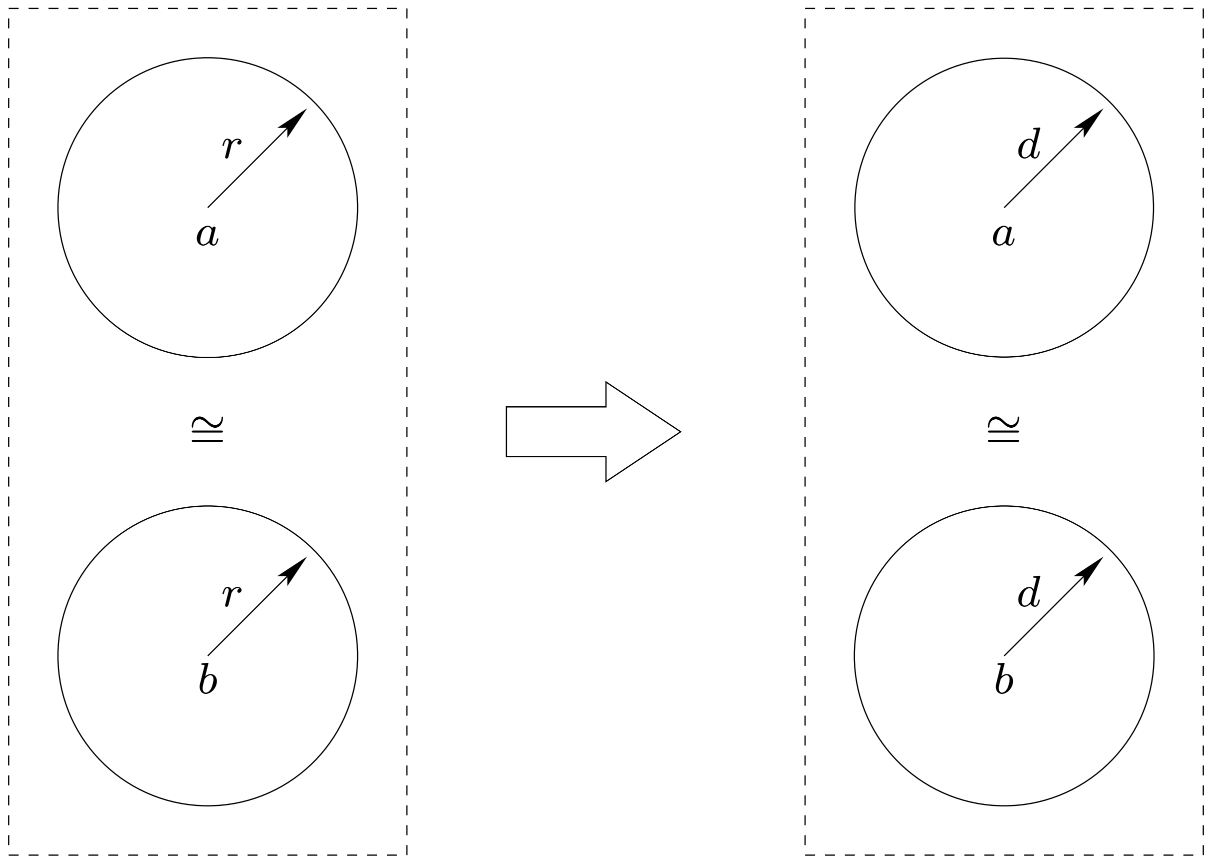
Locality of transformations under isomorphism



source

canonical solution

Locality of transformations under isomorphism



source

canonical solution

Locality of transformations under isomorphism

Locality of a transformation under isomorphism: For every $d \geq 0$ there exists $r \geq 0$ such that, for every instance I of $\mathbf{S}$ and tuples $\bar{a}, \bar{b}$ in I,

$$N_r^I(\bar{a}) \cong N_r^I(\bar{b}) \implies N_d^{\mathcal{F}_{\mathrm{can}}(I)}(\bar{a}) \cong N_d^{\mathcal{F}_{\mathrm{can}}(I)}(\bar{b}).$$

There exist classes of settings where this notion of locality holds.

- LAV setting: each dependency in Σ_{st} is of the form $S(\bar{x}) \to \exists \bar{y}\, \psi_{\mathbf{T}}(\bar{x}, \bar{y})$.

But in general ...

Locality of transformations under isomorphism

Σ_{st}:

$$\forall x \forall y \, (E(x,y) \;\rightarrow\; R(x,y))$$

$$\forall x \forall y \forall z \, (C(x) \wedge E(y,z) \;\rightarrow\; R(y,x) \wedge R(z,x))$$

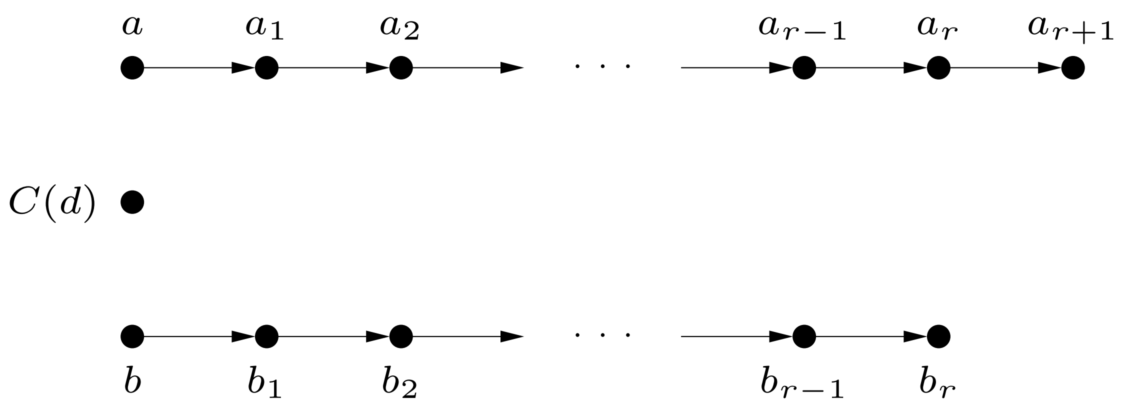
Assume $\mathcal{F}_{\mathrm{can}}$ is local under isomorphism for this setting.

Then there exists $r \geq 0$ such that, for every instance I of $\mathbf{S}$ and a, b in I,

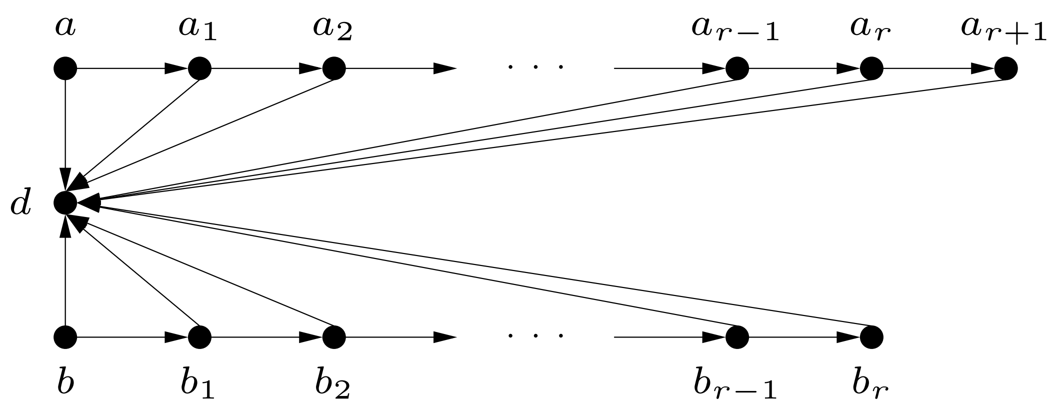
$$N_r^I(a) \cong N_r^I(b) \;\Longrightarrow\; N_2^{\mathcal{F}_{\mathrm{can}}(I)}(a) \cong N_2^{\mathcal{F}_{\mathrm{can}}(I)}(b).$$

Locality of transformations under isomorphism

Source:

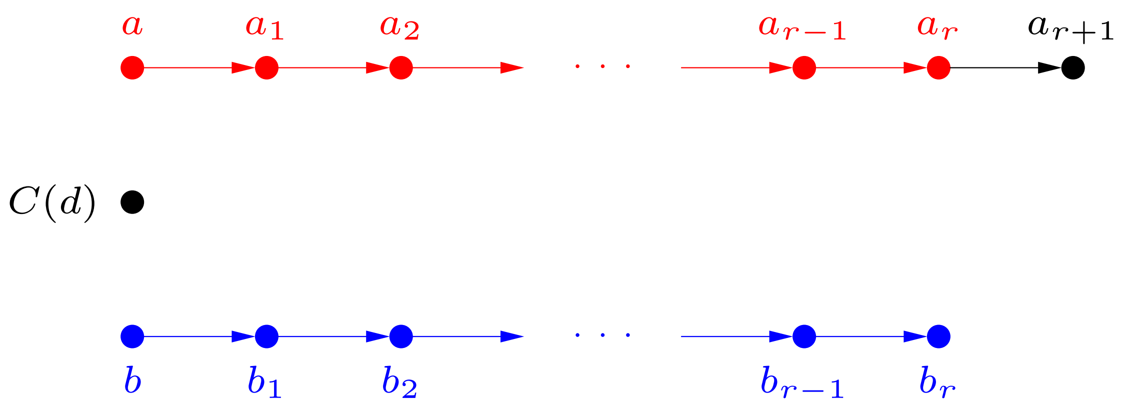


Canonical:

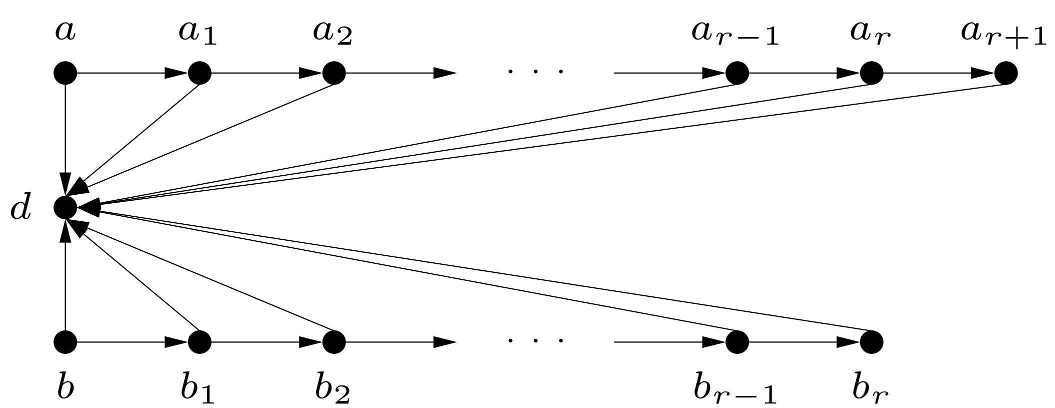


Locality of transformations under isomorphism

Source:

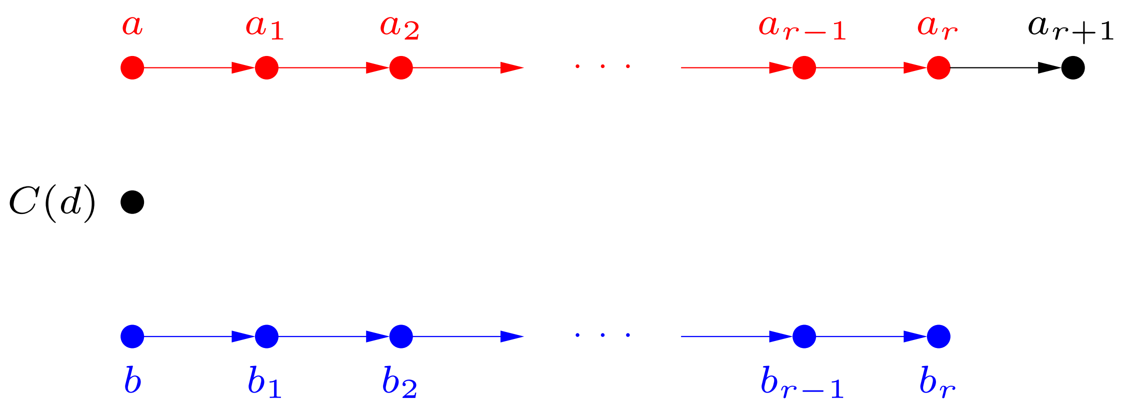


Canonical:

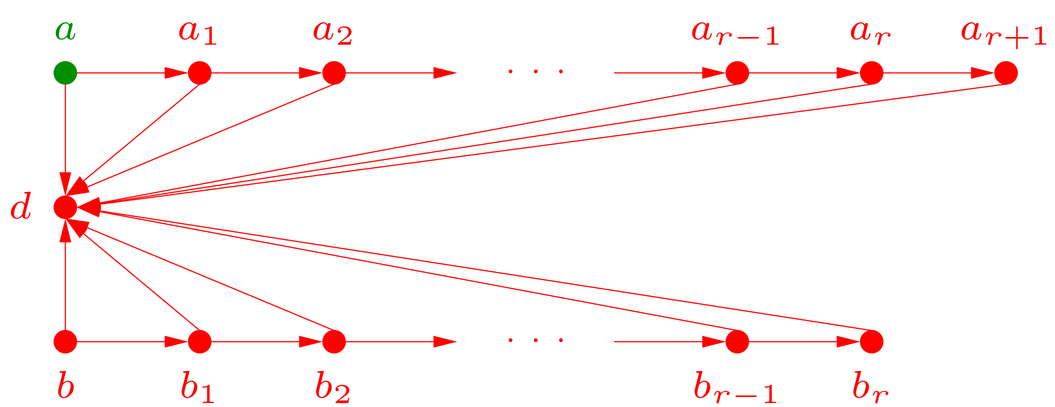


Locality of transformations under isomorphism

Source:

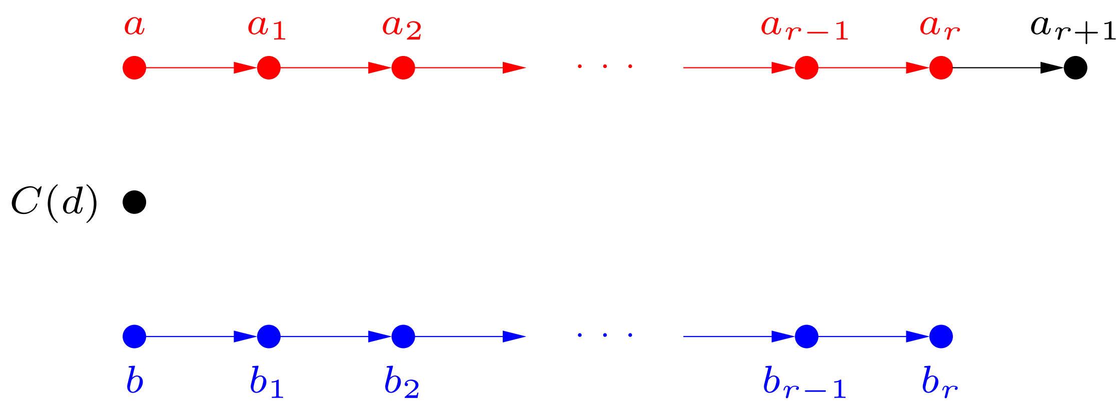


Canonical:

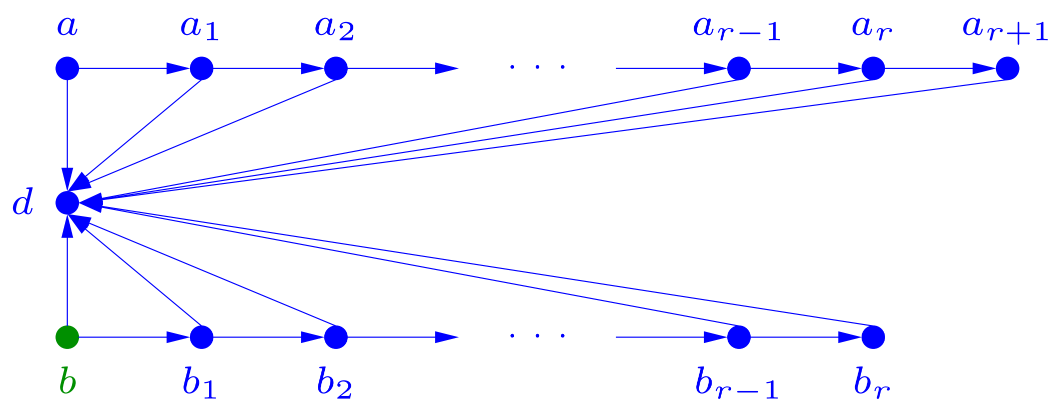


Locality of transformations under isomorphism

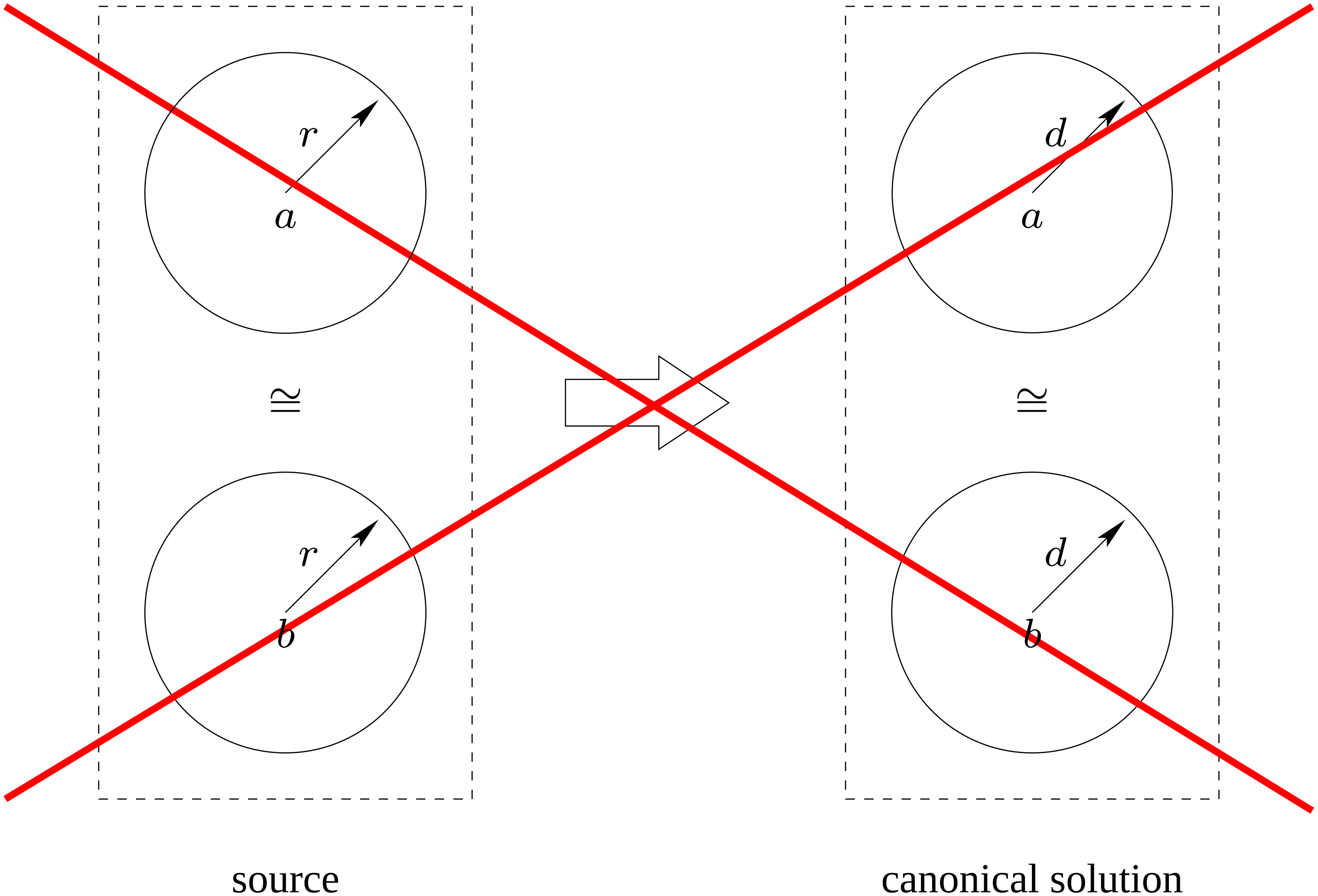
Source:



Canonical:



Locality of transformations under isomorphism



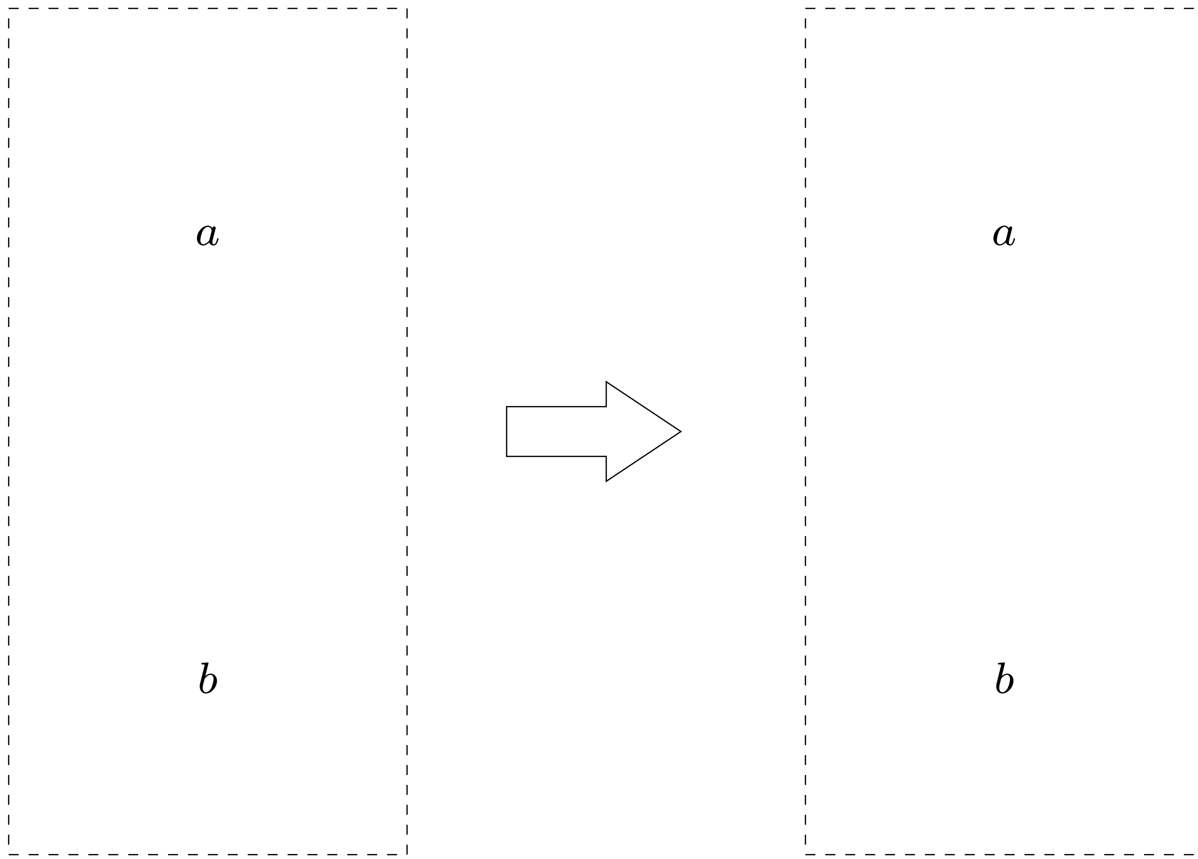
Locality of transformations: Notation

Quantifier rank: Depth of quantifier nesting, denoted $\mathsf{qr}(\phi)$.

Example: $\mathsf{qr}\Big(\exists x\,((\forall y\, P(x,y)) \wedge (\exists u \forall v\, U(x,u,v))\Big) = 3$.

Notion of equivalence: $I_1 \equiv_k I_2$ if I_1 and I_2 agree on all formulas of quantifier rank k.

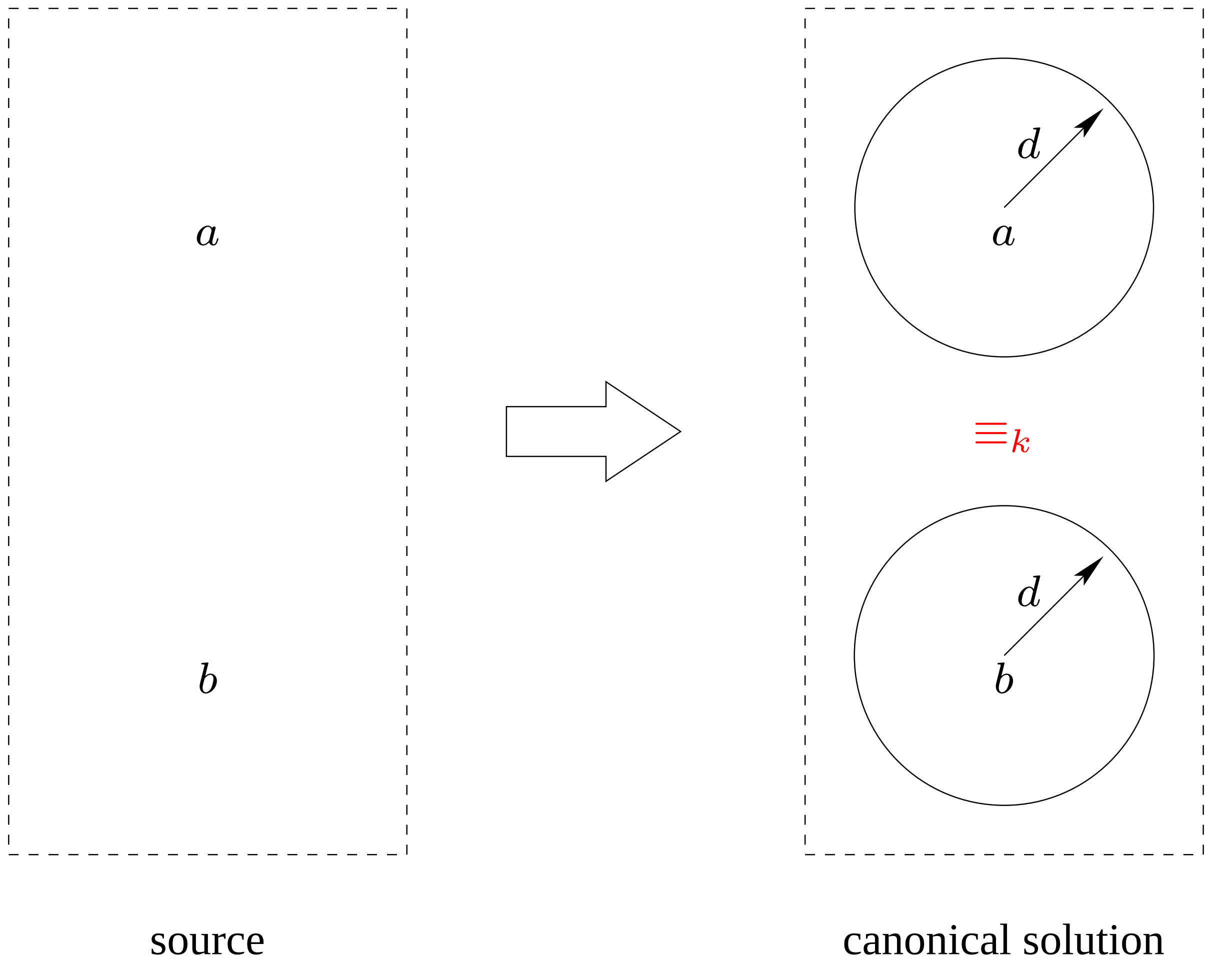
Locality of transformations under logical equivalence



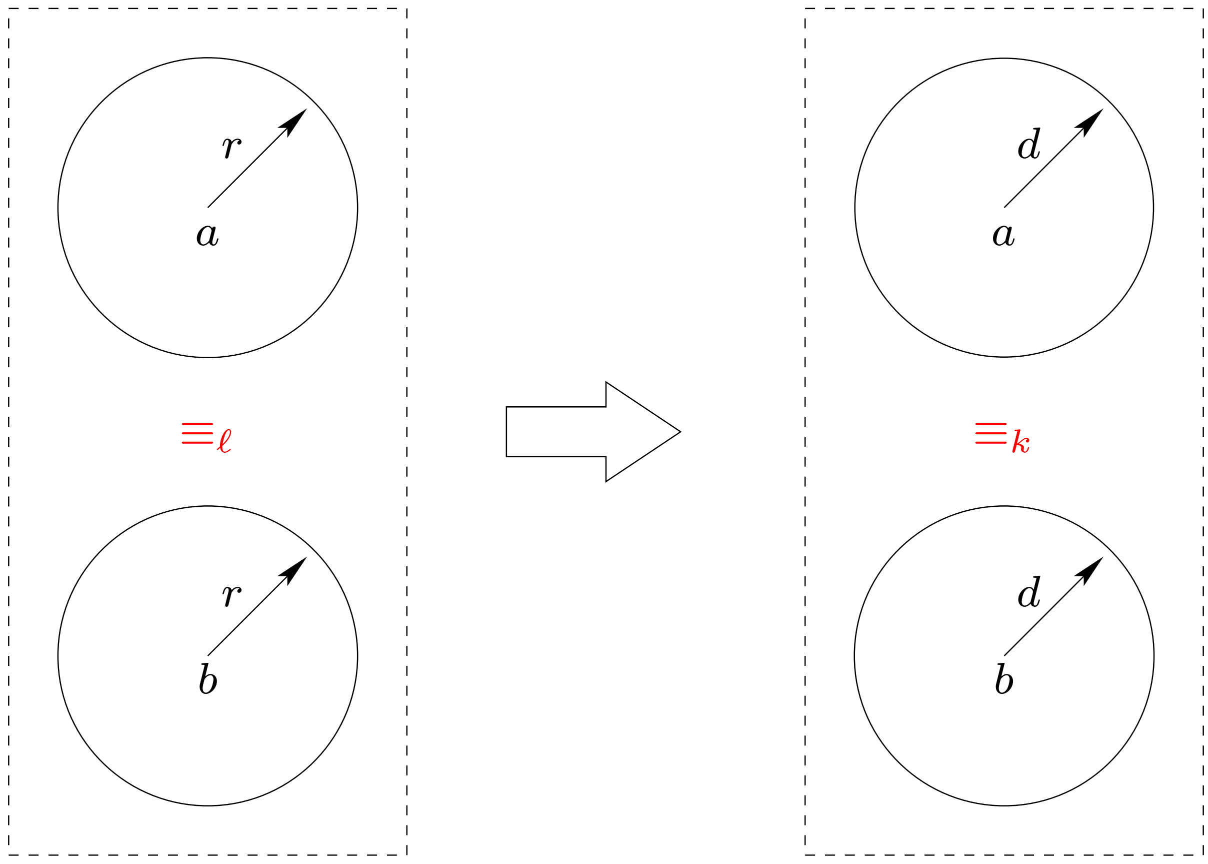
source

canonical solution

Locality of transformations under logical equivalence



Locality of transformations under logical equivalence



source

canonical solution

Locality of transformations under logical equivalence

Locality of a transformation under logical equivalence: For every $d, k \geq 0$ there exists $r, \ell \geq 0$ such that, for every instance I of $\mathbf{S}$ and tuples $\bar{a}, \bar{b}$ in I,

$$N_r^I(\bar{a}) \equiv_\ell N_r^I(\bar{b}) \implies N_d^{\mathcal{F}_{\text{can}}(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\text{can}}(I)}(\bar{b}).$$

Theorem: $\mathcal{F}_{\text{can}}$ satisfies this notion for every data exchange setting.

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

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

What about other transformations?

Core of canonical solution J: Substructure $J^\star$ of J such that there is a homomorphism from J to $J^\star$ and there is no homomorphism from J to a proper substructure of $J^\star$.

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

Core is the smallest solution that is *homomorphically equivalent* to the canonical solution.

- It can be computed in polynomial time (data complexity) [FKP03].

Example: Core

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

Source instance:
$I = \{Employee(peter, 2213477), Employee(peter, 2213479)\}$.

Solutions:

- $\{Dept(peter, 1)\}$.
- $\ldots$
- Canonical solution: $\{Dept(peter, X), Dept(peter, Y)\}$.
- Core: $\{Dept(peter, Z)\}$.

Query rewriting over the core

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

Theorem [FKMP03]: For every data exchange setting and union conjunctive queries Q, there exists Q' such that for every source instance I, *certain*$(Q, I) = Q'(\mathcal{F}_{\mathrm{core}}(I))$.

- Certain answers can be computed more efficiently by using the core.

Rewritability over the core: Can we use locality?

Canonical solution versus core: First attempt

Proposition: There exists a data exchange setting $\mathcal{A} = (\mathbf{S}, \mathbf{T}, \Sigma_{st})$ such that for every data exchange setting $\mathcal{B} = (\mathbf{S}, \mathbf{T}, \Gamma_{st})$, there exists instance I of $\mathbf{S}$ such that:

$$\mathcal{F}^{\mathcal{A}}_{\text{core}}(I) \quad \ncong \quad \mathcal{F}^{\mathcal{B}}_{\text{can}}(I).$$

We need a different approach ...

Expressiveness: Canonical solution versus core

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

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

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

Expressiveness: Canonical solution versus core

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

Expressiveness point of view: Canonical solution is better than the core.

- Canonical solution contains more information than the core.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

What about other semantics?

Usual certain answers semantics sometimes exhibit counterintuitive behavior.

Good solutions: Universal solutions.

- Homomorphically equivalent to the canonical solution.

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

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

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, $\underline{\textit{u-certain}}(Q, I) = Q'(\mathcal{F}_{\text{core}}(I))$.

Query rewriting under the universal solutions semantics

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 solution (core) under the universal solutions semantics, then Q is locally source-dependent under the universal solutions semantics.

Outline

- Motivation: Data exchange.
- First transformation: Canonical solution.
- Locality of queries.
- Locality in data exchange.
- Locality of transformations.
- Second transformation: The core.
- Extension: Other semantics.
- Conclusions.

Conclusions

- Locality notions have been very useful for studying the expressive power of query languages.

- Common data exchange transformations map similar neighborhoods into similar neighborhoods.

- This property can be used to formulate locality notions for data exchange transformations and query languages.

- Locality notions can be used for studying the expressive power of transformations and query languages in data exchange.