

Locally Consistent Transformations and Query Answering in Data Exchange

Marcelo Arenas
University of Toronto

Pablo Barceló
University of Toronto

Ronald Fagin
IBM Almaden

Leonid Libkin
University of Toronto

To appear: 2004 ACM Symposium on Principles of Database Systems (PODS 2004)

Abstract

Data exchange is the problem of taking data structured under a source schema and creating an instance of a target schema. Given a source instance, there may be many solutions – target instances that satisfy the constraints of the data exchange problem. Previous work has identified two classes of desirable solutions: canonical universal solutions, and their cores. Query answering in data exchange amounts to rewriting a query over the target schema to another query that, over a materialized target instance, gives the result that is semantically consistent with the source. A basic question is then whether there exists a transformation sending a source instance into a solution over which target queries can be answered.

We show that the answer is negative for many data exchange transformations that have structural properties similar to the canonical universal solution and the core. Namely, we prove that many such transformations preserve the *local* structure of the data. Using this notion, we further show that every target query rewritable over such a transformation cannot distinguish tuples whose neighborhoods in the source are similar. This gives us a first tool that helps check whether a query is rewritable. We also show that these results are robust: they hold for an extension of relational calculus with grouping and aggregates, and for two different semantics of query answering.

1 Introduction

Data exchange is the problem of materializing an instance that adheres to a target schema, given an instance of a source schema and a specification of the relationship between the source and the target. This is a very old problem [26] that arises in many tasks where data must be transferred between independent applications that do not have the same data format. The need for data exchange has steadily increased over the years. With the proliferation of web data in various for-

mat and with the emergence of e-business applications that need to communicate data yet remain autonomous, data exchange is even more important.

A data exchange setting is a triple $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, where \mathbf{S} is the source schema, \mathbf{T} is the target schema, and Σ_{st} is a set of source-to-target dependencies that express the relationship between \mathbf{S} and \mathbf{T} (some papers also add a set Σ_t of dependencies that express constraints on \mathbf{T} , but here, we will mostly consider data exchange settings with no target constraints). Such a setting gives rise to the following *data exchange problem*: given an instance I over the source schema \mathbf{S} , find an instance J over the target schema \mathbf{T} such that I together with J satisfy the source-to-target dependencies Σ_{st} (when target dependencies are used, J must also satisfy them). Such an instance J is called a *solution* for I in the data exchange setting. In general, there may be many different solutions for a given source instance I . For a data exchange system, the two key issues are:

1. Which solution should be materialized?
2. How should queries be answered over the target?

Papers [8, 9] started a systematic investigation of these issues for data exchange settings in which \mathbf{S} and \mathbf{T} are relational schemas. They isolated a class of solutions, called *universal solutions*, possessing good properties that justify selecting them as the best solutions in data exchange. Specifically, universal solutions have homomorphisms into every possible solution; in particular, they have homomorphisms into each other, and thus are homomorphically equivalent. Universal solutions are the most general among all solutions and, in a precise sense, they represent the entire space of solutions. It was shown in [8] that under fairly general conditions, universal solutions exist, and a *canonical* universal solution can be found in polynomial time, based on the classical chase procedure [4, 25].

Since universal solutions need not be unique, this raises the question of which universal solution to materialize. The answer proposed in [9] is based on using *minimality* as a key criterion for what constitutes the “best” universal solution. Although universal solutions come in different sizes, all of them share a unique (up to isomorphism) common “part”,

which is nothing else but the *core* of each of them, when they are viewed as relational structures [9]. By definition, the core of a structure is the smallest substructure that is also a homomorphic image of the structure. The concept of the core originated in graph theory, where a number of its properties have been established [15]. It was shown in [9] that if the source-to-target dependencies Σ_{st} are tuple-generating dependencies (tgds), then the core of the universal solutions for I is itself a solution for I (they also allow the possibility of having certain sets Σ_t of target dependencies). Hence, the core of the universal solutions for I is the *smallest* universal solution for I , and thus an ideal candidate for the “best” solution, at least in terms of the space required to materialize it. Furthermore, in a number of cases of interest, they show that there is a polynomial-time algorithm for generating the core.

We now turn to discussing query answering, and the related issue of query rewriting [23, 13]. Given a source instance and a data exchange setting, what is the meaning of the “answer” to a query Q over the target schema? Since there may be multiple solutions to the data exchange problem, the standard approach is to define the answer to be the set of *certain answers* [19, 1], that is, those tuples that appear in $Q(J)$ for every solution J . The goal of query answering in data exchange is to find these certain answers based on just *one* materialized target instance.

If Q is a union of conjunctive queries, and J is an arbitrary universal solution, then [8] showed that the certain answers are given exactly by the set of all tuples in $Q(J)$ that are formed entirely of elements from the source. Such nice behavior fails when we go beyond unions of conjunctive queries: it was shown in [8] that there is a Boolean conjunctive query Q with inequalities such that $Q(J)$ does not give the certain answers, no matter which universal solution J is selected, but for some other first-order query Q' (a *rewriting* of Q), the certain answers for Q are given by $Q'(J)$, where J is the canonical universal solution. Unfortunately, query rewritability is not a general phenomenon either, as [8] constructed a Boolean conjunctive query Q with inequalities for which there is no such rewriting Q' .

But the following basic question remains unanswered: is there a transformation \mathcal{F} that maps each source instance I into a solution $\mathcal{F}(I)$ and a rewriting Q' such that the certain answers are given by $Q'(\mathcal{F}(I))$? Of course we want to forbid cheating solutions (like encoding the answer to a Boolean query with a self-loop). But what is a natural condition then to impose on a transformation? Such a condition must ensure a certain degree of “uniformity” of \mathcal{F} (that is, it should not be tailored to deal with a specific query), and be satisfied by the transformations commonly used in data exchange such as $\mathcal{F}_{\text{univ}}$ that maps the source instance I onto the canonical universal solution, or $\mathcal{F}_{\text{core}}$ that maps I onto the core of the universal solutions.

The condition we impose on a transformation \mathcal{F} is that it must be *locally consistent*, that is, points with similar neighborhoods in the source have similar neighborhoods in the target. We make this notion of “locally consistent” precise (in fact, there are two closely related but incomparable properties based on the exact definition of “similarity”), and prove that, in an appropriate data exchange setting, $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ possess both properties.

One of our main results is that the failure of the canonical universal solution to support rewriting is not because there is a “better” choice of solution. Specifically, we show that if the transformation that produces the solution is locally consistent, then there are first-order queries that are not rewritable. This implies that neither the canonical universal solution, nor the core, nor any other “uniformly” generated solution supports rewriting for arbitrary first-order queries. We prove this by showing that queries rewritable over locally consistent transformations cannot distinguish points that have isomorphic neighborhoods in the source instance. Unlike ad hoc techniques employed in [9, 5], this criterion gives us easy ways of showing that a query is not rewritable.

The notion of local consistency introduced in this paper is a new one; although it is inspired by standard notions of locality from logic [11, 14, 10], it is different from them since this is the first notion of locality that applies to transformations that invent new values.

We also prove two extensions of the main results. The first one shows that all the results continue to hold if instead of first-order queries, we use an extension with grouping and aggregate functions, that is, essentially the `select-from-where-groupby-having` fragment of SQL. Secondly, we look at an alternative semantics (proposed in [9]) for query answering in data exchange that, instead of taking certain answers (those tuples that appear in $Q(J)$ for every solution J), takes tuples that appear in $Q(J)$ for every *universal* solution J . We prove that the main results of the paper remain true under this semantics.

Organization Basic notions related to data exchange, universal solutions, cores, and neighborhoods are presented in Section 2. In Section 3 we study structural properties of data exchange transformations. We present a rule-based language that allows us to code many such transformations, and prove local consistency for programs in that language. We derive results for $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ as corollaries. We also briefly consider target dependencies.

In Section 4, we study query rewritability. We show that a query rewritable over any locally consistent transformation cannot distinguish constants whose neighborhoods in the source are isomorphic. We show that this property gives us easy non-rewritability results. We also establish a connection between rewritability over the core, and rewritabil-

ity over the canonical universal solution.

In Section 5, we present extensions of these results to languages with grouping and aggregation, and to the semantics based on universal solutions. Summary and concluding remarks are given in Section 6. Because of space limitations, all proofs are in the appendix.

2 Preliminaries

A *schema* \mathbf{R} is a finite sequence $\langle R_1, \dots, R_k \rangle$ of relation symbols, with each R_i having a fixed arity n_i . An *instance* I of \mathbf{R} assigns to each relation symbol R_i of \mathbf{R} a finite n_i -ary relation $I(R_i)$. The *domain* $\text{dom}(I)$ of instance I is the set of all elements that occur in any of the relations $I(R_i)$ ¹. An instance J of \mathbf{R} is a *subinstance* of I if $\text{dom}(J) \subseteq \text{dom}(I)$ and $J(R_i) \subseteq I(R_i)$, for every i . If one of the inclusions is proper, we refer to J as a *proper subinstance* of I . If \mathbf{R} is a schema, then a *dependency over* \mathbf{R} is a sentence in some logical formalism over \mathbf{R} , typically FO (first-order logic).

2.1 Data exchange setting

Let $\mathbf{S} = \langle S_1, \dots, S_n \rangle$ and $\mathbf{T} = \langle T_1, \dots, T_m \rangle$ be two schemas with no relation symbols in common. We refer to \mathbf{S} as the *source* schema and to the S_i 's as the source relation symbols. We refer to \mathbf{T} as the *target* schema and to the T_j 's as the target relation symbols. We denote by $\langle \mathbf{S}, \mathbf{T} \rangle$ the schema $\langle S_1, \dots, S_n, T_1, \dots, T_m \rangle$. Instances over \mathbf{S} will be called source instances, while instances over \mathbf{T} will be called target instances. If I is a source instance and J is a target instance, then (I, J) denotes an instance K over $\langle \mathbf{S}, \mathbf{T} \rangle$ such that $K(S_i) = I(S_i)$ and $K(T_j) = J(T_j)$, for $i \in [1, n]$ and $j \in [1, m]$.

A *source-to-target dependency* (std) is a sentence of the form

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

where $\varphi_{\mathbf{S}}(\bar{x})$ is a formula over \mathbf{S} in some logical formalism (typically FO) and $\psi_{\mathbf{T}}(\bar{x}, \bar{y})$ is a conjunction of FO atomic formulae over \mathbf{T} .

Definition 2.1 (Data Exchange Setting) A data exchange setting is a triple $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, where \mathbf{S} is a source schema, \mathbf{T} is a target schema, and Σ_{st} is a set of source-to-target dependencies. The data exchange problem associated with $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ is the following: given a source instance I , find a target instance J such that (I, J) satisfies Σ_{st} . Such a J

¹An instance is a special case of an \mathbf{R} -structure \mathbf{A} defined as $(A, R_1^{\mathbf{A}}, \dots, R_k^{\mathbf{A}})$, where A is a set (the *universe*), and $R_i^{\mathbf{A}} \subseteq A^{n_i}$ for each i . Thus, in the case of arbitrary structures, the universe may contain elements that are not present in any of the relations.

is called a solution for I , or simply a solution if I is clear from the context.

We denote by Const an infinite set of all values that may occur in source instances, and, following the data exchange terminology [8, 9], we call those values *constants*. In addition, we also assume an infinite set Var of elements, disjoint from Const . Elements of Var are called *nulls* [8, 9], and they are used to help populate target instances. That is, the domain of a target instance comes from $\text{Const} \cup \text{Var}$.

If I is an instance with values in $\text{Const} \cup \text{Var}$, then $\text{Const}(I)$ denotes the set of all constants occurring in relations in I , and $\text{Var}(I)$ denotes the set of nulls occurring in relations in I . From now on, we assume that there is a way to distinguish constants from nulls. For example, we may assume that the target schema \mathbf{T} contains an auxiliary predicate C whose interpretation is $\text{dom}(I) \cap \text{Const}$.

Papers [8, 9] identified two important subclasses of data exchange, inspired by the *local-as-view (LAV)* and *global-as-view (GAV)* classes of data integration problems [22]:

- *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 some relation symbol in the source schema \mathbf{S} , and, as before, $\psi_{\mathbf{T}}(\bar{x}, \bar{y})$ is a conjunction of atomic formulae over \mathbf{T} .
- *GAV setting*: each dependency in Σ_{st} is of the form $\varphi_{\mathbf{S}}(\bar{x}) \rightarrow T(\bar{x})$, where T is some relation symbol in the target schema \mathbf{T} . If $\varphi_{\mathbf{S}}(\bar{x})$ is a conjunctive query, we speak of the GAV(CQ) setting.²

Example 2.2 Consider a LAV data exchange setting in which $\mathbf{S} = \langle M(\cdot, \cdot), N(\cdot, \cdot) \rangle$, $\mathbf{T} = \langle P(\cdot, \cdot, \cdot), Q(\cdot, \cdot) \rangle$ and Σ_{st} contains the following stds:

$$\begin{aligned} M(x, y) &\rightarrow \exists w \exists z (P(x, y, z) \wedge Q(w, z)), \\ N(x, y) &\rightarrow \exists z P(x, y, z). \end{aligned}$$

Suppose we have a source instance $I = \{M(a, b), N(a, b)\}$.³ Since the stds in Σ_{st} do not completely specify the target, there are multiple solutions that are consistent with the specification. One solution is:

$$J = \{P(a, b, n_1), P(a, b, n_2), Q(n_3, n_1)\},$$

where n_1, n_2, n_3 are values in Var . Another solution, but with no nulls, is $J' = \{P(a, b, a), Q(b, a)\}$. \square

²In [8, 9], the formula $\varphi_{\mathbf{S}}(\bar{x})$ was restricted to being the conjunction of atomic formulae over \mathbf{S} , which is equivalent to the GAV(CQ) setting.

³It is often convenient to define instances by simply listing the tuples attached to the corresponding relation symbols.

2.2 Universal solutions and cores

Let J and J' be two instances over the target schema \mathbf{T} with values in $\text{Const} \cup \text{Var}$. A *homomorphism* $h : J \rightarrow J'$ is a mapping from $\text{Const}(J) \cup \text{Var}(J)$ to $\text{Const}(J') \cup \text{Var}(J')$ such that $h(c) = c$ for all $c \in \text{Const}(J)$, and $\bar{t} \in J(R)$ implies $h(\bar{t}) \in J'(R)$ for all $R \in \mathbf{T}$. Furthermore, we say that J and J' are *homomorphically equivalent* if there are homomorphisms $h : J \rightarrow J'$ and $h' : J' \rightarrow J$.

Definition 2.3 (Universal solution) *If I is a source instance in a data exchange setting, then a universal solution for I is a solution J such that for every solution J' for I , there exists a homomorphism $h : J \rightarrow J'$.*

Example 2.4 The solution J' in Example 2.2 is not universal, since there is no homomorphism from J' to J . But it can be shown that J is a universal solution. \square

It was shown in [8] that universal solutions possess good properties that justify selecting them (as opposed to arbitrary solutions) for the semantics of the data exchange problem. A universal solution is more general than an arbitrary solution because it can be homomorphically mapped into that solution. Moreover, all universal solutions are homomorphically equivalent. Furthermore, results of [8] imply that for the data exchange setting considered in this paper, universal solutions always exist.

To deal with the problem of computing universal solutions, [8] proposes to compute a special kind of universal solution, called a *canonical universal solution*. The algorithm presented in [8] is based on applying the chase, but we shall define canonical universal solutions directly, in Section 3.

The reason one wants to compute a specific solution for the data exchange problem is to be able to evaluate queries over the target schema. It was noted in [8] that universal solutions need not be isomorphic, and thus any decision to choose one is somewhat arbitrary. To deal with this problem, [9] proposed to use the *core* of the universal solutions.

Definition 2.5 (Core) *A subinstance J of an instance I is called a core of I if there is a homomorphism from I to J , but there is no homomorphism from I to a proper subinstance of J .*

It is known [15] that every instance has a unique core (up to isomorphism). It is shown in [9] that every universal solution has the same core (up to isomorphism), and that this core is itself a universal instance. Further, it is shown in [9] that under the assumptions in this paper, the core can be computed in polynomial time.

Example 2.6 In Example 2.2, $J^* = \{P(a, b, n_1), Q(n_3, n_1)\}$ is the core of the universal solutions. \square

2.3 Neighborhoods and locality

The *Gaifman graph* $\mathcal{G}(I)$ of an instance I of \mathbf{R} is the graph whose nodes are the elements of $\text{dom}(I)$, and such that there exists an edge between a and b in $\mathcal{G}(I)$ iff a and b belong to the same tuple of a relation $I(R)$, for some $R \in \mathbf{R}$. For example, if I is an undirected graph, then $\mathcal{G}(I)$ is I itself.

The distance between two elements a and b in I , denoted by $d_I(a, b)$ (or $d(a, b)$, if I is understood), is the distance between them in $\mathcal{G}(I)$. We define $d(\bar{a}, b)$ as the minimum value of $d(a, b)$ where a is an element of \bar{a} .

Given a tuple $\bar{a} = (a_1, \dots, a_m) \in \text{dom}(I)^m$, we define the instance $N_d^I(\bar{a})$, called the *d -neighborhood of \bar{a} in I* , as the restriction of I to the elements at distance at most d from \bar{a} , with the members of \bar{a} treated as distinguished elements. That is, if two neighborhoods $N_d^I(\bar{a})$ and $N_d^I(\bar{b})$ are isomorphic (written $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$), then there is an isomorphism $f : N_d^I(\bar{a}) \rightarrow N_d^I(\bar{b})$ such that $f(a_i) = b_i$, for $1 \leq i \leq m$.

A formula $\varphi(\bar{x})$ in some logical formalism is *local* if there exists a number d such that $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$ implies that $I \models \varphi(\bar{a})$ iff $I \models \varphi(\bar{b})$, for every instance I . It is known [11] that all FO formulae are local. This was generalized to logics that extend FO with counting [24] and aggregate functions [17].

3 Structural Properties of Data Exchange Transformations

In this section we show that data exchange transformations preserve the local character of the data. As a first step towards proving those results, we formulate a rule-based language for specifying transformations such as $\mathcal{F}_{\text{univ}}$, that maps the source instance I onto the canonical universal solution. This language is similar in spirit to languages with oid invention [3, 18, 27] but its rules are nonrecursive. Based on the types of logical formulae used in rules, we establish different results on locality of transformations, and then derive, as corollaries, exact characterizations of locality for various data exchange settings.

3.1 Data exchange programs

A *data exchange program* is a quadruple $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$, where \mathbf{S} (“source”), \mathbf{A} (“auxiliary”) and \mathbf{T} (“target”) are pairwise disjoint relational schemas and \mathcal{R} is a sequence

$\langle r_1, \dots, r_n \rangle$ of rules such that each rule is of the form

$$R_1(\bar{x}_1, \bar{y}_1), \dots, R_k(\bar{x}_k, \bar{y}_k) \quad :- \quad \varphi(\bar{x}), \quad (1)$$

where each R_i is either in \mathbf{A} or in \mathbf{T} , where $\varphi(\bar{x})$ is an FO formula over $\langle \mathbf{S}, \mathbf{A} \rangle$, where variables in the \bar{x}_i 's are among those in \bar{x} , and variables in the \bar{y}_i 's are not in \bar{x} . For example, $R_1(x_1, y_1, y_2), R_2(x_1, y_1) \quad :- \quad \varphi(x_1, x_2)$ is a rule. We refer to $R_1(\bar{x}_1, \bar{y}_1), \dots, R_k(\bar{x}_k, \bar{y}_k)$ as the *head* of the rule, and to $\varphi(\bar{x})$ as the *body* of the rule.

Furthermore, we require that the program be *stratified*. That is, if \mathbf{A}_i is the set of relation symbols from \mathbf{A} used in rules r_1, \dots, r_i , then the formula φ in the body of rule r_{i+1} is over the schema $\langle \mathbf{S}, \mathbf{A}_i \rangle$.

Given a data exchange program $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$, we define the transformation $\mathcal{F}_\Pi : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$ that associates a target schema instance with each source schema instance. For that, we show inductively how to define a mapping $\mathcal{F}_\Pi^i : \text{inst}(\langle \mathbf{S}, \mathbf{A}, \mathbf{T} \rangle) \rightarrow \text{inst}(\langle \mathbf{S}, \mathbf{A}, \mathbf{T} \rangle)$ given by the first i rules of the program. Suppose we are given an instance I of \mathbf{S} , and $J = \mathcal{F}_\Pi^{i-1}(I)$, where $1 \leq i \leq n$ (if $i = 1$, then $J(S) = I(S)$ for every $S \in \mathbf{S}$, and $J(P) = \emptyset$ for every $P \in \langle \mathbf{A}, \mathbf{T} \rangle$). Let the i th rule be given by (1), let \bar{u} be the tuple of variables in \bar{x} that are used in the head of the rule, and let \bar{v} be the tuple of variables in the head of the rule that are not in \bar{x} .

For each tuple \bar{a} of length $|\bar{u}|$ over $\text{dom}(J)$, find all the tuples $\bar{b}_1, \dots, \bar{b}_m$ such that $J \models \varphi(\bar{a}, \bar{b}_j)$, for $1 \leq j \leq m$. Then choose m tuples of length $|\bar{v}|$ of fresh distinct null values $\bar{n}_1, \dots, \bar{n}_m$ over Var . To construct relation R_l , for $l \leq k$, in $\mathcal{F}_\Pi^i(I)$, add tuples $(\pi_{\bar{x}_l}(\bar{a}), \pi_{\bar{y}_l}(\bar{n}_j))$, for $1 \leq j \leq m$, to the relation $J(R_l)$. Here $\pi_{\bar{x}_l}(\bar{a})$ refers to the components of \bar{a} that occur in the positions of \bar{x}_l .

For example, consider again the rule $R_1(x_1, y_1, y_2), R_2(x_1, y_1) \quad :- \quad \varphi(x_1, x_2)$, and assume that $J = \mathcal{F}_\Pi^{i-1}(I)$ has been computed. Consider all $a \in \text{dom}(J)$, and let b_1, \dots, b_m be such that $J \models \varphi(a, b_j)$, for $1 \leq j \leq m$. Then choose $2m$ fresh values n_j^1, n_j^2 for $1 \leq j \leq m$ in Var and add tuples (a, n_j^1, n_j^2) to $J(R_1)$, and (a, n_j^1) to $J(R_2)$, for $1 \leq j \leq m$.

Finally, $\mathcal{F}_\Pi(I)$ is defined to be the restriction of \mathcal{F}_Π^n to the predicates in \mathbf{T} .

Next, we connect data exchange problems with the data exchange setting defined earlier. Given a data exchange setting $\mathcal{DE} = (\mathbf{S}, \mathbf{T}, \Sigma_{st})$, define a data exchange program $\Pi_{\mathcal{DE}} = (\mathbf{S}, \emptyset, \mathbf{T}, \mathcal{R})$, where \mathcal{R} contains a rule $R_1(\bar{x}_1, \bar{y}_1), \dots, R_k(\bar{x}_k, \bar{y}_k) \quad :- \quad \varphi_{\mathbf{S}}(\bar{x})$ for each std $\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists \bar{y} (R_1(\bar{x}_1, \bar{y}_1) \wedge \dots \wedge R_k(\bar{x}_k, \bar{y}_k))$ in Σ_{st} .

Definition 3.1 (Canonical Universal Solution) *The canonical universal solution of instance I in data exchange setting $\mathcal{DE} = (\mathbf{S}, \mathbf{T}, \Sigma_{st})$ is $\mathcal{F}_{\Pi_{\mathcal{DE}}}(I)$. If the data*

exchange setting \mathcal{DE} is understood, we shall denote this transformation $\mathcal{F}_{\Pi_{\mathcal{DE}}}$ by $\mathcal{F}_{\text{univ}}$.

This definition differs from the one given in [8], but it can easily be seen that for data exchange settings as described in this paper, the two definitions coincide.

We define the transformation $\mathcal{F}_{\text{core}}$ such that $\mathcal{F}_{\text{core}}(I)$ is the core of $\mathcal{F}_{\text{univ}}(I)$.

Notice that in the definition of $\Pi_{\mathcal{DE}}$, we did not use any auxiliary relations from \mathbf{A} . One may then ask if those are necessary. The result below shows that they are.

Proposition 3.2 *There is a data exchange program $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ such that \mathcal{F}_Π is different from $\mathcal{F}_{\text{univ}}$, and from $\mathcal{F}_{\text{core}}$, for every data exchange setting $\langle \mathbf{S}', \mathbf{T}', \Sigma_{st} \rangle$.*

3.2 Locally consistent transformations

In this section we introduce the notions of local consistency of transformations from $\text{inst}(\mathbf{S})$ to $\text{inst}(\mathbf{T})$. The first notion says that neighborhoods around elements common to the input and output instances are preserved. Informally, if $a, b \in \text{dom}(I)$ are present in the domain of the resulting instance J of \mathbf{T} , then the isomorphism of sufficiently large neighborhoods of a and b in I guarantees that their neighborhoods are isomorphic in J as well. Formally, we define this as follows.

Definition 3.3 (Local Consistency) *A mapping $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$ is locally consistent if for every $m, d \geq 0$ there exists $d' \geq 0$ such that, for every instance I of \mathbf{S} and m -tuples $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $N_{d'}^I(\bar{a}) \cong N_{d'}^I(\bar{b})$, then*

- 1) $\bar{a} \in \text{dom}(\mathcal{F}(I))^m \Leftrightarrow \bar{b} \in \text{dom}(\mathcal{F}(I))^m$, and
- 2) $N_d^{\mathcal{F}(I)}(\bar{a}) \cong N_d^{\mathcal{F}(I)}(\bar{b})$.

We next present a sufficient condition for a mapping \mathcal{F}_Π associated with a data exchange program Π to be locally consistent. This condition will guarantee local consistency for the LAV setting of data exchange.

We say that a formula $\varphi(\bar{x})$ is r -bounded if for every structure I such that $I \models \varphi(a_1, \dots, a_n)$, it is the case that $d_I(a_i, a_j) \leq r$ for every $i, j \leq n$. A data exchange program Π is r -bounded if every formula in the body of every rule is r -bounded.

Lemma 3.4 *The transformation \mathcal{F}_Π of an r -bounded data exchange program is locally consistent. \square*

Theorem 3.5 *In the LAV setting, both the canonical universal solution transformation $\mathcal{F}_{\text{univ}}$ and the core transformation $\mathcal{F}_{\text{core}}$ are locally consistent.* \square

The result for the canonical universal solution is an immediate consequence of Lemma 3.4, since in a LAV setting \mathcal{DE} , the bodies of rules in $\Pi_{\mathcal{DE}}$ are atomic predicates, which are 1-bounded. The result for the core requires a separate proof, which is given in the appendix. One can also show that local consistency for the core transformation depends crucially on the requirement of the data exchange setting that constants be preserved.

Theorem 3.5 does not extend to the GAV setting, even when restricted to conjunctive queries.

Proposition 3.6 (a) *There are GAV(CQ) settings such that Σ_{st} contains either one dependency of the form $\varphi_{\mathbf{S}}(x, y, z) \rightarrow T(x, y, z)$, or multiple dependencies of the form $\varphi_{\mathbf{S}}(x, y) \rightarrow T(x, y)$, and the corresponding transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ are not locally consistent.*

(b) *If, in the GAV(CQ) setting, Σ_{st} contains only one dependency of the form $\varphi_{\mathbf{S}}(x, y) \rightarrow T(x, y)$, then $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ are locally consistent.*

Since local consistency is a nontrivial property of FO-definable mappings, whether \mathcal{F}_{Π} is locally consistent is undecidable, even in the GAV setting (this easily follows from Trakhtenbrot’s theorem; cf. [6]).

3.3 Local consistency under logical equivalence

We have seen that mappings that arise in the LAV setting are locally consistent, but local consistency may fail even in some simple GAV settings. To overcome this, we introduce a notion of locality based on logical equivalence (in particular, FO-equivalence) rather than isomorphism of neighborhoods, and we prove that in general, the canonical universal solution transformation $\mathcal{F}_{\text{univ}}$ and the core transformation $\mathcal{F}_{\text{core}}$ are locally consistent under FO-equivalence.

The *quantifier rank* of an FO formula is the maximum depth of quantifier nesting in it. If I and J are instances of the same schema, we write $I \equiv_k J$ if I and J satisfy the same FO sentences of quantifier rank up to k . In the new notion of local consistency, we require that $\equiv_{k'}$ -equivalent neighborhoods be sent to \equiv_k -equivalent neighborhoods.

Definition 3.7 (Local Consistency under FO-equivalence) *A mapping $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$ is locally consistent under FO-equivalence if for every $m, d, k \geq 0$ there exist $d', k' \geq 0$ such that, for ev-*

ery instance I of \mathbf{S} and m -tuples $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$, then

- 1) $\bar{a} \in \text{dom}(\mathcal{F}(I))^m \Leftrightarrow \bar{b} \in \text{dom}(\mathcal{F}(I))^m$, and
- 2) $N_d^{\mathcal{F}(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}(I)}(\bar{b})$.

Lemma 3.8 *The transformation \mathcal{F}_{Π} of every data exchange program is locally consistent under FO-equivalence.* \square

Theorem 3.9 *For an arbitrary data exchange setting, both the canonical universal solution transformation $\mathcal{F}_{\text{univ}}$ and the core transformation $\mathcal{F}_{\text{core}}$ are locally consistent under FO-equivalence.* \square

The result for the canonical universal solution is an immediate consequence of Lemma 3.8. The result for the core is considerably harder and relies on the machinery developed for the proof of Theorem 4.8.

Note that the definitions of local consistency and local consistency under FO-equivalence are incomparable: the latter makes a weaker assumption and arrives at a weaker conclusion. Nevertheless, either definition works for our applications in query rewriting, because the statement we need there makes the stronger assumption (isomorphism of neighborhoods) but needs only the weaker conclusion (FO equivalence of neighborhoods).

3.4 Adding target dependencies

Papers [8, 9] considered an extension of the data exchange setting in which dependencies exist for the target schema as well. A solution is then required to satisfy those target dependencies.

Based on familiar classes of dependencies (cf. [4]), we define extensions of the data exchange setting with tgds as well as *equality-generating dependencies* (egds). The tgds over \mathbf{T} are of the form $\forall \bar{x}(\varphi_{\mathbf{T}}(\bar{x}) \rightarrow \exists \bar{y} \psi_{\mathbf{T}}(\bar{x}, \bar{y}))$, where $\varphi_{\mathbf{T}}(\bar{x})$ and $\psi_{\mathbf{T}}(\bar{x}, \bar{y})$ are conjunctions of FO atomic formulae. The egds over \mathbf{T} are of the form $\forall \bar{x}(\varphi_{\mathbf{T}}(\bar{x}) \rightarrow (x_1 = x_2))$, where $\varphi_{\mathbf{T}}(\bar{x})$ is a conjunction of atomic FO formulae, with free variables \bar{x} , and x_1, x_2 are in \bar{x} . If, furthermore, the data exchange setting is restricted to LAV or GAV, we shall speak of LAV+tdg, LAV+egd, etc. settings. The next proposition covers all four possible settings: LAV+tdg, GAV+tdg, LAV+egd, and GAV+egd.

Proposition 3.10 (a) *The transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ of LAV+tdg (or GAV+tdg) settings are not necessarily locally consistent (under FO-equivalence), even if there is only one target dependency.*

(b) *The transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ of GAV+egd settings are locally consistent under FO-equivalence.*

(c) The transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ of LAV+egd settings are not necessarily locally consistent (under FO-equivalence), even if all of the target dependencies are key dependencies.

4 Query Rewriting and Locality

In this section, we study query rewriting in data exchange. We use local consistency to show that rewritable queries have a certain kind of locality property. This property gives an easily applicable tool for proving nonexistence of rewritings over the canonical universal solution and the core.

4.1 The query rewriting problem

Suppose we have a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, and a query Q over the *target* schema \mathbf{T} . What does it mean to answer Q ? Since there are many possible solutions to the data exchange problem, the standard approach is to define the semantics of Q in terms of *certain answers*: that is, for an instance I of \mathbf{S} ,

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

Thus, a tuple \bar{a} is in $\text{certain}(Q, I)$ if it belongs to $Q(J)$ for all solutions J for I .

But how can one find this set $\text{certain}(Q, I)$, given that there are potentially infinitely many solutions? The approach proposed in [8, 9] is to look for some specific transformations $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$, and find conditions under which $\text{certain}(Q, I)$ equals $Q'(\mathcal{F}(I))$. Then Q is rewritable over \mathcal{F} by Q' . Formally,

Definition 4.1 (Query Rewriting) *Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, a mapping $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$ and an m -ary query Q over \mathbf{T} , we say that Q is rewritable over \mathcal{F} if there exists an m -ary FO query Q' over \mathbf{T} such that $\text{certain}(Q, I) = Q'(\mathcal{F}(I))$ for every instance I of \mathbf{S} .*

We shall refer to a query as being rewritable over the canonical universal solution if it is rewritable over $\mathcal{F}_{\text{univ}}$, and rewritable over the core if it is rewritable over $\mathcal{F}_{\text{core}}$. These notions are undecidable in general.

Proposition 4.2 *Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and an FO query Q over \mathbf{T} , it is undecidable whether Q is rewritable over the canonical universal solution, or over the core.* \square

In some cases, we can establish that a class of queries is or is not rewritable. For example, it is known that for every

FO sentence, its asymptotic probability is either 0 or 1 (this is the zero-one law for FO [7]).

Proposition 4.3 *In a data exchange setting, every Boolean query whose asymptotic probability is 0 is rewritable, by false, over both the canonical universal solution and over the core.* \square

However, such partial results do not give us any *techniques* for proving that queries are *not* rewritable. We shall now exhibit such techniques, based on the notions of locality from the previous section.

4.2 Local source-dependency and rewritability

We now show that queries rewritable over locally consistent transformations are guaranteed to satisfy a strong locality criterion on their own, and use these results to show that certain queries are not rewritable over the canonical universal solution or over the core.

Definition 4.4 (Locally source-dependent queries)

Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and a query Q over \mathbf{T} , we say that Q is locally source-dependent if there is $d \geq 0$ such that for every instance I of \mathbf{S} and for every $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$ then

$$(\bar{a} \in \text{certain}(Q, I) \Leftrightarrow \bar{b} \in \text{certain}(Q, I))$$

Theorem 4.5 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting, and Q a query over \mathbf{T} . Assume that Q is rewritable over a transformation \mathcal{F} , where \mathcal{F} is either locally consistent, or locally consistent under FO-equivalence. Then Q is locally source-dependent.* \square

Corollary 4.6 *In a data exchange setting, a target query rewritable over the canonical universal solution or over the core is locally source-dependent.* \square

We now show how this result can be used as a simple tool for proving non-rewritability results, even in very simple data exchange settings. We call a data exchange setting *copying* if \mathbf{S} and \mathbf{T} are two copies of the same schema (that is, $\mathbf{S} = \langle R_1, \dots, R_l \rangle$, $\mathbf{T} = \langle R'_1, \dots, R'_l \rangle$, and R_i and R'_i have the same arity), and $\Sigma_{st} = \{R_i(\bar{x}) \rightarrow R'_i(\bar{x}) \mid i = 1, \dots, l\}$. Note that a copying setting is both LAV and GAV.

Theorem 4.7 *There is a copying data exchange setting and an FO-query that is not rewritable over the canonical universal solution, nor over the core.*

Proof. Let $\mathbf{S} = \langle G(\cdot, \cdot), R(\cdot) \rangle$, $\mathbf{T} = \langle G'(\cdot, \cdot), R'(\cdot) \rangle$ and $\Sigma_{st} = \{G(x, y) \rightarrow G'(x, y), R(x) \rightarrow R'(x)\}$. Define a query $Q(x)$ over the target schema as:

$$R'(x) \vee \exists y \exists z (R'(y) \wedge G'(y, z) \wedge \neg R'(z)).$$

Assume that Q is FO-rewritable over $\mathcal{F}_{\text{univ}}$ or $\mathcal{F}_{\text{core}}$. Then it is locally source-dependent: there exists $d \geq 0$ such that for every source instance I and every $a, b \in \text{dom}(I)$, we have $a \in \underline{\text{certain}}(Q, I)$ iff $b \in \underline{\text{certain}}(Q, I)$ whenever $N_d^I(a) \cong N_d^I(b)$.

Define a source instance I as shown in Figure 1: $I(G)$ is the disjoint union of two cycles of length $2d + 2$, and $I(R) = \{c\}$. Then $N_d^I(a) \cong N_d^I(b)$, which implies that $a \in \underline{\text{certain}}(Q, I)$ iff $b \in \underline{\text{certain}}(Q, I)$. But it is easy to see that $a \in \underline{\text{certain}}(Q, I)$ and $b \notin \underline{\text{certain}}(Q, I)$. This contradiction shows that Q is not rewritable. \square

4.3 Rewritability over the core

We now establish the connection between rewritability over the core and rewritability over the canonical universal solution: we show that the former implies the latter.

Theorem 4.8 *Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, every query Q over the target schema that is rewritable over the core is also rewritable over the canonical universal solution. Moreover, there is a polynomial-time algorithm that, given a rewriting of Q over the core, finds a rewriting of Q over the canonical universal solution.* \square

The local consistency of $\mathcal{F}_{\text{core}}$ under FO equivalence, stated in Theorem 3.9, actually follows from several lemmas developed in the proof of this theorem.

The next proposition says that the converse of Theorem 4.8 does not hold.

Proposition 4.9 *There exists an FO query that is rewritable over the canonical universal solution, but not rewritable over the core.* \square

5 Extensions

Most results of the previous two sections can be extended in two ways. First, as the underlying language for both data exchange programs and query rewritability one can use an extension of FO with grouping and aggregation, corresponding to basic features of SQL `select-from-where-groupby-having` statements. Second, using an alternative semantics of queries over the target schema [9] we show that many results extend for that semantics as well.

5.1 Extended data exchange setting

So far, both data exchange settings and data exchange programs were based on first-order formulae: that is, all stds were of the form $\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists \bar{y} \psi_{\mathbf{T}}(\bar{x}, \bar{y})$, where $\varphi_{\mathbf{S}}(\bar{x})$ is an FO formula, and all formulae in the bodies of rules were FO as well.

We now show how to extend our main results to the setting where these formulae correspond not to relational calculus but to its extension with grouping and aggregates. Such languages are typically defined as an extension of relational algebra (see [20, 21]), but here instead we adopt the logic approach of [17].

Based on the approach of [12, 17], we define an *aggregate operator* to be a sequence $\mathcal{G} = \langle g_0, g_1, g_2, \dots, g_\omega \rangle$ of functions, where each g_n , for $0 < n < \omega$, takes an n -element bag of rational numbers, and returns a number in \mathbb{Q} . The values g_0 and g_ω are constants associated with the output of \mathcal{G} on the empty bag and on infinite bags, respectively (the latter may occur in the definition of the semantics of terms in the logic).

The *aggregate logic* FO_{aggr} over schema \mathbf{R} is two-sorted: first-sort variables range over domains on instances of \mathbf{R} , and second-sort variables range over \mathbb{Q} . It extends FO by

- *numerical terms and predicates:* for every function $f : \mathbb{Q}^n \rightarrow \mathbb{Q}$ and every predicate $P \subseteq \mathbb{Q}^n$, if $t_1(\bar{x}), \dots, t_n(\bar{x})$ are terms of the second (numerical) sort, then so is $f(t_1(\bar{x}), \dots, t_n(\bar{x}))$; furthermore, $P(t_1(\bar{x}), \dots, t_n(\bar{x}))$ is an atomic formula. These have the standard semantics.
- *aggregate terms:* for every aggregate operator \mathcal{G} , a second-sort term $t(\bar{x}, \bar{y})$ and a formula $\varphi(\bar{x}, \bar{y})$, we have a new second-sort term

$$t'(\bar{x}) = \text{Aggr}_{\mathcal{G}} \bar{y} (t(\bar{x}, \bar{y}), \varphi(\bar{x}, \bar{y})).$$

The semantics $t'(\bar{a})$ is defined as follows. If there are infinitely many \bar{b} such that $\varphi(\bar{a}, \bar{b})$ holds, then the value of $t'(\bar{a})$ is g_ω . Otherwise, let $\bar{b}_1, \dots, \bar{b}_m$ enumerate all the \bar{b} such that $\varphi(\bar{a}, \bar{b})$ holds. Then $t'(\bar{a})$ is defined as g_m applied to the bag $\{\{t(\bar{a}, \bar{b}_1), \dots, t(\bar{a}, \bar{b}_m)\}\}$.

Example 5.1 Let R be a ternary relation whose tuples are (d, e, s) , where d is the department name, e is the employee name, and s is the salary. The query that computes the total salary for each department is given by the following FO_{aggr} formula $\varphi(d, v)$:

$$(\exists e \exists s R(d, e, s)) \wedge (v = \text{Aggr}_{\mathcal{G}_{\text{SUM}}} (e, s)(s, R(d, e, s))),$$

where \mathcal{G}_{SUM} is the sequence $\langle g_0, g_1, g_2, \dots, g_\omega \rangle$ such that $g_0 = g_\omega = 0$ and $g_n(\{\{a_1, \dots, a_n\}\}) = a_1 + \dots + a_n$ for positive integers n . \square

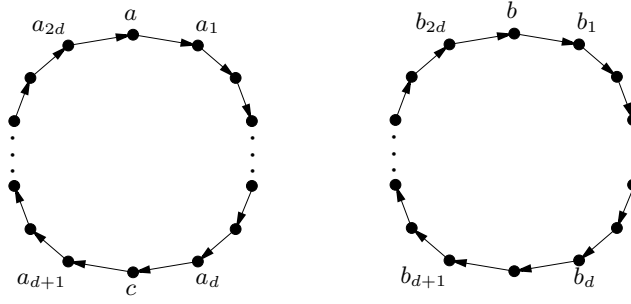


Figure 1: Instance I of Theorem 4.7.

We define an $\text{FO}_{\text{aggr}}\text{-data exchange setting}$ to be a data exchange setting in which every std is of the form $\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists \bar{y} \psi_{\mathbf{T}}(\bar{x}, \bar{y})$, where $\varphi_{\mathbf{S}}(\bar{x})$ is an FO_{aggr} formula with all free variables of the first sort. Likewise, we define an $\text{FO}_{\text{aggr}}\text{-data exchange program}$ as one in which all formulae in the bodies of rules are FO_{aggr} formulae with all free variables of the first sort. Just as in the case of FO, we define the canonical universal solution of an $\text{FO}_{\text{aggr}}\text{-data exchange setting}$ as the result of an FO_{aggr} data exchange program obtained by converting each std $\varphi_{\mathbf{S}}(\bar{x}) \rightarrow \exists \bar{y} (R_1(\bar{x}_1, \bar{y}_1) \wedge \dots \wedge R_k(\bar{x}_k, \bar{y}_k))$ into a rule $R_1(\bar{x}_1, \bar{y}_1), \dots, R_k(\bar{x}_k, \bar{y}_k) :- \varphi_{\mathbf{S}}(\bar{x})$.

Theorem 5.2 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be an $\text{FO}_{\text{aggr}}\text{-data exchange setting}$. Every query over \mathbf{T} that is $\text{FO}_{\text{aggr}}\text{-rewritable}$ over the canonical universal solution, or over the core, is locally source-dependent.* \square

The proof is based on a modified version of local consistency, in which we use equivalence with respect to a certain counting extension of FO [16, 24].

5.2 Universal solutions semantics

It was argued in [9] that since the universal solutions are the preferred solutions in data exchange, it may be more fundamental and meaningful to consider semantics based on them. Furthermore, it can be shown that the usual certain answers semantics sometimes exhibit rather counterintuitive behavior (an example is given in the appendix).

Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, an m -ary query Q over \mathbf{T} , and a source instance I , we define the *universal solutions semantics* of Q as

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

Clearly, $\underline{\text{certain}}(Q, I) \subseteq \underline{\text{u-certain}}(Q, I)$. We now show that the main results of this paper are preserved when one

considers this new semantics of answering queries over the target.

Given a mapping $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$, we say that Q is $\text{FO}_{\text{aggr}}\text{-rewritable}$ over \mathcal{F} under the universal solutions semantics if there exists an m -ary FO_{aggr} -query Q' over \mathbf{T} such that $\underline{\text{u-certain}}(Q, I) = Q'(\mathcal{F}(I))$ for every instance I of \mathbf{S} .

We say that a query Q over \mathbf{T} is *locally source-dependent under the universal solutions semantics* if there is $d \geq 0$ such that for every instance I of \mathbf{S} and every $\bar{a}, \bar{b} \in \text{dom}(I)^m$, whenever $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$ then

$$(\bar{a} \in \underline{\text{u-certain}}(Q, I) \Leftrightarrow \bar{b} \in \underline{\text{u-certain}}(Q, I)).$$

The next theorem says that Theorem 5.2 extends to the universal solutions semantics.

Theorem 5.3 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be an $\text{FO}_{\text{aggr}}\text{-data exchange setting}$. Every query over \mathbf{T} that is $\text{FO}_{\text{aggr}}\text{-rewritable}$ over the canonical universal solution, or over the core, under the universal solutions semantics, is locally source-dependent, under the universal solutions semantics.* \square

Thus, Theorem 5.3 can be used as a tool for proving non-rewritability under the new semantics (an example is given in the appendix).

We conclude with a result that shows the incomparability of rewritability under the usual semantics and the universal solutions semantics.

Proposition 5.4 *Let \mathcal{F} be either $\mathcal{F}_{\text{univ}}$ or $\mathcal{F}_{\text{core}}$.*

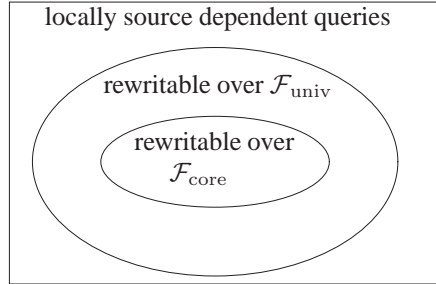
- 1) *There is an FO-query Q that is rewritable (even in FO) over \mathcal{F} under the usual semantics, but is not $\text{FO}_{\text{aggr}}\text{-rewritable}$ over \mathcal{F} under the universal solutions semantics.*
- 2) *There is an FO-query Q that is rewritable (even in FO) over \mathcal{F} under the universal solutions semantics, but is not $\text{FO}_{\text{aggr}}\text{-rewritable}$ over \mathcal{F} under the usual semantics.*

Transformation	LAV Setting		GAV Setting		General Setting	
	locally consistent	locally consistent under \equiv	locally consistent	locally consistent under \equiv	locally consistent	locally consistent under \equiv
canonical universal solution $\mathcal{F}_{\text{univ}}$	yes	yes	no	yes	no	yes
core $\mathcal{F}_{\text{core}}$	yes	yes	no	yes	no	yes

Summary of local consistency results

Rewritable over	in logic	
	FO	FO _{aggr}
$\mathcal{F}_{\text{univ}}$	yes	yes
$\mathcal{F}_{\text{core}}$	yes	yes

Is a query locally-source-dependent?



Summary of rewritability results

Figure 2: Summary of the main results

6 Conclusions

Figure 2 summarizes the main results of the paper. The first table shows when the canonical universal solution and core transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ are locally consistent (“under \equiv ” means “under FO equivalence”, but instead of FO one can use FO_{aggr} as well). The second table gives four classes of locally source-dependent queries, based on the logic and transformation they are rewritable over. The final picture shows the relationship between different classes of rewritable queries. Unlike isolated results on rewriting that exist in the literature, our results give easily applicable tools for studying these notions.

In the future, we would like to develop tools for studying data exchange transformation and query rewriting in the presence of target dependencies, and to extend techniques from relational databases to other data formats.

Acknowledgment We thank Michael Benedikt and Phokion Kolaitis for their comments.

References

- [1] S. Abiteboul, O. Duschka. Complexity of answering queries using materialized views. In *PODS 1998*, pages 254–263.
- [2] S. Abiteboul, R. Hull and V. Vianu. *Foundations of Databases*, Addison Wesley, 1995.
- [3] S. Abiteboul, P. Kanellakis. Object identity as a query language primitive. *J. ACM* 45 (1998), 798–842.
- [4] C. Beeri and M. Y. Vardi. A proof procedure for data dependencies. *J. ACM* 31(4) (1984), 718–741.
- [5] O. Duschka, A. Levy. Recursive plans for information gathering. *IJCAI 1997*, pages 778–784.
- [6] H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer Verlag, 1995.
- [7] R. Fagin. Probabilities on finite models. *J. Symbolic Logic* 41, 1 (1976), 50–58.
- [8] R. Fagin, Ph. Kolaitis, R. Miller, L. Popa. Data exchange: semantics and query answering. In *ICDT’03*, pp. 207–224.
- [9] R. Fagin, Ph. Kolaitis, L. Popa. Data exchange: getting to the core. In *PODS’03*, pages 90–101.
- [10] R. Fagin, L. Stockmeyer, M. Vardi. On monadic NP vs monadic co-NP, *Inf. & Comput.*, 120 (1994), 78–92.
- [11] H. Gaifman. On local and non-local properties, *Proceedings of the Herbrand Symposium, Logic Colloquium ’81*, North Holland, 1982.
- [12] E. Grädel and Y. Gurevich. Metafinite model theory. *Information and Computation* 140 (1998), 26–81.
- [13] A. Halevy. Theory of answering queries using views. *SIGMOD Record* 29 (2000), 40–47.
- [14] W. Hanf. Model-theoretic methods in the study of elementary logic. In J.W. Addison et al, eds, *The Theory of Models*, North Holland, 1965, pages 132–145.
- [15] P. Hell and J. Nešetřil. The core of a graph. *Discrete Math.* 109 (1992), 117–126.
- [16] L. Hella. Logical hierarchies in PTIME. *Information and Computation*, 129 (1996), 1–19.
- [17] L. Hella, L. Libkin, J. Nurmonen and L. Wong. Logics with aggregate operators. *J. ACM*, 48 (2001), 880–907.
- [18] R. Hull, M. Yoshikawa. ILOG: declarative creation and manipulation of object identifiers. *VLDB 1990*, pages 455–468.
- [19] T. Imielinski, W. Lipski. Incomplete information in relational databases. *J. ACM* 31(1984), 761–791.

- [20] A. Klug. Equivalence of relational algebra and relational calculus query languages having aggregate functions. *J. ACM* 29 (1982), 699–717.
- [21] K. Larsen. On grouping in relational algebra. *Int. J. Foundations of Comp. Sci.* 10 (1999), 301–311.
- [22] M. Lenzerini. Data integration: a theoretical perspective. In *PODS'02*, pages 233–246.
- [23] A. Levy, A. Mendelzon, Y. Sagiv, D. Srivastava. Answering queries using views. *PODS'95*, pages 95–104.
- [24] L. Libkin. Logics with counting and local properties. *ACM Trans. on Computational Logic*, 1 (2000), 33–59.
- [25] D. Maier, A. O. Mendelzon, Y. Sagiv. Testing implications of data dependencies. *TODS* 4 (1979), 455–469.
- [26] N. Shu, B. Housel, R. Taylor, S. Ghosh, V. Lum. EXPRESS: a data extraction, processing, and restructuring system. *TODS* 2 (1977), 134–174.
- [27] J. Van den Bussche, D. Van Gucht, M. Andries, M. Gyssens. On the completeness of object-creating database transformation languages. *J. ACM* 44 (1997), 272–319.

A Appendix

Preliminaries: structures, Ehrenfeucht-Fraïssé games, bijective games

In many proofs, we need a definition of structures whose vocabularies include both constant and relation symbols, and games on structures that characterize elementary equivalence in various logics.

A *signature* σ is a collection of constant symbols c_1, \dots, c_n and relation symbols P_1, \dots, P_m , where each relation symbol has an associated arity. A σ -*structure* $\mathbf{A} = (A, P_1^{\mathbf{A}}, \dots, P_m^{\mathbf{A}}, c_1^{\mathbf{A}}, \dots, c_n^{\mathbf{A}})$ consists of a universe A together with an interpretation of each constant c_i of σ as an element $c_i^{\mathbf{A}}$ of A , and each k -ary relation symbol P_i of σ as a set $P_i^{\mathbf{A}} \subseteq A^k$.

If σ is a signature, we write σ_l to refer to the signature that extends σ with l new constant symbols. If \mathbf{A} is a structure and $\bar{a} = (a_1, \dots, a_l)$ is a tuple of elements of A , we write (\mathbf{A}, \bar{a}) for the σ_l -structure in which all the constant and relation symbols from σ are interpreted as in \mathbf{A} , and the new l constant symbols are interpreted as a_1, \dots, a_l .

Note that if σ contains only relation symbols, then a neighborhood $N_d^{\mathbf{A}}(\bar{a})$, with $|\bar{a}| = l$, is viewed as a σ_l -structure, where \bar{a} interpret the l new constant symbols. Hence, isomorphism of neighborhoods must preserve constants.

Let \mathbf{A} and \mathbf{B} be two σ -structures. Then two tuples $\bar{a} = a_1, \dots, a_n$ and $\bar{b} = b_1, \dots, b_n$ define a *partial isomorphism* $\mathbf{A} \rightarrow \mathbf{B}$ if the following hold:

- For any $i, j \leq n$, $a_i = a_j$ iff $b_i = b_j$.
- For any constant symbol c of the signature and $i \leq n$, $a_i = c^{\mathbf{A}}$ iff $b_i = c^{\mathbf{B}}$.
- For any k -ary relation symbol P of the signature and any sequence $[i_1, \dots, i_k]$ of not necessarily distinct numbers from $[1, n]$, $(a_{i_1}, \dots, a_{i_k}) \in P^{\mathbf{A}}$ iff $(b_{i_1}, \dots, b_{i_k}) \in P^{\mathbf{B}}$.

The *Ehrenfeucht-Fraïssé (EF) game* is played by two players, called the *spoiler* and the *duplicator*, on two σ -structures \mathbf{A}, \mathbf{B} . In each round i , the spoiler selects either a point $a_i \in A$, or $b_i \in B$, and the duplicator responds by selecting $b_i \in B$, or $a_i \in A$, respectively. The duplicator wins after n rounds if the relation $\{(a_i, b_i) \mid 1 \leq i \leq n\}$ is a partial isomorphism $\mathbf{A} \rightarrow \mathbf{B}$, otherwise the spoiler wins. It is well known that the duplicator has a winning strategy in the k -round game on \mathbf{A} and \mathbf{B} iff \mathbf{A} and \mathbf{B} agree on all FO sentences of quantifier rank k , that is, iff $\mathbf{A} \equiv_k \mathbf{B}$. Furthermore, the duplicator has a winning strategy in the k -round game on (\mathbf{A}, \bar{a}) and (\mathbf{B}, \bar{b}) iff for every FO formula $\varphi(\bar{x})$ of quantifier rank up to k , $\mathbf{A} \models \varphi(\bar{a})$ iff $\mathbf{B} \models \varphi(\bar{b})$.

A stronger version of the game, called *bijective Ehrenfeucht-Fraïssé game*, was introduced in [16]. Again, the spoiler and the duplicator play on two structures \mathbf{A} and \mathbf{B} . For the n -round game, in each round $i = 1, \dots, n$, the duplicator selects a bijection $f_i : A \rightarrow B$, and the spoiler selects a point $a_i \in A$ (if $|A| \neq |B|$, then the spoiler wins). The winning condition is the same: if after the last round the relation $\{(a_i, f_i(a_i)) \mid 1 \leq i \leq n\}$ is a partial isomorphism $\mathbf{A} \rightarrow \mathbf{B}$, then the duplicator wins; otherwise the spoiler wins. We shall see later that this version of game corresponds to a powerful counting extension of FO [17].

The following will be used widely through the rest of the paper. Given two tuples $\bar{a} = (a_1, \dots, a_n)$ and $\bar{b} = (b_1, \dots, b_m)$, by slightly abusing notation we write $\bar{a} \subseteq \bar{b}$ instead of $\{a_1, \dots, a_n\} \subseteq \{b_1, \dots, b_m\}$, $\bar{a} \cap \bar{b}$ instead of $\{a_1, \dots, a_n\} \cap \{b_1, \dots, b_m\}$, $\bar{a} \cup \bar{b}$ instead of $\{a_1, \dots, a_n\} \cup \{b_1, \dots, b_m\}$ and $\bar{a} \setminus \bar{b}$ instead of $\{a_1, \dots, a_n\} \setminus \{b_1, \dots, b_m\}$. Also, we refer by $\bar{a}\bar{b}$ to the tuple $(a_1, \dots, a_n, b_1, \dots, b_m)$.

Proof of Proposition 3.2

Let $\mathbf{S} = \langle S(\cdot, \cdot) \rangle$, $\mathbf{A} = \langle R(\cdot, \cdot), N(\cdot) \rangle$ and $\mathbf{T} = \langle T(\cdot, \cdot) \rangle$. Assume that \mathcal{R} contains the rules $R(x, z)$, $R(z, y)$, $N(z) :- S(x, y)$ and $T(x, y) :- \exists z(R(x, z) \wedge R(z, y) \wedge N(x) \wedge N(y))$, and that $(\mathbf{S}', \mathbf{T}', \Sigma_{st})$ is an arbitrary data exchange setting. Then there exists a constant k , that is at most the maximum number of conjuncts in the right-hand side of a source-to-target dependency in Σ_{st} , such that for every instance I' of \mathbf{S}' , it is the case that every path in the Gaifman graph of $\mathcal{F}_{\text{univ}}(I')$ containing only null values is of length at most k . Furthermore, by definition of $\mathcal{F}_{\text{core}}$, the same property holds for $\mathcal{F}_{\text{core}}(I')$ [9]. Thus, if we construct an instance I of \mathbf{S} such that the Gaifman graph of $\mathcal{F}_{\Pi}(I)$

contains a path of null values of length $k + 1$, then for every instance I' of \mathbf{S}' , $\mathcal{F}_\Pi(I) \not\cong \mathcal{F}_{\text{univ}}(I')$ and $\mathcal{F}_\Pi(I) \not\cong \mathcal{F}_{\text{core}}(I')$. Next we show how to define such an instance.

Let I be an instance of \mathbf{S} such that $I(S) = \{(a_i, a_{i+1}) \mid i \in [0, k + 2]\}$, where each $a_i \in \text{Const}$. Then $\mathcal{F}_\Pi(I)$ is an instance of \mathbf{T} such that $\mathcal{F}_\Pi(I)(T) = \{(b_i, b_{i+1}) \mid i \in [0, k + 1]\}$, where each b_i is a null value. Thus, the Gaifman graph of $\mathcal{F}_\Pi(I)$ contains a path of null values of length $k + 1$, which concludes the proof of the proposition.

Proof of Lemma 3.4

We start with the following auxiliary result.

Lemma A.1 *For every data exchange program $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ there exists a data exchange program $\Pi' = (\mathbf{S}, \mathbf{A}', \mathbf{T}, \mathcal{R}')$ such that:*

1. *Every rule of \mathcal{R}' has exactly one conjunct in its head.*
2. *Every predicate in $\langle \mathbf{A}', \mathbf{T} \rangle$ appears in the head of exactly one rule of \mathcal{R}' .*
3. *For every instance I of \mathbf{S} , $\mathcal{F}_\Pi(I) \cong \mathcal{F}_{\Pi'}(I)$.*

Moreover, if Π is r -bounded, for some $r \geq 1$, then Π' is r -bounded.

Proof: We show how to transform a rule of the form

$$R_1(\bar{x}_1, \bar{y}_1), \dots, R_k(\bar{x}_k, \bar{y}_k) \quad :- \quad \varphi(\bar{x}), \quad (2)$$

where $\bar{x}_1 \cup \dots \cup \bar{x}_k \subseteq \bar{x}$ and $(\bar{y}_1 \cup \dots \cup \bar{y}_k) \cap \bar{x} = \emptyset$, into an equivalent set of rules of the form shown above. The extension of this technique to the general case is straightforward.

Let $\bar{y} = \bar{y}_1 \cup \dots \cup \bar{y}_k$ and R be an auxiliary $(|\bar{x}| + |\bar{y}|)$ -ary predicate not mentioned in (2). Furthermore, for every $j \in [1, k]$, $\pi_{\bar{x}_j}(\bar{x})$ refers to the components of \bar{x} that occur in the positions of \bar{x}_j , and likewise for $\pi_{\bar{y}_j}$. Then the following set of rules satisfy the conditions stated on the Lemma and they are equivalent to (2):

$$\begin{aligned} R(\bar{x}, \bar{y}) & \quad :- \quad \varphi(\bar{x}), \\ T(\bar{u}, \bar{v}) & \quad :- \quad \bigvee_{j \in [1, k], R_j = T} \exists \bar{x} \exists \bar{y} (R(\bar{x}, \bar{y}) \wedge \bar{u} = \pi_{\bar{x}_j}(\bar{x}) \wedge \bar{v} = \pi_{\bar{y}_j}(\bar{y})), \quad \text{for every } T \in \{R_1, \dots, R_k\}. \end{aligned}$$

We note that if φ is r -bounded, for some $r \geq 1$, then the new set of rules is also r -bounded since all the rules shown above are either r -bounded or 1-bounded. \square

Now we prove Lemma 3.4. Assume that $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ is an r -bounded data exchange program, where \mathcal{R} contains n rules of the form described in Lemma A.1 and $\langle \mathbf{A}, \mathbf{T} \rangle = \langle R_1, \dots, R_n \rangle$, where R_i appears in the head of the i -th rule of \mathcal{R} . We will show by induction on the number n of rules in \mathcal{R} that \mathcal{F}_Π^n is locally consistent. Given that $\mathcal{F}_\Pi(I)$ is the restriction of the predicates in $\mathcal{F}_\Pi^n(I)$ to the predicates in \mathbf{T} , we conclude that \mathcal{F}_Π is locally consistent. We note that \mathcal{F}_Π^n satisfies condition 1) of Definition 3.3 by Lemma A.2 and, hence, in the following paragraphs we only show that \mathcal{F}_Π^n satisfies condition 2) of this definition.

First, assume that $n = 1$, that is, $\mathcal{R} = \{R_1(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})\}$, where $\varphi(\bar{x}, \bar{z})$ is an r -bounded FO formula over \mathbf{S} and $|\langle \bar{x}, \bar{z} \rangle| = l$. Let $d, m > 0$ and define d' as $d \cdot r + \text{lr}(\varphi)$, where $\text{lr}(\varphi)$ is the locality rank of φ , that is, a number that depends on the quantifier rank of φ and l such that for every instance I of \mathbf{S} and $\bar{a}_1, \bar{a}_2 \in \text{dom}(I)^l$, if $N_{\text{lr}(\varphi)}^I(\bar{a}_1) \cong N_{\text{lr}(\varphi)}^I(\bar{a}_2)$, then $I \models \varphi(\bar{a}_1)$ iff $I \models \varphi(\bar{a}_2)$. Let I be an instance of \mathbf{S} and $\bar{a}, \bar{b} \in \text{dom}(I)^m$. We will show that if $N_{d'}^I(\bar{a}) \cong N_{d'}^I(\bar{b})$, then $N_d^{\mathcal{F}_\Pi^1(I)}(\bar{a}) \cong N_d^{\mathcal{F}_\Pi^1(I)}(\bar{b})$.

Let $f : N_{d'}^I(\bar{a}) \rightarrow N_{d'}^I(\bar{b})$ be an isomorphism and define $g : N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{a}) \rightarrow N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{b})$ as follows. Let c be an element (constant or null) of $N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{a})$. Then, given that the rule in the program has one predicate in its head, there exists $P(\bar{c})$ in $N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{a})$, where $P \in \langle \mathbf{S}, R_1 \rangle$, and a sequence $P_1(\bar{c}_1), \dots, P_k(\bar{c}_k)$ of elements of $N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{a})$ such that c is the j -th element of \bar{c} , $k \leq d$, $P_k(\bar{c}_k) = P(\bar{c})$, $\bar{a} \cap \bar{c}_1 \neq \emptyset$ and $\bar{c}_i \cap \bar{c}_{i+1} \cap \text{Const} \neq \emptyset$ ($i \in [1, k-1]$). By definition of $\mathcal{F}_{\Pi}^1(I)$, for every $i \in [1, k]$, there exists a tuple \bar{c}'_i in I such that either $P_i(\bar{c}'_i)$ is in I and $\bar{c}'_i = \bar{c}_i$ (if $P_i \in \mathbf{S}$) or $I \models \varphi(\bar{c}'_i)$ and $R_1(\bar{c}_i) :- \varphi(\bar{c}'_i)$ is an instantiation of the rule in \mathcal{R} (if $P_i = R_1$). Given that φ is r -bounded, for every $i \in [1, k]$, all the elements in \bar{c}'_i are at distance at most $d \cdot r$ from \bar{a} in $N_{d'}^I(\bar{a})$. Moreover, it can be seen that for any tuple \bar{e} of elements in $N_{d \cdot r}^I(\bar{a})$, $N_{\text{lr}(\varphi)}^I(\bar{e}) \cong N_{\text{lr}(\varphi)}^I(f(\bar{e}))$ since $d' = d \cdot r + \text{lr}(\varphi)$ and $N_{d'}^I(\bar{a}) \cong N_{d'}^I(\bar{b})$. Thus, for every $i \in [1, n]$, if $P_i(\bar{c}'_i)$ is in I , $P_i(f(\bar{c}'_i))$ is in I , and if $I \models \varphi(\bar{c}'_i)$, then $I \models \varphi(f(\bar{c}'_i))$. Let $P_1(\bar{e}_1), \dots, P_k(\bar{e}_k)$ be a sequence of elements of $N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{b})$ generated from $f(\bar{c}'_1), \dots, f(\bar{c}'_k)$ by using the same rules that were used to generate $P_1(\bar{c}_1), \dots, P_k(\bar{c}_k)$ from $\bar{c}'_1, \dots, \bar{c}'_k$. We define $g(c)$ as the j -th element of \bar{e}_k .

It is not hard to see that g is an isomorphism and, therefore, $N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{a}) \cong N_d^{\mathcal{F}_{\Pi}^1(I)}(\bar{b})$.

We now consider the inductive step. Assume that the property holds for every data exchange program containing at most $n-1$ rules and assume that \mathcal{R} contains rules r_1, \dots, r_n whose heads mention predicates R_1, \dots, R_n , respectively. Given that the transformation \mathcal{F}_{Π}^n for the program $(\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ is the composition of the transformation $\mathcal{F}_{\Pi_1}^{n-1}$ for the program $\Pi_1 = (\mathbf{S}, \mathbf{A}, \langle R_1, \dots, R_{n-1} \rangle, \langle r_1, \dots, r_{n-1} \rangle)$ and the transformation $\mathcal{F}_{\Pi_2}^1$ for the program $\Pi_2 = (\langle \mathbf{S}, R_1, \dots, R_{n-1} \rangle, \emptyset, \langle R_n \rangle, \langle r_n \rangle)$, and these two transformations are locally consistent by induction hypothesis, we conclude that the transformation \mathcal{F}_{Π}^n for $(\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ is also locally consistent.

Proof of Theorem 3.5

That $\mathcal{F}_{\text{univ}}$ is locally consistent follows from Lemma 3.4 since LAV settings are 1-bounded. Thus, we only need to show that $\mathcal{F}_{\text{core}}$ is locally consistent for LAV settings. The proof proceeds by making use of the algorithm given in [9] for computing the core in the LAV setting.

Let J be an instance with nulls. The *Gaifman graph of the nulls of J* is an undirected graph in which (1) the nodes are all the nulls of J , and (2) there is an edge between two nulls whenever the nulls belong to the same tuple of some relation in J . A *block* of nulls is the set of nulls in a connected component of the Gaifman graph of nulls. If v is a null of J , then we may refer to the block of nulls that contains v as the *block of v* . Note that, by the definition of blocks, the set of all nulls of J is partitioned into disjoint blocks.

Let h be a homomorphism of an instance J . Denote the result of applying h to J by $h(J)$. If $h(J)$ is a subinstance of J , then we call h an *endomorphism* of J . An endomorphism h of J is *useful* if $h(J) \neq J$ (i.e., $h(J)$ is a proper subinstance of J).

Let J and J' be two instances such that the nulls of J' form a subset of the nulls of J . Let h be some endomorphism of J' , and let B be a block of nulls of J . We say that h is *J -local for B* if $h(x) = x$ whenever $x \notin B$. (Since all the nulls of J' are among the nulls of J , it makes sense to consider whether or not a null x of J' belongs to the block B of J .) We say that h is *J -local* if it is J -local for B , for some block B of J .

We now present the algorithm of [9] for computing the core of the universal solutions in the LAV setting, when given the canonical universal solution J .

1. Compute the blocks of J , and initialize J' to be J .
2. Check whether there exists a useful J -local endomorphism h of J' . If not, then stop with result J' .
3. Update J' to be $h(J')$, and return to Step 2.

Let b be the maximal number of existentially quantified variables over all stds in Σ_{st} . As noted in [9], it follows easily from the construction of the canonical universal solution J (by chasing with Σ_{st}) that b is an upper bound on the size of a block in J .

By Lemma 3.4, the canonical universal solution transformation $\mathcal{F}_{\text{univ}}$ of a LAV setting is locally consistent. We shall show that the mapping that maps the canonical universal solution onto the core is locally consistent. Since the composition of locally consistent transformations is locally consistent, this is enough to prove the theorem. Let I be a source database, let J be a canonical universal solution, and let J_0 be the core of J . Assume $m \geq 0$ and $d \geq 1$ (we do not allow $d = 0$ for technical convenience, and it is clear that this restriction is unimportant). We need only show that whenever \bar{a} and \bar{b} are m -tuples with isomorphic $(d + b - 1)$ -neighborhoods in the canonical universal solution J , then \bar{a} and \bar{b} have isomorphic d -neighborhoods in the core. To simplify the wording, let us phrase this by saying that we need only show that the $(d + b - 1)$ -neighborhood $N_{d+b-1}^J(\bar{a})$ determines the d -neighborhood $N_d^{J_0}(\bar{a})$.

Since we are assuming a LAV setting, the stds in Σ_{st} are of the form $\forall \bar{x}(R(\bar{x}) \rightarrow \exists \bar{y}\varphi_{\mathbf{T}}(\bar{x}, \bar{y}))$, where $R(\bar{x})$ is an atomic formula, and where $\varphi_{\mathbf{T}}(\bar{x}, \bar{y})$ is a conjunction of FO atomic formulae. Let v be a null in $N_d^J(\bar{a})$. Then there is such a std σ in Σ_{st} and there is a tuple \bar{e} of constants where $R(\bar{e})$ holds in the source instance I , such that v , and every member of its block, is generated by chasing σ starting with $R(\bar{e})$. If every member of \bar{e} were at least distance d from \bar{a} in I (that is, outside of $N_d^I(\bar{a})$) then it is easy to see that v would be at least distance $d + 1$ from \bar{a} in J . But this is false, since $v \in N_d^J(\bar{a})$. So some member of \bar{e} is in $N_{d-1}^I(\bar{a})$, and hence every member of \bar{e} is in $N_d^I(\bar{a})$. We see from the definition of b that chasing with σ causes every member of the block of v to be in $N_{d+b-1}^J(\bar{a})$. Hence, if h is an endomorphism of J , then every member of the block B of v is mapped into $N_{d+b-1}^J(\bar{a})$ (this is because paths of length, say, m , that begin with a member a of \bar{a} are mapped by each endomorphism into paths of length at most m that begin with a). So every J -local endomorphism maps the nulls of B (each of which is in $N_{d+b-1}^J(\bar{a})$) into points in $N_{d+b-1}^J(\bar{a})$.

We now show that there is enough information in $N_{d+b-1}^J(\bar{a})$ to produce $N_d^{J_0}(\bar{a})$. Hence, the $(d + b - 1)$ -neighborhood $N_{d+b-1}^J(\bar{a})$ determines the d -neighborhood $N_d^{J_0}(\bar{a})$, which as we noted is sufficient to prove the theorem.

Let us call a block *special* if it contains a null in $N_d^J(\bar{a})$. Consider a modified version of the algorithm for producing the core where the special blocks are selected first. Thus, the modified version of the algorithm selects a non-special block with a useful J -local endomorphism in Step 2 only when there is no special block with a useful J -local endomorphism. We can think of this algorithm as consisting of two phases. In the first phase, only special blocks are selected, and in the second phase, non-special blocks are selected. Since, as we showed, every J -local endomorphism maps the nulls of a special block B (each of which is in $N_{d+b-1}^J(\bar{a})$) into points in $N_{d+b-1}^J(\bar{a})$, it follows easily that there is enough information in $N_{d+b-1}^J(\bar{a})$ to carry out the first phase of the algorithm. Let C be the neighborhood about \bar{a} of radius d in J' at the end of the first phase (J' is as defined in the algorithm). It is fairly easy to see that C is also the neighborhood about \bar{a} of radius d in J' at the end of the second phase (intuitively, no changes take place in C in the second phase, because a useful J -local endomorphism can only remove tuples, not add tuples.) Since J' at the end of the second phase is the core, it follows that C is $N_d^{J_0}(\bar{a})$. So indeed, there is enough information in $N_{d+b-1}^J(\bar{a})$ to produce $N_d^{J_0}(\bar{a})$, which was to be shown.

Proof of Proposition 3.6

Since in the GAV setting, $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ coincide (since no nulls are produced), it suffices to prove the result for $\mathcal{F}_{\text{univ}}$.

(a) Consider a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, where $\mathbf{S} = \langle E(\cdot, \cdot), U(\cdot) \rangle$, $\mathbf{T} = \langle R(\cdot, \cdot, \cdot) \rangle$ and Σ_{st} contains a single dependency $\{E(x, y) \wedge U(z) \rightarrow R(x, y, z)\}$. We will show that if $d = 2$, then for every $d' \geq 0$ there exists an instance I of \mathbf{S} and elements $a, b \in \text{dom}(I)$ for which $N_{d'}^I(a) \cong N_{d'}^I(b)$ and $N_d^{\mathcal{F}_{\text{univ}}(I)}(a) \not\cong N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$. For a given $d' \geq 0$ set I to be the disjoint union of a point c under predicate U and two successor relations S_1 and S_2 under predicate E of length $2d' + 2$ and $2d' + 4$ respectively. Choose a to be the middle point of S_1 and b the middle point of S_2 . Then $N_{d'}^I(a) \cong N_{d'}^I(b)$ but $N_d^{\mathcal{F}_{\text{univ}}(I)}(a) \not\cong N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$.

For the second statement of part (a), consider a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ with $\mathbf{S} = \langle E(\cdot, \cdot), R(\cdot), G(\cdot), B(\cdot) \rangle$, and $\mathbf{T} = \langle T(\cdot, \cdot) \rangle$. The set Σ_{st} has the following stds:

$$\begin{aligned} R(x) \wedge G(y) &\rightarrow T(x, y) \\ R(x) \wedge B(y) \wedge E(x, y) &\rightarrow T(x, y) \\ B(x) \wedge G(y) \wedge E(x, y) &\rightarrow T(x, y) \\ G(x) \wedge G(y) \wedge E(x, y) &\rightarrow T(x, y) \end{aligned}$$

In this case $\mathcal{F}_{\text{univ}}(I)$ is the result of the evaluation of the following query in I :

$$(R(x) \wedge G(y)) \vee (R(x) \wedge B(y) \wedge E(x, y)) \vee (B(x) \wedge G(y) \wedge E(x, y)) \vee (G(x) \wedge G(y) \wedge E(x, y)).$$

Now, for each $n > 1$, we create an instance I_n of S as follows. Its universe is $\{c, c', a_1, \dots, a_n, b_1, \dots, b_{n+1}\}$. The interpretation of R is $\{c, c'\}$, the interpretation of B is $\{a_1, b_1\}$ and the interpretation of G is $\{a_2, \dots, a_n, b_2, \dots, b_{n+1}\}$. The relation E includes the following edges:

$$(c, a_1), (c', b_1), (a_i, a_{i+1}), i < n, (b_i, b_{i+1}), i \leq n.$$

Suppose that the transformation $\mathcal{F}_{\text{univ}}$ for this data exchange setting is locally consistent. Choose d so that $N_d^I(a) \cong N_d^I(b)$ would imply $N_1^{\mathcal{F}_{\text{univ}}(I)}(a) \cong N_1^{\mathcal{F}_{\text{univ}}(I)}(b)$, and let $n > d$. Then $N_d^{I_n}(c) \cong N_d^{I_n}(c')$, but it is easy to see that, in $\mathcal{F}_{\text{univ}}(I_n)$, the 1-neighborhoods of c and c' are not isomorphic.

(b) Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, suppose Σ_{st} contains a single dependency $\varphi_{\mathbf{S}}(x, y) \rightarrow T(x, y)$, where $\varphi_{\mathbf{S}}$ is a conjunctive query $\exists \bar{z} \bigwedge_i \alpha_i(x, y, \bar{z})$ and each α_i is an atom. In this case $\mathcal{F}_{\text{univ}}(I)$ is the result of $\varphi_{\mathbf{S}}$ on I . We note that $\mathcal{F}_{\text{univ}}$ satisfies condition 1) of definition 3.3 by Lemma A.2 and, hence, in the following paragraphs we only show that $\mathcal{F}_{\text{univ}}$ satisfies condition 2) of this definition.

Let $G(\varphi_{\mathbf{S}})$ be the graph of this conjunctive query, where the nodes are the variables x, y, \bar{z} , and there is an edge between them if they belong to the same atom. Suppose x and y are in the same connected component of $G(\varphi_{\mathbf{S}})$. Let d be the distance between them in $G(\varphi_{\mathbf{S}})$. Then the data exchange program corresponding to this data exchange setting is d -bounded, and hence, by Lemma 3.4, defines a locally consistent transformation.

Now assume x and y are in two different connected components of $G(\varphi_{\mathbf{S}})$. Then $\varphi_{\mathbf{S}}(x, y)$ is equivalent to $\alpha(x) \wedge \beta(y)$, where α and β are FO (in fact, conjunctive) queries. If $\alpha(I)$ and $\beta(I)$ are the sets of elements satisfying α and β respectively, then $\mathcal{F}_{\text{univ}}(I) = \alpha(I) \times \beta(I)$. In particular, for any $r \geq 1$ and any $a, N_r^{\mathcal{F}_{\text{univ}}(I)}(a) = N_1^{\mathcal{F}_{\text{univ}}(I)}(a)$.

By Gaifman's theorem [11], there is a number d that depends on α and β (and thus is determined by $\varphi_{\mathbf{S}}$) such that if $N_d^I(a) \cong N_d^I(b)$, then a and b agree on α and β . Hence, $N_d^I(a) \cong N_d^I(b)$ implies that $N_1^{\mathcal{F}_{\text{univ}}(I)}(a) \cong N_1^{\mathcal{F}_{\text{univ}}(I)}(b)$ and thus $N_r^{\mathcal{F}_{\text{univ}}(I)}(a) \cong N_r^{\mathcal{F}_{\text{univ}}(I)}(b)$ for any r . This finishes the proof.

Example: core and local consistency

We mentioned that without the requirement that constants be mapped to constants, $\mathcal{F}_{\text{core}}$ need not be locally consistent. We now give an example of this. Define a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ as follows: $\mathbf{S} = \langle E(\cdot, \cdot), R(\cdot), G(\cdot) \rangle$, $\mathbf{T} = \langle E'(\cdot, \cdot), R'(\cdot), G'(\cdot) \rangle$ and $\Sigma_{st} = \{E(x, y) \rightarrow E'(x, y), R(x) \rightarrow R'(x), G(x) \rightarrow G'(x)\}$. For each n , define the instance I_n of \mathbf{S} whose domain is $\{a_j^i \mid 1 \leq i \leq 4, 1 \leq j \leq n\} \cup \{c, c'\}$. The relation E has edges (a_j^i, a_{j+1}^i) for each $i \leq 4$, and $j < n$. Furthermore, there are edges

$$(a_n^1, c), (a_n^2, c), (a_n^3, c'), (a_n^4, c').$$

The relation G is interpreted as $\{a_1^1, a_1^2, a_1^3\}$ and R as $\{a_1^4\}$.

Let J_n be the canonical universal solution of I_n ($n \geq 0$). We note that for any d and $n > d$, $N_d^{J_n}(c) \cong N_d^{J_n}(c')$.

Without the requirement that a homomorphism h be identity on the constants (which is the definition used in graph theory [15]), $\mathcal{F}_{\text{core}}(I_n)$ is isomorphic to the substructure of J_n obtained by removing $\{a_j^2 \mid 1 \leq j \leq n\}$. But in this structure, even $N_1^{\mathcal{F}_{\text{core}}(I_n)}(c)$ and $N_1^{\mathcal{F}_{\text{core}}(I_n)}(c')$ are not isomorphic. \square

Proof of Lemma 3.8

We begin by proving a lemma and two propositions.

Lemma A.2 *For every data exchange program $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ with just one rule $R(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})$, there exists $d, k \geq 0$ such that for every instance I of \mathbf{S} and every $a, b \in \text{dom}(I)$, if $N_d^I(a) \equiv_k N_d^I(b)$, then $a \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$ iff $b \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$.*

Proof: It is known that for every FO formula $\psi(\bar{u})$, where $|\bar{u}| = m$, there exists d', k' such that, for every instance I and m -tuples \bar{a}, \bar{b} , if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$, then $I \models \psi(\bar{a})$ iff $I \models \psi(\bar{b})$. Furthermore, d' and k' are functions of m , the quantifier rank of $\psi(\bar{u})$ ($\text{qr}(\psi)$) and m_ψ , where m_ψ is the maximum arity of a predicate in ψ . We denote values d', k' by $f(m, \text{qr}(\psi), m_\psi)$ and $g(m, \text{qr}(\psi), m_\psi)$, respectively.

Define $d = f(1, \text{qr}(\varphi) + |\bar{x}| + |\bar{z}|, m_{\mathbf{S}})$ and $k = g(1, \text{qr}(\varphi) + |\bar{x}| + |\bar{z}|, m_{\mathbf{S}})$, where $m_{\mathbf{S}}$ is the maximum arity of a predicate in \mathbf{S} . We will show that for every instance I of \mathbf{S} and every $a, b \in \text{dom}(I)$, if $N_d^I(a) \equiv_k N_d^I(b)$, then $a \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$ iff $b \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$. Assume that $N_d^I(a) \equiv_k N_d^I(b)$ and that $a \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$. Then without loss of generality we can assume that $I \models \exists \bar{u} \exists \bar{z} \varphi(a, \bar{u}, \bar{z})$, where $|\bar{u}| = |\bar{x}| - 1$. Thus, by definition of d and k and considering that $\text{qr}(\exists \bar{u} \exists \bar{z} \varphi(a, \bar{u}, \bar{z})) = |\bar{u}| + |\bar{z}| + \text{qr}(\varphi) < |\bar{x}| + |\bar{z}| + \text{qr}(\varphi)$, $I \models \exists \bar{u} \exists \bar{z} \varphi(b, \bar{u}, \bar{z})$ and, therefore, $b \in \text{dom}(\mathcal{F}_{\Pi}^n(I))$. \square

Proposition A.3 *Let $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ be a data exchange program with just one rule of the form $R(\bar{x}) :- \varphi(\bar{x})$. Then \mathcal{F}_{Π}^n is locally consistent under FO-equivalence.*

Proof: Condition 1) of definition 3.7 is satisfied by Lemma A.2. To prove that condition 2) of this definition is also satisfied we need to introduce some terminology. We know that for an arbitrary (finite) instance I , an arbitrary m -tuple \bar{a} in $\text{dom}(I)$ and every $d, k \geq 0$, there exists an FO formula $\theta_{I, \bar{a}}^{d, k}(\bar{u})$ such that, for every m -tuple \bar{b} in $\text{dom}(I)$, $I \models \theta_{I, \bar{a}}^{d, k}(\bar{b})$ iff $N_d^I(\bar{a}) \equiv_k N_d^I(\bar{b})$. Furthermore, for every instance I and m -tuple \bar{a} , the quantifier rank of $\theta_{I, \bar{a}}^{d, k}$ is a function of d and k .

Let I be an instance of \mathbf{S} , $m_{\mathbf{S}}$ the maximum arity of a predicate in \mathbf{S} and $\psi(\bar{u})$ the formula obtained from $\theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{u})$ by replacing each occurrence of R by φ , where $\bar{a} \in \text{dom}(I)^m$. Define $d' = f(m, \text{qr}(\psi), m_{\mathbf{S}})$ and $k' = g(m, \text{qr}(\psi), m_{\mathbf{S}})$, where f and g are as in the proof of Lemma A.2. We observe that d' and k' are independent of I and \bar{a} since the quantifier rank of $\theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}$ is a function of d and k . We will show that for every instance I of \mathbf{S} and pair of tuples $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$ then $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. By definition of d' and k' , if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$, then $I \models \psi(\bar{a})$ iff $I \models \psi(\bar{b})$. But for every tuple $\bar{c} \in \text{dom}(I)^m$, $I \models \psi(\bar{c})$ iff $\mathcal{F}_{\Pi}^n(I) \models \theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{c})$. Thus, if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$, then $\mathcal{F}_{\Pi}^n(I) \models \theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{a})$ iff $\mathcal{F}_{\Pi}^n(I) \models \theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{b})$. Since it is always the case that $\mathcal{F}_{\Pi}^n(I) \models \theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{a})$, if $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$ then $\mathcal{F}_{\Pi}^n(I) \models \theta_{\mathcal{F}_{\Pi}^n(I), \bar{a}}^{d, k}(\bar{b})$ and, therefore, $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. \square

Proposition A.4 *Let $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ be a data exchange program with just one rule of the form $R(\bar{x}, \bar{y}) :- S(\bar{x}, \bar{z})$, where $S \in \mathbf{S}$ and $\bar{y} \neq \emptyset$. Then \mathcal{F}_{Π}^n is locally consistent under FO-equivalence.*

Proof: Condition 1) of definition 3.7 is trivially satisfied. Next we prove that condition 2) of this definition is also satisfied.

Let $d, k \geq 0$. Choose $d' = d$ and $k' = m_{\mathbf{S}} \cdot (k + d)$, where $m_{\mathbf{S}}$ is the maximum arity of a predicate in \mathbf{S} . We will show that $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$ whenever $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$. In fact, we will show how to play a k -round game between $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$ from a k' -round game between $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$. The game is as follows. In each round $i \in [1, k]$ the spoiler chooses one element c_i in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ (the case he/she chooses an element c_i in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$ is completely symmetric).

- If $c_i \notin \text{dom}(I)$, then it can be unequivocally identified with an instantiation $R(\bar{c}, c_i, \bar{n}) :- S(\bar{c}, \bar{e})$ of the rule of the program. It is not hard to see that $S(\bar{c}, \bar{e}) \in N_{d'}^I(\bar{a})$. Then in the game between $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$ the spoiler plays every element in \bar{c} and \bar{e} (the number of elements in $\bar{c} \cup \bar{e}$ is at most $m_{\mathbf{S}}$), and the duplicator responds with a tuple $\bar{c}' \bar{e}'$ in $N_{d'}^I(\bar{b})$. Since $N_{d'}^I(\bar{a}) \equiv_{k'} N_{d'}^I(\bar{b})$ and $k' = m_{\mathbf{S}} \cdot (k + d)$, $S(\bar{c}', \bar{e}') \in N_{d'}^I(\bar{b})$ and, therefore, the null c_i is naturally and univocally associated with an element $c'_i \notin \text{dom}(I)$ through the instantiation $R(\bar{c}', c'_i, \bar{n}') :- S(\bar{c}', \bar{e}')$. Set c'_i to be the duplicator response to c_i in the i -th round of the game between $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$.
- If $c_i \in \text{dom}(I)$, we know that it also belongs to $N_{d'}^I(\bar{a})$. The spoiler plays c_i in the game between $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$. The duplicator responds with c'_i in $N_{d'}^I(\bar{b})$. Set c'_i to be the duplicator response to c_i in the i -th round of the game between $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$.

Claim A.5 *The response c'_i to the element c_i , defined by the above strategy, is in $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$.*

Proof: The proof of this is direct from the fact that $k' \geq m_S \cdot d$. □

We show that the strategy shown above is a winning strategy for the duplicator by induction on the move $i \leq k$. For $i = 0$ the proof goes as follows. By contradiction assume that (\bar{a}, \bar{b}) is not a partial isomorphism in $\mathcal{F}_\Pi^n(I)$. Then without loss of generality, we can assume that there exists $T(\bar{a}') \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$ such that $T(\bar{b}') \notin N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, where $\bar{a}' \subseteq \bar{a}$, \bar{b}' is the corresponding subset of \bar{b} and $T \in \langle \mathbf{S}, R \rangle$. Then $T \neq R$ since tuples \bar{a} and \bar{b} contain only elements in $\text{dom}(I)$. Moreover, $T \notin \mathbf{S}$ since (\bar{a}, \bar{b}) is a partial isomorphism between $N_d^I(\bar{a})$ and $N_d^I(\bar{b})$, which leads to a contradiction.

Assume that $i - 1$ moves of the game have been played by following the strategy shown above. For the i -th move, $i \in [1, k]$, the spoiler chooses c_i in $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$. Call $\bar{c} = c_1, \dots, c_{i-1}$ and $\bar{c}' = c'_1, \dots, c'_{i-1}$ to the elements played so far in the game between $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$. Suppose first $c_i \in \text{dom}(I)$. Then c_i is in $N_d^I(\bar{a})$. Assume on the contrary that c'_i does not work as a winning duplicator response for the game between $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$. Then without loss of generality, we can assume that there is a tuple $T(\bar{e}, c_i, \bar{n}) \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$, such that $T(\bar{e}', c'_i, \bar{n}') \notin N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, where $\bar{e} \subseteq \bar{a}\bar{c} \cap \text{dom}(I)$, $\bar{n} \subseteq \bar{c} \setminus \text{dom}(I)$, and \bar{e}' and \bar{n}' are the responses to \bar{e} and \bar{n} in $\bar{b}\bar{c}'$. If $T \in \mathbf{S}$ then \bar{n} is empty, and $T(\bar{e}', c'_i) \in N_d^I(\bar{b})$. This implies by induction hypothesis that $T(\bar{e}', c'_i) \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, which is a contradiction. If $T = R$, we know that $T(\bar{e}, c_i, \bar{n}) = R(\bar{e}, c_i, \bar{n})$ comes from an instantiation $R(\bar{e}, c_i, \bar{n}) :- S(\bar{e}, c_i, \bar{e}_1)$ of the rule of the program. Since $|\bar{n}| \geq 1$, all the elements of \bar{e}_1 were already played in the game between $N_d^I(\bar{a})$ and $N_d^I(\bar{b})$. Let \bar{e}'_1 in $N_d^I(\bar{b})$ be the response of the duplicator to \bar{e}_1 . Then $S(\bar{e}', c'_i, \bar{e}'_1) \in N_d^I(\bar{b})$ and, therefore, by Claim A.5 we obtain that $c'_i \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, and by induction hypothesis $T(\bar{e}', c'_i, \bar{n}') \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, which leads to a contradiction.

Assume now $c_i \notin \text{dom}(I)$. Then c_i comes from an instantiation $R(\bar{e}, c_i, \bar{n}) :- S(\bar{e}, \bar{e}_1)$ of the rule of the program, where $\bar{e}\bar{e}_1 \subseteq \text{dom}(I)$ and \bar{n} is a tuple of values not in $\text{dom}(I)$. To c_i we attach its natural response c'_i as previously described. Suppose on the contrary that c'_i does not work as a winning duplicator response for the game between $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$. Then $\bar{e} \subseteq \bar{a}\bar{c} \cap \text{dom}(I)$, $\bar{n} \subseteq \bar{c} \setminus \text{dom}(I)$, $R(\bar{e}, c_i, \bar{n}) \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{a})$ and $R(\bar{e}', c'_i, \bar{n}') \notin N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, where \bar{e}' and \bar{n}' are the responses to \bar{e} and \bar{n} in $\bar{b}\bar{c}'$. Since $c_i \notin \text{dom}(I)$, all the elements of \bar{e}_1 were already played in the game between $N_d^I(\bar{a})$ and $N_d^I(\bar{b})$. Let \bar{e}'_1 in $N_d^I(\bar{b})$ be the response of the duplicator to \bar{e}_1 . Then $S(\bar{e}', c'_i, \bar{e}'_1) \in N_d^I(\bar{b})$ and, therefore, by Claim A.5 we know that $c'_i \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, and by induction hypothesis $R(\bar{e}', c'_i, \bar{n}') \in N_d^{\mathcal{F}_\Pi^n(I)}(\bar{b})$, which leads to a contradiction. □

Proof of Lemma 3.8: Let $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ be a data exchange program and assume that \mathcal{R} contains n rules of the form described in Lemma A.1. We note that for every rule $R(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})$ in the program, if $\bar{y} = \emptyset$, then this rule is equivalent to a rule $R(\bar{x}) :- \exists \bar{z} \varphi(\bar{x}, \bar{z})$. Otherwise $\bar{y} \neq \emptyset$ and $R(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})$ can be decomposed into two different rules: $R'(\bar{x}, \bar{z}) :- \varphi(\bar{x}, \bar{z})$ and $R(\bar{x}, \bar{y}) :- R'(\bar{x}, \bar{z})$. Then we can assume without loss of generality that \mathcal{R} contains only rules of the form $T(\bar{x}) :- \varphi(\bar{x})$, where $\varphi(\bar{x})$ is a first-order formula over $\langle \mathbf{S}, \mathbf{A} \rangle$ and $T \in \langle \mathbf{A}, \mathbf{T} \rangle$, or of the form $T(\bar{x}, \bar{y}) :- S(\bar{x}, \bar{z})$, where $\bar{y} \neq \emptyset$, $S \in \langle \mathbf{S}, \mathbf{A} \rangle$ and $T \in \langle \mathbf{A}, \mathbf{T} \rangle$. We will prove by induction on the number n of rules of \mathcal{R} that \mathcal{F}_Π^n is locally consistent under FO-equivalence. Given that for every instance I of \mathbf{S} , $\mathcal{F}_\Pi^n(I)$ is the restriction of the predicates in $\mathcal{F}_\Pi^n(I)$ to the predicates in \mathbf{T} , we conclude that \mathcal{F}_Π^n is locally consistent under FO-equivalence. We note that \mathcal{F}_Π^n satisfies condition 1) of definition 3.7 by Lemma A.2 and, hence, in the following paragraph we only show that \mathcal{F}_Π^n satisfies condition 2) of this definition.

For the basis case, take the number of rules in Π to be one. Suppose first that the rule in \mathcal{R} is of the form $R(\bar{x}) :- \varphi(\bar{x})$. Then from Proposition A.3 we know that \mathcal{F}_Π^n is locally consistent under FO-equivalence. Suppose now the rule in \mathcal{R} is of the form $T(\bar{x}, \bar{y}) :- R(\bar{x}, \bar{z})$, where $\bar{y} \neq \emptyset$. Then from Proposition A.4 we obtain that \mathcal{F}_Π^n is locally consistent under FO-equivalence. We now consider the inductive step. Assume that the property holds for every data exchange program containing at most $n - 1$ rules and assume that \mathcal{R} contains rules r_1, \dots, r_n whose heads mention predicates R_1, \dots, R_n , respectively. Given that the transformation \mathcal{F}_Π^n for the program $(\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ is the composition of the transformation $\mathcal{F}_{\Pi_1}^{n-1}$ for the program $\Pi_1 = (\mathbf{S}, \mathbf{A}, \langle R_1, \dots, R_{n-1} \rangle, \langle r_1, \dots, r_{n-1} \rangle)$ and the transformation $\mathcal{F}_{\Pi_2}^1$ for the program $\Pi_2 = (\langle \mathbf{S}, R_1, \dots, R_{n-1} \rangle, \emptyset, \langle R_n \rangle, \langle r_n \rangle)$, and these two transformations are locally consistent under FO-equivalence by induction hypothesis, we conclude that the transformation \mathcal{F}_Π^n for $(\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ is also locally consistent under FO-equivalence. □

Proof of Theorem 3.9

The proof for $\mathcal{F}_{\text{univ}}$ is a direct consequence of Lemma 3.8. The proof for $\mathcal{F}_{\text{core}}$ is postponed until we obtain the proof of Theorem 4.8.

Proof of Proposition 3.10

(a) Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st}, \Sigma_t)$ be a LAV+tgcd (GAV+tgcd) setting defined as follows: $\mathbf{S} = \langle S(\cdot, \cdot), M(\cdot) \rangle$, $\mathbf{T} = \langle T(\cdot, \cdot), N(\cdot) \rangle$, $\Sigma_{st} = \{S(x, y) \rightarrow T(x, y), M(x) \rightarrow N(x)\}$ and $\Sigma_t = \{T(x, y) \wedge T(y, z) \rightarrow T(x, z)\}$. Let $m = 1$, $d = 1$ and $k = 1$. We will show that for these values there is no $d', k' \geq 0$ such that for every instance I of S and for every $a, b \in \text{dom}(I) \cap \text{dom}(\mathcal{F}_{\text{univ}}(I))$, if $N_{d'}^I(a) \equiv_{k'} N_{d'}^I(b)$, then $N_d^{\mathcal{F}_{\text{univ}}(I)}(a) \equiv_k N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$. On the contrary, assume that such d', k' exist and let I be an instance of \mathbf{S} defined as the disjoint union of two successor relations of length $d' + 1$: $I(S) = \{(a_i, a_{i+1}) \mid 1 \leq i \leq d'\} \cup \{(b_i, b_{i+1}) \mid 1 \leq i \leq d'\}$. Furthermore, assume that $I(M) = \{a_{d'+1}\}$. Then $N_{d'}^I(a_1) \equiv_{k'} N_{d'}^I(b_1)$ since $N_{d'}^I(a_1) \cong N_{d'}^I(b_1)$. In this case $\mathcal{F}_{\text{univ}}(I)$ is the disjoint union of two linear orders: $\mathcal{F}_{\text{univ}}(I)(T) = \{(a_i, a_j) \mid 1 \leq i < j \leq d'\} \cup \{(b_i, b_j) \mid 1 \leq i < j \leq d'\}$. Thus, $N_d^{\mathcal{F}_{\text{univ}}(I)}(a_1) \not\equiv_k N_d^{\mathcal{F}_{\text{univ}}(I)}(b_1)$ since $\mathcal{F}_{\text{univ}}(I)(N) = \{a_{d'+1}\}$ and $a_{d'+1}$ is at distance 1 from a_1 .

We note that the previous proof also shows that the transformation $\mathcal{F}_{\text{core}}$ of LAV+tgcd (GAV+tgcd) settings is not necessarily locally consistent (under FO-equivalence) since in the setting shown above $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ coincide.

(b) Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st}, \Sigma_t)$ be a GAV+egd setting, where Σ_t is a set of equality generating dependencies over \mathbf{T} , and I an instance of \mathbf{S} . If I has a canonical universal solution J , then J is also a canonical universal solution of I in the GAV setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$. Thus, $(\mathbf{S}, \mathbf{T}, \Sigma_{st}, \Sigma_t)$ is locally consistent under FO-equivalence by Corollary 3.9.

Given that the transformations $\mathcal{F}_{\text{univ}}$ and $\mathcal{F}_{\text{core}}$ of GAV+egd settings coincide, the previous proof also shows that the transformation $\mathcal{F}_{\text{core}}$ of GAV+egd settings is locally consistent under FO-equivalence.

(c) Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st}, \Sigma_t)$ be a LAV+egd setting, where $\mathbf{S} = \langle E(\cdot, \cdot), V(\cdot) \rangle$, $\mathbf{T} = \langle E'(\cdot, \cdot), V'(\cdot), R_1(\cdot, \cdot), R_2(\cdot, \cdot) \rangle$, Σ_{st} contains the following source-to-target dependencies:

$$\begin{aligned} E(x, y) &\rightarrow E'(x, y), \\ V(x) &\rightarrow V'(x), \\ E(x, y) &\rightarrow \exists u_1 \exists u_2 \exists u_3 (R_1(x, u_1) \wedge R_1(y, u_2) \wedge R_2(u_1, u_3) \wedge R_2(u_2, u_3)), \end{aligned}$$

and Σ_t contains the following key dependencies:

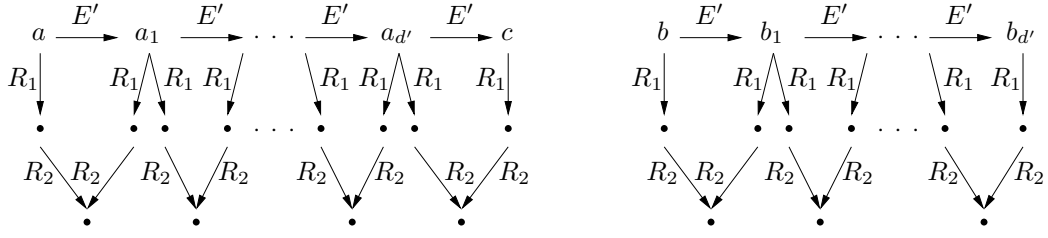
$$\begin{aligned} R_1(x, y) \wedge R_1(x, z) &\rightarrow y = z, \\ R_2(x, y) \wedge R_2(x, z) &\rightarrow y = z, \\ R_2(y, x) \wedge R_2(z, x) &\rightarrow y = z. \end{aligned}$$

Let $m = 1$, $d = 2$ and $k \geq 1$. We will show that for these values there is no $d', k' \geq 0$ such that for every instance I of S and for every $a, b \in \text{dom}(I) \cap \text{dom}(\mathcal{F}_{\text{univ}}(I))$, if $N_{d'}^I(a) \equiv_{k'} N_{d'}^I(b)$, then $N_d^{\mathcal{F}_{\text{univ}}(I)}(a) \equiv_k N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$. On the contrary, assume that such d', k' exist. Define a database instance I with domain $\{a, a_1, \dots, a_{d'}, b, b_1, \dots, b_{d'}, c\}$ as follows. $I(V) = \{c\}$ and $I(E)$ contains the following tuples:

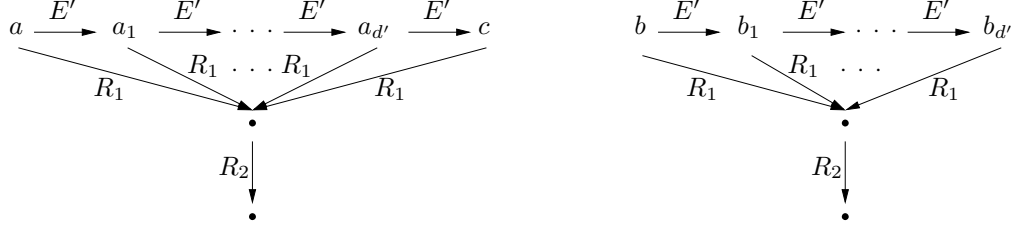
$$a \xrightarrow{E} a_1 \xrightarrow{E} \dots \xrightarrow{E} a_{d'} \xrightarrow{E} c \qquad b \xrightarrow{E} b_1 \xrightarrow{E} \dots \xrightarrow{E} b_{d'}$$

As shown in the figure, $I(E)$ is a union of two paths, one containing $d' + 2$ elements with first element a and last element c and another path containing $d' + 1$ elements with first element b . Observe that $N_{d'}^I(a) \equiv_{k'} N_{d'}^I(b)$, since $N_{d'}^I(a) \cong N_{d'}^I(b)$.

The canonical universal solution $\mathcal{F}_{\text{univ}}(I)$ of I can be constructed by first applying the set of source-to-target dependencies Σ_{st} :



and then applying the set of key dependencies Σ_t :



In the figures shown above, the symbol \bullet is used to represent null values. Observe that $V'(\mathcal{F}_{\text{univ}}(I)) = \{c\}$ since the only source-to-target dependency mentioning predicate V' is $V(x) \rightarrow V'(x)$. Thus, $N_d^{\mathcal{F}_{\text{univ}}(I)}(a) \not\equiv_k N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$, since the distance between a and c is at most 2 and there is no a point c' in $N_d^{\mathcal{F}_{\text{univ}}(I)}(b)$ such that $V'(c')$ holds. This leads to a contradiction.

We note that the previous proof also shows that the transformation $\mathcal{F}_{\text{core}}$ of LAV+egd settings is not necessarily locally consistent (under FO-equivalence) since for the instance I shown above, $\mathcal{F}_{\text{univ}}(I) = \mathcal{F}_{\text{core}}(I)$.

Proof of Proposition 4.3

Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting, where $\mathbf{T} = \langle R_1^{a_1}, \dots, R_m^{a_m} \rangle$ and a_i is the arity of predicate R_i ($1 \leq i \leq m$), and Q be a Boolean query over \mathbf{T} . For $n \geq 0$, define $\mu_n(\mathbf{T})$ and $\mu_n(\mathbf{T}, Q)$ as follows:

$$\begin{aligned} \mu_n(\mathbf{T}) &= |\{J \mid J \text{ is an instance of } \mathbf{T} \text{ and } \text{dom}(J) = \{1, \dots, n\}\}|, \\ \mu_n(\mathbf{T}, Q) &= |\{J \mid J \text{ is an instance of } \mathbf{T}, \text{dom}(J) = \{1, \dots, n\} \text{ and } Q(J) = \text{true}\}|. \end{aligned}$$

Then we say that the asymptotic probability of Q is 0 if

$$\lim_{n \rightarrow \infty} \frac{\mu_n(\mathbf{T}, Q)}{\mu_n(\mathbf{T})} = 0. \quad (3)$$

We will prove that if the asymptotic probability of Q is 0, then for every instance I of \mathbf{S} , $\text{certain}(Q, I) = \text{false}$. From this it immediately follows that Q is rewritable over the canonical universal solution and over the core.

On the contrary, assume that the property is false, that is, there exists an instance I of \mathbf{S} such that $\text{certain}(Q, I) = \text{true}$, and let J be an arbitrary solution for I , with $\text{dom}(J) = \{1, \dots, k\}$. We show that this assumption leads to a contradiction by considering the following fact: $Q(J') = \text{true}$, for every instance J' of \mathbf{T} containing J , since J is a solution for I , $\text{certain}(Q, I) = \text{true}$ and every instance J' of \mathbf{T} containing J is also a solution for I . To establish a contradiction we consider two cases.

1. First assume that all the predicates in T are unary ($a_i = 1$ for every $i \in [1, m]$). In this case, for every $n > k$:

$$\begin{aligned} \mu_n(\mathbf{T}) &= (2^m - 1)^n, \\ \mu_n(\mathbf{T}, Q) &\geq (2^m - 1)^{n-k}. \end{aligned}$$

Thus, if the $\lim_{n \rightarrow \infty} \mu_n(\mathbf{T}, Q) / \mu_n(\mathbf{T})$ exists, then this limit is at least:

$$\lim_{n \rightarrow \infty} \frac{(2^m - 1)^{n-k}}{(2^m - 1)^n} = \frac{1}{(2^m - 1)^k} > 0.$$

Hence, (3) does not hold, which is a contradiction.

2. Assume that there is at least one non-unary predicate in \mathbf{T} , say $a_1 > 1$. For every $n \geq 0$:

$$\sum_{i=0}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}) = \prod_{i=1}^m 2^{n^{a_i}}$$

and, therefore,

$$\mu_n(\mathbf{T}) = \prod_{i=1}^m 2^{n^{a_i}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}). \quad (4)$$

Furthermore, given that $Q(J') = \text{true}$ for every instance J' of \mathbf{T} containing J and $\text{dom}(J) = \{1, \dots, k\}$, we conclude that

$$\mu_n(\mathbf{T}, Q) \geq \prod_{i=1}^m 2^{n^{a_i} - k^{a_i}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}). \quad (5)$$

Therefore, by (4) and (5)

$$\lim_{n \rightarrow \infty} \frac{\mu_n(\mathbf{T}, Q)}{\mu_n(\mathbf{T})} \geq \frac{\prod_{i=1}^m 2^{n^{a_i} - k^{a_i}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T})}{\prod_{i=1}^m 2^{n^{a_i}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T})} = \frac{1 / \prod_{i=1}^m 2^{k^{a_i}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}) / \prod_{i=1}^m 2^{n^{a_i}}}{1 - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}) / \prod_{i=1}^m 2^{n^{a_i}}}.$$

Thus, if we prove that

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T})}{\prod_{i=1}^m 2^{n^{a_i}}} = 0, \quad (6)$$

then we conclude that the minimum value of $\lim_{n \rightarrow \infty} \mu_n(\mathbf{T}, Q) / \mu_n(\mathbf{T})$ could be $1 / \prod_{i=1}^m 2^{k^{a_i}}$, if this limit exists, and, therefore, (3) does not hold, which is a contradiction. Next we prove that (6) is true.

For every $n \geq 0$, let $\mu_n(R_1)$ be the number of instances of R_1 with $\{1, \dots, n\}$ as domain. As in the previous case,

$$\mu_n(R_1) = 2^{n^{a_1}} - \sum_{i=1}^n \binom{n}{i} \mu_{n-i}(R_1).$$

Thus, given that for every $i \in [1, n]$:

$$\mu_{n-i}(R_1) \leq 2^{(n-i)^{a_1}} \leq 2^{(n-1)^{a_1}},$$

we conclude that

$$\begin{aligned} \mu_n(R_1) &\geq 2^{n^{a_1}} - \sum_{i=1}^n \binom{n}{i} 2^{(n-1)^{a_1}} \\ &= 2^{n^{a_1}} - 2^{(n-1)^{a_1}} \cdot (2^n - 1) \\ &= 2^{n^{a_1}} + 2^{(n-1)^{a_1}} - 2^{(n-1)^{a_1} + n}. \end{aligned}$$

Hence, for every $n \geq 0$

$$\mu_n(\mathbf{T}) \geq \mu_n(R_1) \cdot \prod_{i=2}^m 2^{n^{a_i}} \geq (2^{n^{a_1}} + 2^{(n-1)^{a_1}} - 2^{(n-1)^{a_1} + n}) \cdot \prod_{i=2}^m 2^{n^{a_i}}.$$

Thus, by (4) we conclude that

$$\sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T}) \leq \prod_{i=1}^m 2^{n^{a_i}} - (2^{n^{a_1}} + 2^{(n-1)^{a_1}} - 2^{(n-1)^{a_1} + n}) \cdot \prod_{i=2}^m 2^{n^{a_i}}$$

and, hence,

$$\frac{\sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T})}{\prod_{i=1}^m 2^{n^{a_i}}} \leq 1 - \frac{2^{n^{a_1}} + 2^{(n-1)^{a_1}} - 2^{(n-1)^{a_1}+n}}{2^{n^{a_1}}}.$$

Given that $(n-1)^{a_1} = n^{a_1} - a_1 n^{a_1-1} + p(n)$, where $p(n)$ is a polynomial of degree $a_1 - 2$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \binom{n}{i} \mu_{n-i}(\mathbf{T})}{\prod_{i=1}^m 2^{n^{a_i}}} &\leq 1 - \lim_{n \rightarrow \infty} \frac{2^{n^{a_1}} + 2^{(n-1)^{a_1}} - 2^{(n-1)^{a_1}+n}}{2^{n^{a_1}}} \\ &= 1 - \lim_{n \rightarrow \infty} \left(1 + \frac{1}{2^{a_1 n^{a_1-1} - p(n)}} - \frac{1}{2^{a_1 n^{a_1-1} - n - p(n)}} \right) \\ &= 0. \end{aligned}$$

This concludes the proof of the proposition.

Proof of Theorem 4.5

We prove the theorem for the case of locally consistent transformations. The proof for transformations that are locally consistent under FO-equivalence is similar.

Let Q' be a first-order rewriting of Q over \mathcal{F} , that is, an m -ary FO query over \mathbf{T} such that for every instance I of \mathbf{S} , $\text{certain}(Q, I) = Q'(\mathcal{F}(I))$. Assume that \mathcal{F} is locally consistent. By Gaifman's theorem [11], there exists a constant d' such that for every instance J of \mathbf{T} and m -tuples \bar{a}, \bar{b} in J , if $N_{d'}^J(\bar{a}) \cong N_{d'}^J(\bar{b})$, then $\bar{a} \in Q'(J)$ if and only if $\bar{b} \in Q'(J)$. Given that \mathcal{F} is locally consistent, there exists $d \geq 0$ such that for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} in I , if $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$, then (1) $\bar{a} \in \text{dom}(\mathcal{F}(I))$ if and only if $\bar{b} \in \text{dom}(\mathcal{F}(I))$ and (2) $N_{d'}^{\mathcal{F}(I)}(\bar{a}) \cong N_{d'}^{\mathcal{F}(I)}(\bar{b})$. From this we conclude that Q is locally source-dependent since for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} in I ,

$$\begin{aligned} N_d^I(\bar{a}) \cong N_d^I(\bar{b}) &\Rightarrow N_{d'}^{\mathcal{F}(I)}(\bar{a}) \cong N_{d'}^{\mathcal{F}(I)}(\bar{b}) \\ &\Rightarrow \bar{a} \in Q'(\mathcal{F}(I)) \text{ iff } \bar{b} \in Q'(\mathcal{F}(I)) \\ &\Rightarrow \bar{a} \in \text{certain}(Q, I) \text{ iff } \bar{b} \in \text{certain}(Q, I). \end{aligned}$$

Proof of Theorem 4.8

Theorem 4.8 is a consequence of the following proposition.

Proposition A.6 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting and $\varphi(\bar{x})$ an FO formula over \mathbf{T} . Then there exists an FO formula $\varphi'(\bar{x})$ such that for every instance I of \mathbf{S} and every $\bar{a} \in \text{dom}(\mathcal{F}_{\text{univ}}(I))^m$, $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{a})$ iff there exists a core J' of $\mathcal{F}_{\text{univ}}(I)$ such that $\bar{a} \in \text{dom}(J')^m$ and $J' \models \varphi(\bar{a})$. Furthermore, if $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ is considered to be fixed, then φ' can be computed in time $O(\|\varphi\|^3)$.*

To prove this proposition, we assume that we are given a fixed data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and we let q be the maximum number of conjuncts in the right-hand side of a source-to-target dependency in Σ_{st} . Furthermore, we let $m_{\mathbf{T}}$ be the maximum arity of a predicate in \mathbf{T} .

For every canonical universal solution J of an instance of \mathbf{S} and for every null value $c \in \text{dom}(J)$, define $B_{\text{univ}}(c) \subseteq \text{dom}(J)$ as follows. For every $c' \in \text{dom}(J)$, $c' \in B_{\text{univ}}(c)$ if there exists tuples $P_1(\bar{c}_1), \dots, P_n(\bar{c}_n)$ in J , $n \geq 1$, such that $c \in \bar{c}_1$, $\bar{c}_i \cap \bar{c}_{i+1} \cap \text{Var} \neq \emptyset$ ($i \in [1, n-1]$) and $c' \in \bar{c}_n$. We note that by definition of the canonical universal solution, $c' \in B_{\text{univ}}(c)$ iff there exists a path $c_1 \cdots c_n$, $1 \leq n \leq q$, in the Gaifman graph of J such that $c_1 = c$, c_i is a null value ($i \in [1, n-1]$) and $c_n = c'$. Moreover, by slightly abusing the notation, we say that $X \subseteq B_{\text{univ}}(c)$ is a core of $B_{\text{univ}}(c)$ if $J|_X$ is a core of $J|_{B_{\text{univ}}(c)}$, where for every $Y \subseteq \text{dom}(J)$, $J|_Y$ is the substructure of J induced by the elements in Y . The following follows directly from the definitions.

Claim A.7 *Let J be a canonical universal solution of an instance of \mathbf{S} and J^* be a core of J . Then*

a) for every $c \in \text{Var}(J^*)$, $B_{\text{univ}}(c) \cap \text{dom}(J^*)$ is a core of $B_{\text{univ}}(c)$.

b) every substructure of J^* is a core of itself.

In this section we will extensively use the following predicates. Let \bar{x} and \bar{y} be two q -tuples of variables. Then $\text{Hom}(\bar{x}, \bar{y})$ is a formula such that for any instance J of \mathbf{T} and tuples \bar{a} and \bar{b} in J , $J \models \text{Hom}(\bar{a}, \bar{b})$ iff there is a homomorphism from $J|_{\bar{a}}$ to $J|_{\bar{b}}$ (this is equivalent to say that (\bar{a}, \bar{b}) defines a partial homomorphism $J \rightarrow J$ that is the identity on the constants), and $\text{Iso}(\bar{x}, \bar{y})$ is such that $J \models \text{Iso}(\bar{a}, \bar{b})$ iff there is an isomorphism from $J|_{\bar{a}}$ to $J|_{\bar{b}}$ that is the identity on the constants (this is equivalent to say that (\bar{a}, \bar{b}) defines a partial isomorphism $J \rightarrow J$ that is the identity on the constants). It is not hard to see that each one of the previous predicates is definable by a quantifier-free FO formula.

Lemma A.8 *There exists an FO formula $\theta_{\text{core}}(u, \bar{v})$, where $|\bar{v}| = q \cdot m_{\mathbf{T}}$, such that for every canonical universal solution J of some instance of \mathbf{S} , and for every tuple $c\bar{c} \subseteq \text{dom}(J)$, $J \models \theta_{\text{core}}(c, \bar{c})$ if and only if $c \in \text{Var}(J)$ and \bar{c} is a core of $B_{\text{univ}}(c)$ containing c .*

Proof: By definition of B_{univ} , for every canonical universal solution J and every $c \in \text{Var}(J)$, $B_{\text{univ}}(c)$ contains at most $q \cdot m_{\mathbf{T}}$ elements and is a subset of the set of points at distance at most q from c in the Gaifman graph of J . Thus, by definition of B_{univ} and given that Σ_{st} is assumed to be fixed, B_{univ} is FO definable, that is, there is an FO formula $\alpha(x, \bar{y})$, where $|\bar{y}| = q \cdot m_{\mathbf{T}}$, such that for every canonical universal solution J of some instance of \mathbf{S} , for every $c \in \text{Var}(J)$ and for every tuple $\bar{c} \subseteq \text{dom}(J)$, $J \models \alpha(c, \bar{c})$ if and only if $\bar{c} = B_{\text{univ}}(c)$ (if $B_{\text{univ}}(c)$ contains less than q elements, then \bar{c} has some repeated elements). Then formula $\theta_{\text{core}}(u, \bar{v})$, where $|\bar{v}| = q \cdot m_{\mathbf{T}}$, is defined as follows:

$$\neg C(u) \wedge u \in \bar{v} \wedge \exists \bar{y} (\alpha(u, \bar{y}) \wedge \bar{v} \subseteq \bar{y} \wedge \text{Hom}(\bar{y}, \bar{v}) \wedge \bigwedge_{\{\bar{z} \mid \bar{z} \subseteq \bar{v}\}} \neg \text{Hom}(\bar{v}, \bar{z})).$$

□

Lemma A.9 *Let J be a canonical universal solution of an instance of \mathbf{S} , J^* a core of J , $c \in \text{Var}(J^*)$ and $B_{\text{core}}^*(c) = \text{dom}(J^*) \cap B_{\text{univ}}(c)$. For every substructure J' of J such that $|J'| \leq |B_{\text{core}}^*(c)|$, if there is a homomorphism from $B_{\text{core}}^*(c)$ to J' , then there exists a homomorphism from J' to $B_{\text{core}}^*(c)$.*

Proof: Assume that there is a homomorphism h from $B_{\text{core}}^*(c)$ to a substructure J' of J such that $|J'| \leq |B_{\text{core}}^*(c)|$, and assume that there is no homomorphism from J' to $B_{\text{core}}^*(c)$. Consider h' to be a homomorphism from J to J^* that is the identity in J^* (the existence of this homomorphism is proved in [15]), and let $f : J^* \rightarrow J^*$ be a homomorphism defined as follows:

$$f(x) = \begin{cases} h'(h(x)) & x \in B_{\text{core}}^*(c) \\ x & \text{otherwise} \end{cases}$$

By the hypothesis h' is not a homomorphism from J' to $B_{\text{core}}^*(c)$. Thus, given that $|J'| \leq |B_{\text{core}}^*(c)|$, there is at least one null value c' in $B_{\text{core}}^*(c)$ which is not the image of any element in J' under the homomorphism f . Hence, f maps J^* into a proper subset of itself, which contradicts the fact that J^* is a core. □

Lemma A.10 *Let J be a canonical universal solution of an instance of \mathbf{S} , $c \in \text{Var}(J)$ and $X \subseteq \text{dom}(J)$ a core of $B_{\text{univ}}(c)$ that contains c , and assume that for every substructure J' of J , if $|J'| \leq |X|$ and there is a homomorphism from $J|_X$ to J' then there exists a homomorphism from J' to $J|_X$. Then c belongs to some core of J .*

Proof: Consider an arbitrary core J^* of J and let h be the homomorphism from J to J^* that is the identity on every element of J^* . We claim that there is an isomorphism from $J|_{h(X)}$ to $J|_X$ that is the identity in the constants. This shows that there is an isomorphism from J^* to $(J^* \setminus J|_{h(X)}) \cup J|_X$ and, hence, that c belongs to a core.

We know that h is a homomorphism from $J|_X$ to $J|_{h(X)}$ and $|h(X)| \leq |X|$. By hypothesis there is a homomorphism h' from $J|_{h(X)}$ to $J|_X$. Hence h must be one-to-one; otherwise the composition of $(h' \circ h)$ will map $J|_X$ to a proper substructure of itself, contradicting the fact that $J|_X$ is a core. By the same reason (see Claim A.7.b), h' is one-to-one, and we conclude that h is one-to-one and onto. Since h' is the identity on the constants, the image of a null under h must be a null; otherwise the composition $(h' \circ h)$ would map $J|_X$ to a proper substructure of itself, contradicting the fact that $J|_X$ is a core. This proves that h is an isomorphism that is the identity on the constants. \square

Recall that C is a unary predicate whose interpretation over an instance I is $\text{Const}(I)$. We define a formula $\text{Core}_{\text{Var}}(x)$ as follows:

$$\neg C(x) \wedge \forall \bar{y} \forall \bar{z} (\theta_{\text{core}}(x, \bar{y}) \wedge \text{Hom}(\bar{y}, \bar{z}) \wedge \text{Onto}(\bar{y}, \bar{z}) \rightarrow \text{Hom}(\bar{z}, \bar{y})),$$

where $|\bar{z}| = q \cdot m_{\mathbf{T}}$ and $\text{Onto}(\bar{y}, \bar{z})$ tests whether the cardinality of the elements instantiating \bar{y} is bigger than the cardinality of the elements instantiating \bar{z} . Since the size of \bar{y} is fixed from $\theta_{\text{core}}(x, \bar{y})$ and is equal to the size of \bar{z} , $\text{Onto}(\bar{y}, \bar{z})$ is FO expressible. From Lemmas A.9 and A.10 we obtain that for a canonical universal solution J and $a \in \text{dom}(J)$, $J \models \text{Core}_{\text{Var}}(a)$ iff $a \in \text{Var}(J)$ and there is a core of J containing a .

Given a canonical universal solution J of some instance of \mathbf{S} and $n > 0$, the predicate $\text{Core}_{\text{Var}}(x_1, \dots, x_n)$ is defined as follows. For every $a_1, \dots, a_n \in \text{dom}(J)$, $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_n)$ if and only if there exists a core J^* of J such that $\{a_1, \dots, a_n\} \subseteq \text{Var}(J^*)$. Next we prove that $\text{Core}_{\text{Var}}(x_1, \dots, x_n)$ is first-order definable, but first we need to define one extra formula.

Let $\beta(x, y)$ be an FO formula defined as follows. For every canonical universal solution J of an instance of \mathbf{S} , $J \models \beta(a, b)$ iff for every \bar{a} and \bar{b} such that $J \models \theta_{\text{core}}(a, \bar{a}) \wedge \theta_{\text{core}}(b, \bar{b})$ the following holds:

- If \bar{a} and \bar{b} do not contain exactly the same elements and for every \bar{a}' and \bar{b}' such that $J \models \theta_{\text{core}}(a, \bar{a}') \wedge \theta_{\text{core}}(b, \bar{b}')$ it is not the case that $J|_{\bar{a}'} \cong J|_{\bar{b}'}$, then there is no homomorphism from $J|_{\bar{a}}$ to $J|_{\bar{b}}$.

It is not hard to see that $\beta(x, y)$ can be expressed in FO as follows:

$$\forall \bar{x} \forall \bar{y} (\theta_{\text{core}}(x, \bar{x}) \wedge \theta_{\text{core}}(y, \bar{y}) \rightarrow (\neg \text{Eq}(\bar{x}, \bar{y}) \wedge \neg \exists \bar{u} \exists \bar{v} (\theta_{\text{core}}(x, \bar{u}) \wedge \theta_{\text{core}}(y, \bar{v}) \wedge \text{Iso}(\bar{u}, \bar{v})) \rightarrow \neg \text{Hom}(\bar{x}, \bar{y}))),$$

where predicate $\text{Eq}(\bar{x}, \bar{y})$ is an FO formula checking whether \bar{x} and \bar{y} are the same set of elements. This formula exists since $|\bar{x}| = |\bar{y}| = q \cdot m_{\mathbf{T}}$.

Next lemma shows that $\text{Core}_{\text{Var}}(x_1, \dots, x_n)$ can be defined in FO, for every $n > 0$.

Lemma A.11 *The formula $\text{Core}_{\text{Var}}(x_1, \dots, x_n)$ is definable in FO by*

$$\bigwedge_{1 \leq i \leq n} \text{Core}_{\text{Var}}(x_i) \wedge \bigwedge_{1 \leq j \neq k \leq n} \beta(x_j, x_k).$$

Proof: Assume first that for elements $a_1, \dots, a_n \in \text{Var}(J)$, $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_n)$, where J is a canonical universal solution of an instance of \mathbf{S} . Hence, for some core J^* of J it is the case that $a_1, \dots, a_n \in \text{Var}(J^*)$. This implies that $J \models \text{Core}_{\text{Var}}(a_i)$, for every $i \in [1, n]$.

Towards a contradiction assume that $J \models \neg \beta(a_j, a_k)$, for some $j, k \in [1, n]$, $j \neq k$. Then there exists \bar{c} and \bar{d} such that $J \models \theta_{\text{core}}(a_j, \bar{c}) \wedge \theta_{\text{core}}(a_k, \bar{d}) \wedge \neg \text{Eq}(\bar{c}, \bar{d})$, and for every \bar{c}' and \bar{d}' , $J \models \theta_{\text{core}}(a_j, \bar{c}') \wedge \theta_{\text{core}}(a_k, \bar{d}') \rightarrow \neg \text{Iso}(\bar{c}', \bar{d}')$. In addition, there is a homomorphism $h : J|_{\bar{c}} \rightarrow J|_{\bar{d}}$.

Define $B_{\text{core}}^*(a_j)$ and $B_{\text{core}}^*(a_k)$ as $B_{\text{core}}^*(a_j) = B_{\text{univ}}(a_j) \cap \text{dom}(J^*)$ and $B_{\text{core}}^*(a_k) = B_{\text{univ}}(a_k) \cap \text{dom}(J^*)$ respectively. Then $B_{\text{core}}^*(a_j) \cap B_{\text{core}}^*(a_k) \subseteq \text{Const}(J)$. Otherwise suppose there is a null n in both $B_{\text{core}}^*(a_j)$ and $B_{\text{core}}^*(a_k)$. Then there is a path $a_j \cdots n \cdots a_k$ in the Gaifman graph of J^* such that all the elements in this path are nulls. Then $B_{\text{core}}^*(a_j) = B_{\text{core}}^*(a_k)$, which contradicts the fact that there are no two cores of $B_{\text{univ}}(a_j)$ and $B_{\text{univ}}(a_k)$ containing a_j and a_k , respectively, such that these cores are isomorphic.

We know there exist isomorphisms $f_j : B_{\text{core}}^*(a_j) \rightarrow J|_{\bar{c}}$ and $f_k : J|_{\bar{d}} \rightarrow B_{\text{core}}^*(a_k)$. Define $h' : J^* \rightarrow J^*$ in the following way:

$$h'(x) = \begin{cases} f_k(h(f_j(x))) & x \in B_{\text{core}}^*(a_j) \\ x & \text{otherwise} \end{cases}$$

It is not hard to see that h' is a homomorphism and that h' maps J^* into a proper subset of itself since $B_{\text{core}}^*(a_j) \cap B_{\text{core}}^*(a_k) \subseteq \text{Const}(J)$. This is a contradiction because J^* is a core.

Assume now that for elements $a_1, \dots, a_n \in \text{Var}(J)$,

$$J \models \bigwedge_{1 \leq i \leq n} \text{Core}_{\text{Var}}(a_i) \wedge \bigwedge_{1 \leq j \neq k \leq n} \beta(a_j, a_k),$$

where J is a canonical universal solution of an instance of **S**. We prove that $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_n)$ by induction on n . The case $n = 1$ is trivial.

Assume that for elements $a_1, \dots, a_{n+1} \in \text{Var}(J)$,

$$J \models \bigwedge_{1 \leq i \leq n+1} \text{Core}_{\text{Var}}(a_i) \wedge \bigwedge_{1 \leq j \neq k \leq n+1} \beta(a_j, a_k).$$

Hence,

$$J \models \bigwedge_{1 \leq i \leq n} \text{Core}_{\text{Var}}(a_i) \wedge \bigwedge_{1 \leq j \neq k \leq n} \beta(a_j, a_k) \wedge \text{Core}_{\text{Var}}(a_{n+1}) \wedge \bigwedge_{1 \leq i \leq n} \beta(a_i, a_{n+1}) \wedge \beta(a_{n+1}, a_i).$$

By induction hypothesis, $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_n)$. This means that there exists a core J_1^* of J that contains elements a_1, \dots, a_n . In addition, there is a core J_2^* of J such that $a_{n+1} \in \text{dom}(J_2^*)$ since $J \models \text{Core}_{\text{Var}}(a_{n+1})$.

Let $f : J_1^* \rightarrow J_2^*$ be an isomorphism that is the identity on the constants. We know that for every $c \in \text{dom}(J_1^*)$ and $B_{\text{core}}^*(c) = B_{\text{univ}}(c) \cap \text{dom}(J_1^*)$, $f(B_{\text{core}}^*(c)) = B_{\text{core}}^{**}(f(c))$, where $B_{\text{core}}^{**}(f(c)) = B_{\text{univ}}(f(c)) \cap \text{dom}(J_2^*)$.

In order to prove that $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_{n+1})$ we consider two cases:

- It is the case that for every $i \in [1, n]$, $B_{\text{core}}^{**}(a_{n+1}) \neq f(B_{\text{core}}^*(a_i))$. Let $J^* = (J_1^* \setminus f^{-1}(B_{\text{core}}^{**}(a_{n+1}))) \cup B_{\text{core}}^{**}(a_{n+1})$. It is not hard to see that J^* is a core of J and that $a_1, \dots, a_{n+1} \in \text{dom}(J^*)$.
- For some $i \in [1, n]$, $B_{\text{core}}^{**}(a_{n+1}) = f(B_{\text{core}}^*(a_i))$. Then there is a homomorphism from $J|_{B_{\text{core}}^{**}(a_{n+1})}$ to $J|_{B_{\text{core}}^*(a_i)}$. Since $J \models \beta(a_{n+1}, a_i)$ and $J|_{B_{\text{core}}^{**}(a_{n+1})} \cong J|_{B_{\text{core}}^*(a_i)}$, we deduce that $B_{\text{core}}^{**}(a_{n+1}) = B_{\text{core}}^*(a_i)$ and, therefore, $a_{n+1} \in B_{\text{core}}^*(a_i)$. Thus J_1^* contains elements a_1, \dots, a_{n+1} , implying $J \models \text{Core}_{\text{Var}}(a_1, \dots, a_{n+1})$.

□

From $\text{Core}_{\text{Var}}(x_1, \dots, x_n)$ we define a formula $\text{Core}(x_1, \dots, x_n)$ as follows:

$$\begin{aligned} & \left(\bigwedge_{i=1}^n C(x_i) \right) \vee \left(\bigwedge_{i=1}^n \neg C(x_i) \wedge \text{Core}_{\text{Var}}(x_1, \dots, x_n) \right) \vee \\ & \exists y_1 \dots \exists y_n \exists z_1 \dots \exists z_n \left(\bigwedge_{i=1}^n \bigvee_{j=1}^n y_i = x_j \wedge \bigwedge_{i=1}^n C(y_i) \wedge \bigwedge_{i=1}^n \bigvee_{j=1}^n z_i = x_j \wedge \bigwedge_{i=1}^n \neg C(z_i) \wedge \right. \\ & \quad \left. \bigwedge_{i=1}^n \bigwedge_{j=1}^n y_i \neq z_j \wedge \bigwedge_{i=1}^n \bigvee_{j=1}^n (x_i = y_j \vee x_i = z_j) \wedge \text{Core}_{\text{Var}}(z_1, \dots, z_n) \right). \end{aligned}$$

Intuitively, for a canonical universal solution J and a tuple \bar{a} , $J \models \text{Core}(\bar{a})$ iff there is a core J^* of J that contains all the elements in \bar{a} .

Proof of Proposition A.6: Take an arbitrary FO formula $\varphi(\bar{x})$. The algorithm transforms first $\varphi(\bar{x})$ into an equivalent FO formula in prenex normal form. Then it transforms the latter into an equivalent formula $\varphi^*(\bar{x})$ of the form

$$(\neg)\exists y_1 (\neg)\exists y_2 \dots (\neg)\exists y_p \psi(\bar{x}, y_1, \dots, y_p),$$

where ψ is a quantifier-free formula and (\neg) means that the negation may or may not be present in front of the existential quantifier. We claim that $\varphi'(\bar{x})$ defined as

$$\text{Core}(\bar{x}) \wedge (\neg)\exists y_1 (\text{Core}(\bar{x}, y_1) \wedge (\neg)\exists y_2 (\text{Core}(\bar{x}, y_1, y_2) \wedge (\dots \wedge (\neg)\exists y_p (\text{Core}(\bar{x}, y_1, \dots, y_p) \wedge \psi(\bar{x}, y_1, \dots, y_p)) \dots)))$$

satisfies that for every instance I of \mathbf{S} and every $\bar{a} \in \text{dom}(\mathcal{F}_{\text{univ}}(I))^m$, $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{a})$ iff there exists a core J' of $\mathcal{F}_{\text{univ}}(I)$ such that $\bar{a} \in \text{dom}(J')^m$ and $J' \models \varphi(\bar{a})$. We prove this by induction on p .

Let $J = \mathcal{F}_{\text{univ}}(I)$ and $\bar{a} \in \text{dom}(J)^m$. For $p = 0$ we have that $\varphi'(\bar{x}) = \text{Core}(\bar{x}) \wedge \varphi(\bar{x})$. Assume first $J \models \text{Core}(\bar{a}) \wedge \varphi(\bar{a})$. Since $J \models \text{Core}(\bar{a})$, there is a core J' of J that contains all the elements in \bar{a} . Moreover, since φ is a boolean combination of atomic formulas, it must be the case that $J' \models \varphi(\bar{a})$. On the other hand, assume there is a core J' of J such that $J' \models \varphi(\bar{a})$ and all the elements of \bar{a} are in J' . Since J' is a substructure of J and φ is a boolean combination of atomic formulas, $J \models \text{Core}(\bar{a}) \wedge \varphi(\bar{a})$.

Going from p to $p + 1$ we have two cases to analyze depending on the existential quantifier binding y_1 . The first case is that it occurs positively. The analysis then goes as follows. Assume that

$$J \models \text{Core}(\bar{a}) \wedge \exists y_1 (\text{Core}(\bar{a}, y_1) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, y_1, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, y_1, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, y_1, y_2, \dots, y_{p+1})) \dots))).$$

Then for some $c \in \text{dom}(J)$,

$$J \models \text{Core}(\bar{a}, c) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, c, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, c, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, c, y_2, \dots, y_{p+1})) \dots)).$$

By the induction hypothesis there is a core J' of J such that it contains all the elements in $\bar{a}c$ and

$$J' \models (\neg)\exists y_2 \dots (\neg)\exists y_{p+1} \psi(\bar{a}, c, y_2, \dots, y_{p+1}),$$

implying that $J' \models \exists y_1 (\neg)\exists y_2 \dots (\neg)\exists y_{p+1} \psi(\bar{a}, y_1, y_2, \dots, y_{p+1})$.

Assume on the other hand that there is a core J' of J such that $J' \models \exists y_1 (\neg)\exists y_2 \dots (\neg)\exists y_{p+1} \psi(\bar{a}, y_1, y_2, \dots, y_{p+1})$ and J' contains all the elements in \bar{a} . Then for some $c \in \text{dom}(J')$, $J' \models (\neg)\exists y_2 \dots (\neg)\exists y_{p+1} \psi(\bar{a}, c, y_2, \dots, y_{p+1})$. By the induction hypothesis

$$J \models \text{Core}(\bar{a}, c) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, c, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, c, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, c, y_2, \dots, y_{p+1})) \dots)).$$

The latter shows that for some $c \in \text{dom}(J)$,

$$J \models \text{Core}(\bar{a}) \wedge \exists y_1 (\text{Core}(\bar{a}, y_1) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, y_1, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, y_1, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, y_1, y_2, \dots, y_{p+1})) \dots))).$$

The second case is that the quantifier binding y_1 occurs negatively. Assume that

$$J \models \text{Core}(\bar{a}) \wedge \neg \exists y_1 (\text{Core}(\bar{a}, y_1) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, y_1, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, y_1, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, y_1, y_2, \dots, y_{p+1})) \dots))).$$

Then for every $c \in \text{dom}(J)$, $J \models \text{Core}(\bar{a})$ and

$$J \not\models \text{Core}(\bar{a}, c) \wedge (\neg)\exists y_2 (\text{Core}(\bar{a}, c, y_2) \wedge (\dots \wedge (\neg)\exists y_{p+1} (\text{Core}(\bar{a}, c, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, c, y_2, \dots, y_{p+1})) \dots)).$$

Take J' to be a core of J that contains all the elements in \bar{a} . By the induction hypothesis, for every $c \in \text{dom}(J')$,

$$J' \not\models (\neg)\exists y_2 \dots (\neg)\exists y_{p+1} \psi(\bar{a}, c, y_2, \dots, y_{p+1}),$$

implying that $J' \models \neg \exists y_1 (\neg \exists y_2 \dots (\neg \exists y_{p+1} \psi(\bar{a}, y_1, y_2, \dots, y_{p+1}))$.

Assume on the other hand that

$$J \not\models \text{Core}(\bar{a}) \wedge \neg \exists y_1 (\text{Core}(\bar{a}, y_1) \wedge (\neg \exists y_2 (\text{Core}(\bar{a}, y_1, y_2) \wedge (\dots \wedge (\neg \exists y_{p+1} (\text{Core}(\bar{a}, y_1, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, y_1, y_2, \dots, y_{p+1}))) \dots))).$$

Then, either $J \not\models \text{Core}(\bar{a})$, which trivially satisfies that there is no core J' of J such that $J' \models \neg \exists y_1 (\neg \exists y_2 \dots (\neg \exists y_{p+1} \psi(\bar{a}, y_1, y_2, \dots, y_{p+1}))$, or $J \models \text{Core}(\bar{a})$ and

$$J \not\models \neg \exists y_1 (\text{Core}(\bar{a}, y_1) \wedge (\neg \exists y_2 (\text{Core}(\bar{a}, y_1, y_2) \wedge (\dots \wedge (\neg \exists y_{p+1} (\text{Core}(\bar{a}, y_1, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, y_1, y_2, \dots, y_{p+1}))) \dots))).$$

The latter implies there is $c \in \text{dom}(J)$ such that

$$J \models \text{Core}(\bar{a}, c) \wedge (\neg \exists y_2 (\text{Core}(\bar{a}, c, y_2) \wedge (\dots \wedge (\neg \exists y_{p+1} (\text{Core}(\bar{a}, c, y_2, \dots, y_{p+1}) \wedge \psi(\bar{a}, c, y_2, \dots, y_{p+1}))) \dots))).$$

By the induction hypothesis there is a core J' of J such that J' contains all the elements in $\bar{a}c$ and

$$J' \models (\neg \exists y_2 \dots (\neg \exists y_{p+1} \psi(\bar{a}, c, y_2, \dots, y_{p+1})),$$

implying that $J' \models \exists y_1 (\neg \exists y_2 \dots (\neg \exists y_{p+1} \psi(\bar{a}, y_1, y_2, \dots, y_{p+1}))$. This concludes the first part of the proof.

To finish we need to show that φ' can be computed in time $O(\|\varphi\|^3)$. First we note that the formula φ^* is of size $O(\|\varphi\|)$ and can be computed in time $O(\|\varphi\|^2)$. Thus, we only need to show that φ' can be computed from φ^* in time $O(\|\varphi^*\|^3)$. Given that

$$\|\varphi'\| = \|\varphi^*\| + \|\text{Core}(\bar{x})\| + \sum_{i=1}^p \|\text{Core}(y_1, \dots, y_i)\|,$$

and for every $i \in [1, p]$, $\|\text{Core}(y_1, \dots, y_i)\|$ is $O(i^2 + \|\text{Core}_{\text{var}}(y_1, \dots, y_i)\|)$ and $\|\text{Core}_{\text{var}}(y_1, \dots, y_i)\|$ is $O(i \cdot \|\text{Core}_{\text{var}}(x)\| + i^2 \cdot \|\beta(x, y)\|)$, we conclude that φ' can be computed in time $O(\|\varphi^*\| + |\bar{x}|^2 + |\bar{x}| \cdot \|\text{Core}_{\text{var}}(x)\| + |\bar{x}|^2 \cdot \|\beta(x, y)\| + p^3 + p^2 \cdot \|\text{Core}_{\text{var}}(x)\| + p^3 \cdot \|\beta(x, y)\|)$ and, hence, φ' can be computed in time $O(\|\varphi^*\|^3 \cdot (\|\text{Core}_{\text{var}}(x)\| + \|\beta(x, y)\|))$. Thus, if $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ is considered to be fixed, then φ' can be computed in time $O(\|\varphi^*\|^3)$ since the sizes of $\text{Core}_{\text{var}}(x)$ and $\beta(x, y)$ depend only on the size of the data exchange. \square

Proof of Theorem 4.8: Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting and Q be a m -ary query over the target schema \mathbf{T} . Assume that Q is rewritable over the core, that is, there exists a first-order formula $\varphi(\bar{x})$ such that for every instance I of \mathbf{S} and every $\bar{a} \in \text{dom}(I)^m$, $\bar{a} \in \underline{\text{certain}}(Q, I)$ if and only if $\mathcal{F}_{\text{core}}(I) \models \varphi(\bar{a})$.

Let $\varphi'(\bar{x})$ be a formula constructed from $\varphi(\bar{x})$ as shown in Proposition A.6. Then given that the canonical universal solution and its cores share the same constants, all the cores of an instance are isomorphic and for every $\bar{a} \in \underline{\text{certain}}(Q, I)$, \bar{a} is in $\text{dom}(I)^m \cap \text{dom}(\mathcal{F}_{\text{univ}}(I))^m$, we conclude that $\varphi'(\bar{x})$ is a rewriting of Q over the canonical universal solution. \square

Proof of Theorem 4.8: Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting and Q be a m -ary query over the target schema \mathbf{T} . Assume that Q is rewritable over the core, that is, there exists a first-order formula $\varphi(\bar{x})$ such that for every instance I of \mathbf{S} and every $\bar{a} \in \text{dom}(I)^m$, $\bar{a} \in \underline{\text{certain}}(Q, I)$ if and only if $\mathcal{F}_{\text{core}}(I) \models \varphi(\bar{a})$.

Let $\varphi'(\bar{x})$ be a formula constructed from $\varphi(\bar{x})$ as shown in Proposition A.6. Then given that the canonical universal solution and its cores share the same constants, all the cores of an instance are isomorphic and for every $\bar{a} \in \underline{\text{certain}}(Q, I)$, \bar{a} is in $\text{dom}(I)^m \cap \text{dom}(\mathcal{F}_{\text{univ}}(I))^m$, we conclude that $\varphi'(\bar{x})$ is a rewriting of Q over the canonical universal solution. \square

Proof of Theorem 3.9 (continued)

We prove that $\mathcal{F}_{\text{core}}$ is locally consistent under FO-equivalence. Since we know that $\mathcal{F}_{\text{univ}}$ is locally consistent under FO-equivalence, and that constants are preserved from canonical solutions to their cores, it is enough to prove that for every

$m, d, k \geq 0$ there exist $d', k' \geq 0$ such that, for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} over $\text{dom}(I)^m$, if $N_{d'}^{\mathcal{F}_{\text{univ}}(I)}(\bar{a}) \equiv_{k'} N_{d'}^{\mathcal{F}_{\text{univ}}(I)}(\bar{b})$, then $N_d^{\mathcal{F}_{\text{core}}(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\text{core}}(I)}(\bar{b})$.

We know that for an arbitrary (finite) instance I , an arbitrary m -tuple \bar{a} in $\text{dom}(I)$ and every $d, k \geq 0$, there exists an FO formula $\varphi_{I, \bar{a}}^{d, k}(\bar{u})$ such that, for every m -tuple \bar{b} in $\text{dom}(I)$, $I \models \varphi_{I, \bar{a}}^{d, k}(\bar{b})$ iff $N_d^I(\bar{a}) \equiv_k N_d^I(\bar{b})$. Furthermore, for every instance I and m -tuple \bar{a} , the quantifier rank of $\varphi_{I, \bar{a}}^{d, k}$ is a function of d and k (meaning that it is independent of I).

Fix $d, k \geq 0$ and consider $\varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}$ for an arbitrary instance I of \mathbf{S} and any $\bar{a} \in \text{dom}(I)^m$. From Proposition A.6 we know that there exists a formula $\varphi'(\bar{x})$ such that for every $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{b})$ then $\mathcal{F}_{\text{core}}(I) \models \varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}(\bar{b})$. The quantifier rank of φ' just depends on the quantifier rank of $\varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}$ and Σ_{st} , meaning that it is independent of I .

Choose d', k' such that $N_{d'}^{\mathcal{F}_{\text{univ}}(I)}(\bar{a}) \equiv_{k'} N_{d'}^{\mathcal{F}_{\text{univ}}(I)}(\bar{b})$ implies that $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{a})$ iff $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{b})$. The latter implies that $\mathcal{F}_{\text{core}}(I) \models \varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}(\bar{a})$ iff $\mathcal{F}_{\text{core}}(I) \models \varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}(\bar{b})$. Since it is always the case that $\mathcal{F}_{\text{core}}(I) \models \varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}(\bar{a})$, we obtain that $\mathcal{F}_{\text{core}}(I) \models \varphi_{\mathcal{F}_{\text{core}}(I), \bar{a}}^{d, k}(\bar{b})$. From this we conclude that $N_d^{\mathcal{F}_{\text{core}}(I)}(\bar{a}) \equiv_k N_d^{\mathcal{F}_{\text{core}}(I)}(\bar{b})$.

Proof of Proposition 4.9

Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting, where $\mathbf{S} = \langle R(\cdot, \cdot), S(\cdot, \cdot), N(\cdot), M(\cdot) \rangle$, $\mathbf{T} = \langle S'(\cdot, \cdot), N'(\cdot), M'(\cdot), P(\cdot), T(\cdot) \rangle$ and Σ_{st} is defined as follows. Predicates S', N' and M' are defined by means of the following rules:

$$\begin{aligned} S(x, y) &\rightarrow S'(x, y), \\ N(x) &\rightarrow N'(x), \\ M(x) &\rightarrow M'(x). \end{aligned}$$

Predicate P is defined to be:

$$\begin{aligned} \exists x R(x, x) &\rightarrow \exists u P(u), \\ \exists x \exists y (\neg R(x, y) \wedge \neg R(y, x) \wedge x \neq y) &\rightarrow \exists u P(u), \\ \exists x \exists y \exists z (R(x, y) \wedge R(y, z) \wedge \neg R(x, z)) &\rightarrow \exists u P(u), \\ \neg \forall x \forall y (S(x, y) \leftrightarrow R(x, y) \wedge \neg \exists z (R(x, z) \wedge R(z, y))) &\rightarrow \exists u P(u), \\ \exists x (N(x) \wedge M(x)) &\rightarrow \exists u P(u). \end{aligned}$$

Basically, if R is not a linear order on the domain of an instance I of \mathbf{S} or S does not correspond to the successor relation of R or there is an element in the intersection of N and M , then P contains at least one element in every solution for I . Finally, predicate T is defined as follows:

$$\begin{aligned} \exists x \exists y (R(x, y) \wedge N(x) \wedge M(y)) &\rightarrow \exists u T(u), \\ \exists x \exists y S(x, y) &\rightarrow \exists u T(u). \end{aligned}$$

Let Q be the following domain independent FO query:

$$\exists x P(x) \vee \exists x \exists y (N'(x) \wedge S'(x, y) \wedge \neg N'(y)) \vee \exists x (N'(x) \wedge M'(x)). \quad (7)$$

We will prove that Q is FO-rewritable over the canonical universal solution and that Q is not FO-rewritable over the core.

Let Q' be the query $\exists x P(x) \vee \exists x \exists y (T(x) \wedge T(y) \wedge x \neq y)$. We show next that Q' is a rewriting of Q over the canonical universal solution, that is, for every instance I of \mathbf{S} with canonical universal solution J , $Q'(J)$ holds iff certain(Q, I) = *true*.

- Assume that $Q'(J)$ holds. If $J \models \exists x P(x)$, then every solution J' of I satisfies this sentence, since there is a homomorphism from J to J' , and, therefore, certain(Q, I) = *true*. Thus, assume that $J \not\models \exists x P(x)$ and $J \models \exists x \exists y (T(x) \wedge T(y) \wedge x \neq y)$. Then $I(R)$ is a linear order, $I(S)$ is the successor relation of this order and there exists $a, b \in \text{dom}(I)$ such that $R(a, b)$, $N(a)$ and $M(b)$. Let J' be a solution of I . To prove that certain(Q, I) = *true*,

we show that $J' \models \exists x \exists y (N'(x) \wedge S'(x, y) \wedge \neg N'(y)) \vee \exists x (N'(x) \wedge M'(x))$. Assume that $J' \not\models \exists x \exists y (N'(x) \wedge S'(x, y) \wedge \neg N'(y))$ and, hence, $J' \models \forall x \forall y (N'(x) \wedge S'(x, y) \rightarrow N'(y))$. Then, $N'(b)$ is in J' since $N'(a)$ is in J' and a' appears before than b' in the successor relation S' . We conclude that $J' \models \exists x (N'(x) \wedge M'(x))$ since $M'(b)$ is in J' .

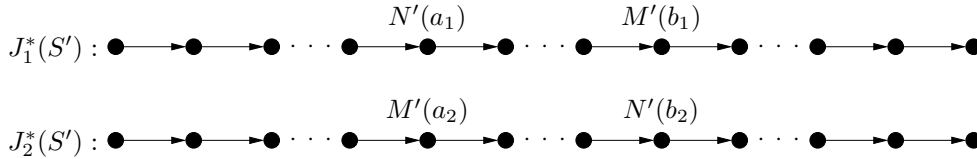
- Assume that $\text{certain}(Q, I) = \text{true}$ and that $Q'(J)$ does not hold. Then $J \not\models \exists x P(x)$, $J \models \exists x \exists y (N'(x) \wedge S'(x, y) \wedge \neg N'(y)) \vee \exists x (N'(x) \wedge M'(x))$ and $J \not\models \exists x \exists y (T(x) \wedge T(y) \wedge x \neq y)$. Hence, $I(R)$ is a linear order, $I(S)$ is the successor relation of this order and all the elements in $I(M)$ appears before than all the elements in $I(N)$ in the order $I(R)$. Let $\{J_n\}_{n \geq 0}$ be a sequence of solutions for I recursively defined as follows: $J_0 := J$ and

$$J_{n+1} := J_n \cup \{N'(b) \mid \text{there exists } a \in \text{dom}(J) \text{ s.t. } N'(a) \text{ is in } J_n \text{ and } S'(a, b) \text{ is in } J\}.$$

Then $J' = \bigcup_{n \geq 0} J_n$ is a solution for I such that $J' \not\models \exists x \exists y (N'(x) \wedge S'(x, y) \wedge \neg N'(y))$. Furthermore, given that all the elements in $I(M)$ appears before than all the elements in $I(N)$ in the order $I(R)$, $J' \not\models \exists x (N'(x) \wedge M'(x))$, which contradicts the fact that $\text{certain}(Q, I) = \text{true}$.

Now we prove that Q is not FO-expressible over the core. On the contrary, assume that there exists a first-order sentence φ such that for every instance I of \mathbf{S} with core solution J^* , $J^* \models \varphi$ iff $\text{certain}(Q, I) = \text{true}$. Let k be the quantifier rank of φ . Define instance I_1, I_2 of \mathbf{S} as follows. $I_i(S)$ is the successor relation of a linear order $I_i(R)$ ($i = 1, 2$) containing k' elements, where k' is a function of k (to be defined later), $I_1(N) = \{a_1\}$, $I_1(M) = \{b_1\}$, $I_2(N) = \{b_2\}$ and $I_2(M) = \{a_2\}$. Furthermore, $(a_i, b_i) \in I_i(R)$ ($i = 1, 2$), the distance between a_i and b_i is $k'/2$ ($i = 1, 2$) and the distance between the first point of $I_i(R)$ and a_i is $k'/4$ ($i = 1, 2$). It is easy to see that $\text{certain}(Q, I_1) = \text{true}$ and $\text{certain}(Q, I_2) = \text{false}$.

The core solutions J_1^*, J_2^* for I_1, I_2 are as follows:



Furthermore, $J_1^*(P), J_2^*(P)$ are empty and $J_1^*(T), J_2^*(T)$ contain only one null value. Thus, if k' is long enough, then $J_1^* \equiv_k J_2^*$ and, therefore, $J_1^* \models \varphi$ iff $J_2^* \models \varphi$. We conclude that $\text{certain}(Q, I_1) = \text{certain}(Q, I_2)$, which leads to a contradiction⁴.

An example of a non-local query under the universal solution semantics

Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting, where $\mathbf{S} = \langle E(\cdot, \cdot), A(\cdot), B(\cdot) \rangle$, $\mathbf{T} = \langle G(\cdot, \cdot), H(\cdot, \cdot), T(\cdot), U(\cdot) \rangle$ and Σ_{st} contains the following source-to-target dependencies: $A(x) \rightarrow T(x), B(x) \rightarrow T(x), B(x) \rightarrow U(x), E(x, y) \rightarrow G(x, y)$, together with:

$$E(x, y) \wedge B(z) \rightarrow H(x, z).$$

Furthermore, let $Q(x)$ be the FO-query over the target schema defined as the disjunction of

$$\exists y \exists z (H(x, y) \wedge H(x, z) \wedge T(y) \wedge T(z) \wedge y \neq z)$$

and $\psi_1, \psi_2, \psi_3, \psi_4$, where ψ_1, ψ_2, ψ_3 and ψ_4 are FO-sentences defined as follows:

$$\begin{aligned} \psi_1 &:= \exists y \exists z (U(y) \wedge U(z) \wedge y \neq z), \\ \psi_2 &:= \exists y \exists z (T(y) \wedge H(y, z) \wedge \neg T(z)), \\ \psi_3 &:= \exists u \exists v \exists y \exists z (H(v, u) \wedge G(v, y) \wedge H(y, z) \wedge T(u) \wedge \neg U(u) \wedge \neg T(z)), \\ \psi_4 &:= \exists y \exists z (H(y, z) \wedge \forall w (H(y, w) \rightarrow z = w)). \end{aligned}$$

⁴We note that $J_1 \not\equiv_k J_2$, where J_1, J_2 are the canonical universal solutions for I_1, I_2 , since J_1 contains two null values in T while J_2 contains only one.

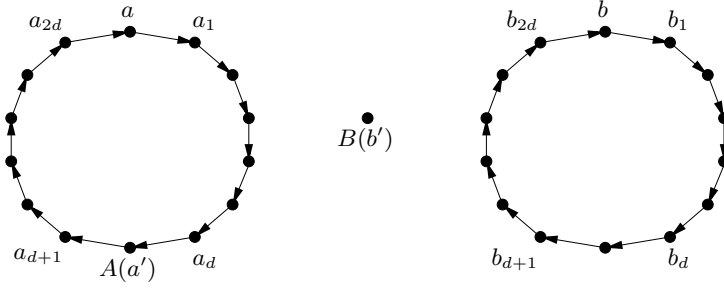


Figure 3: Instance I in the example of a non-local query under the universal solution semantics.

Assume that $Q(x)$ is FO_{aggr} -rewritable over the canonical universal solution under the universal solutions semantics. Then there exists $d \geq 0$ such that for every source instance I and every $a, b \in \text{dom}(I)$, $a \in \underline{\text{u-certain}}(Q, I)$ iff $b \in \underline{\text{u-certain}}(Q, I)$ whenever $N_d^I(a) \cong N_d^I(b)$. Define a source instance I as shown in Figure 3. In the instance, $I(E)$ is the disjoint union of two cycles of length $2d + 2$, $I(A) = \{a'\}$ and $I(B) = \{b'\}$, where b' is an isolated point. We observe that $N_d^I(a) \cong N_d^I(b)$ and, therefore, it should be the case that $a \in \underline{\text{u-certain}}(Q, I)$ iff $b \in \underline{\text{u-certain}}(Q, I)$. Thus, if we prove that $a \in \underline{\text{u-certain}}(Q, I)$ and $b \notin \underline{\text{u-certain}}(Q, I)$, our original assumption that Q is FO_{aggr} -rewritable becomes false. In the next paragraph we show this.

To show that $b \notin \underline{\text{u-certain}}(Q, I)$, we need to exhibit a universal solution J_0 for I such that $J_0 \not\models Q(b)$. Let J_0 as shown in Figure 4, where $c, c', c_1, \dots, c_{2d}, e, e', e_1, \dots, e_{2d}$ are null values, the solid lines represent G -edges and the dashed lines represent H -edges. For the sake of readability, the same element b' appears twice in this figure. Furthermore, define $J_0(U)$ as $\{b'\}$ and $J_0(T)$ as $\{a', b', c, c', c_1, \dots, c_{2d}\}$. It is not hard to see that J_0 is a solution for I . Moreover, since the homomorphism h defined as $h(c_0) = b'$, for every null $c_0 \in \text{dom}(J_0)$, and $h(c_0) = c_0$, for every constant $c_0 \in \text{dom}(J_0)$, maps J_0 into the canonical universal solution, we conclude that J_0 is a universal solution. It can be seen that $J_0 \not\models \psi_1 \vee \psi_2 \vee \psi_3 \vee \psi_4$ and, therefore, $J_0 \not\models Q(b)$ since $e \notin J_0(T)$.

To show that $a \in \underline{\text{u-certain}}(Q, I)$, we need to show that for every universal solution J for I , $J \models Q(a)$. Let J be a universal solution for I . Assume that J does not satisfy $\psi_1 \vee \psi_2 \vee \psi_3 \vee \psi_4$. Given that $b' \in I(B)$ and J is a universal solution for I , we know that $b' \in J(U)$, every point in the cycle containing a in $J(G)$ is connected to b' by means of an H -edge and none of these points is in $J(U)$. Thus, given that $J \models \neg\psi_4$, the cycle in $J(G)$ containing a is as shown in the left hand side of Figure 4, where $c, c', c_1, \dots, c_{2d}$ are null values, the solid lines represent G -edges and the dashed lines represent H -edges. For the sake of simplicity we show these null values as distinct elements, even though some of them can represent the same null value.

Since $a' \in I(A)$, $a' \in J(T)$. Thus, $c' \in J(T)$ given that $J \models \neg\psi_2$. Furthermore, $c' \notin J(U)$ and $c_i \notin J(U)$, for every $i \in [d + 1, 2d]$, since $J \models \neg\psi_1$ and $b' \in J(U)$. We conclude that $c', c_{d+1}, c_{d+2}, \dots, c_{2d}, c \in J(T)$ since $J \models \neg\psi_3$ and, hence,

$$J \models \exists y \exists z (H(a, y) \wedge H(a, z) \wedge T(y) \wedge T(z) \wedge y \neq z).$$

Proof of Theorem 5.2.

For proving this result we need to introduce the notion of local consistency under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence, where $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ is a logic first presented in [16] (but not exactly in the same way that is presented here) and later studied in [24].

The logic $\mathcal{L}_{\infty\omega}$ is the extension of FO with infinitary disjunctions. This means that if $\varphi_i(\bar{x})$ is a formula in $\mathcal{L}_{\infty\omega}$, for all $i \in I$, then $\bigvee_{i \in I} \varphi_i(\bar{x})$ is also a formula in $\mathcal{L}_{\infty\omega}$. We define logic $\mathcal{L}_{\infty\omega}(\mathbf{C})$ over schema \mathbf{R} as two-sorted: first-sort variables range over domains on instances of \mathbf{R} , and second-sort variables range over \mathbb{N} . We refer to the second sort variables with symbols $i, j, k, \bar{i}, \bar{j}, \bar{k}$, etc. Logic $\mathcal{L}_{\infty\omega}(\mathbf{C})$ extends $\mathcal{L}_{\infty\omega}$ by

- *numerical terms and predicates*: There is a second sort constant for every $k \in \mathbb{N}$. Also, if $t_1(\bar{x}), \dots, t_n(\bar{x})$ are terms of the second (numerical) sort, $P(t_1(\bar{x}), \dots, t_n(\bar{x}))$ is an atomic formula with the standard semantics.

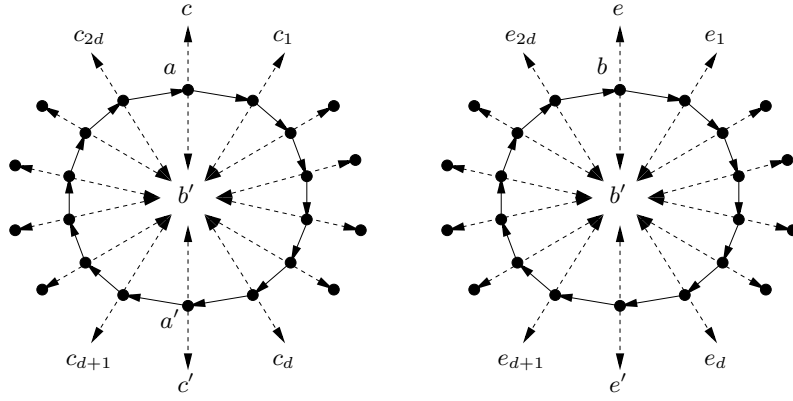


Figure 4: Instance J_0 in the example of a non-local query under the universal solution semantics.

- *counting quantifiers*: If $\varphi(y, \bar{x}, \bar{j})$ is an $\mathcal{L}_{\infty\omega}(\mathbf{C})$ formula, then $\psi(\bar{x}, k, \bar{j}) \equiv \exists ky \varphi(y, \bar{x}, \bar{j})$ is also an $\mathcal{L}_{\infty\omega}(\mathbf{C})$ formula. Note that $\exists ky$ binds y but not k . Furthermore, if c is a fixed natural number, then $\theta(\bar{x}, \bar{k}) \equiv \exists cy \varphi(y, \bar{x}, \bar{k})$ is also a formula in this logic (for example, $\exists 7y \varphi(y, \bar{x}, \bar{k})$ is an $\mathcal{L}_{\infty\omega}(\mathbf{C})$ formula).

The semantics of formula ψ is as follows (the semantics of θ is defined similarly). Suppose we have a structure \mathbf{A} with first sort universe $\{a_1, \dots, a_n\}$. Choose a first sort interpretation \bar{a} for \bar{x} and second sort interpretations \bar{j}_0 for \bar{j} and k_0 for k . Then $\mathbf{A} \models \psi(\bar{a}, k_0, \bar{j}_0)$ iff

$$|\{b \in \{a_1, \dots, a_n\} \mid \mathbf{A} \models \varphi(b, \bar{a}, \bar{j}_0)\}| \geq k_0.$$

The logic $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ is the restriction of $\mathcal{L}_{\infty\omega}(\mathbf{C})$ to formulae of bounded quantifier depth. As in the case of Ehrenfeucht-Fraïssé games and FO logic, it is possible to relate bijective games with $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ expressibility. Indeed, it was shown in [16] and [24] that the duplicator has a winning strategy in the k -round game on structures (\mathbf{A}, \bar{a}) and (\mathbf{B}, \bar{b}) iff for every $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ formula $\varphi(\bar{x}, \bar{j})$ with quantifier rank at most k and every tuple \bar{j}_0 of natural numbers, $\mathbf{A} \models \varphi(\bar{a}, \bar{j}_0)$ iff $\mathbf{B} \models \varphi(\bar{b}, \bar{j}_0)$. In the following we represent this by $(\mathbf{A}, \bar{a}) \equiv_k^{\text{bij}} (\mathbf{B}, \bar{b})$.

It is shown in [17] that $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ is strictly more expressive than FO_{aggr} . This implies that if the duplicator has a winning strategy in the k -round game on structures (\mathbf{A}, \bar{a}) and (\mathbf{B}, \bar{b}) then for every FO_{aggr} formula $\varphi(\bar{x}, \bar{j})$ of quantifier rank k and every tuple \bar{j}_0 of rational numbers, $\mathbf{A} \models \varphi(\bar{a}, \bar{j}_0)$ iff $\mathbf{B} \models \varphi(\bar{b}, \bar{j}_0)$.

We say that a mapping $\mathcal{F} : \text{inst}(\mathbf{S}) \rightarrow \text{inst}(\mathbf{T})$ is *locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence* if for every $m, d, k \geq 0$ there exist $d', k' \geq 0$ such that, for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} over $\text{dom}(I)^m$, if $N_{d'}^I(\bar{a}) \equiv_{k'}^{\text{bij}} N_{d'}^I(\bar{b})$, then

- 1) $\bar{a} \in \text{dom}(\mathcal{F}(I))^m \Leftrightarrow \bar{b} \in \text{dom}(\mathcal{F}(I))^m$, and
- 2) $N_d^{\mathcal{F}(I)}(\bar{a}) \equiv_k^{\text{bij}} N_d^{\mathcal{F}(I)}(\bar{b})$.

We first show that the transformation $\mathcal{F}_{\text{univ}}$ is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence. This follows directly from Propositions A.12 and A.13 below.

Proposition A.12 *Let $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ be a program with just one rule of the form $R(\bar{x}) :- \varphi(\bar{x})$. Then \mathcal{F}_{Π}^* is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence.*

Proof: Similar to the proof of Proposition A.3, but with bijective games instead of Ehrenfeucht-Fraïssé games. \square

For the following results we need some definitions in advance. Fix $d \geq 0$, and define $s_d(l)$ inductively in the following way: $s_d(0) = d$ and $s_d(l+1) = 3 \cdot s_d(l) + 1$. Also, fix $k \geq 0$ and define $r_k(l)$ inductively as $r_k(0) = k$ and $r_k(l+1) = r_k(l) + 2 \lceil \log(2 \cdot s_d(l) + 1) \rceil + m_{\mathbf{S}} + 1$, where $m_{\mathbf{S}}$ is the maximum arity of a predicate in \mathbf{S} .

Proposition A.13 Let $\Pi = (\mathbf{S}, \mathbf{A}, \mathbf{T}, \mathcal{R})$ be a data exchange program with just one rule of the form $R(\bar{x}, \bar{y}) :- S(\bar{x}, \bar{z})$, where $S \in \mathbf{S}$ and $\bar{y} \neq \emptyset$. Then \mathcal{F}_{Π}^n is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence.

Proof: Condition 1) is trivially satisfied. For condition 2) we proceed as follows. Let us denote the arity of S by l and let $m, d, k \geq 0$. Define inductively the following quantities:

- $s(l, 0) = d$ and $s(l, i + 1) = s_{s(l, i)}(l)$;
- $r(l, 0) = k$ and $r(l, i + 1) = r_{r(l, i)}(l)$.

Choose $d' = s(l, k)$ and $k' = \max\{r(l, k), m_{\mathbf{S}} \cdot d\}$. We will show that for every instance I of \mathbf{S} and $\bar{a}, \bar{b} \in \text{dom}(I)^m$, if $N_{d'}^I(\bar{a}) \equiv_{k'}^{\text{bij}} N_{d'}^I(\bar{b})$, then $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a}) \equiv_k^{\text{bij}} N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. More precisely, we will show how to play a k -round bijective game on $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$ from a k' -round bijective game on $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$. The game is as follows. Let $i \in [1, k]$ and assume that for every $j \in [1, i - 1]$, (\bar{p}_j, \bar{p}'_j) are the elements determined by the strategy in the round j of the bijective game on $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$, and assume that there exist bijections g_1, \dots, g_{i-1} from $\text{dom}(I)^l$ to itself such that for every $j \in [1, i - 1]$, $g_j(\bar{p}_j) = \bar{p}'_j$ and for every $\bar{e} \in \text{dom}(I)^l$,

$$N_{s(l, k-j)}^I(\bar{a}\bar{p}_1 \cdots \bar{p}_{j-1}\bar{e}) \equiv_{r(l, k-j)}^{\text{bij}} N_{s(l, k-j)}^I(\bar{b}\bar{p}'_1 \cdots \bar{p}'_{j-1}g_j(\bar{e})).$$

In round i the duplicator chooses a bijection $f : \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})) \rightarrow \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b}))$ as follows. Given that $N_{s(l, k-i+1)}^I(\bar{a}\bar{p}_1 \cdots \bar{p}_{i-1}) \equiv_{r(l, k-i+1)}^{\text{bij}} N_{s(l, k-i+1)}^I(\bar{b}\bar{p}'_1 \cdots \bar{p}'_{i-1})$, we know by Lemma A.15 (below) that there is a bijection $g_i : \text{dom}(I)^l \rightarrow \text{dom}(I)^l$ such that for every $\bar{e} \in \text{dom}(I)^l$,

$$N_{s(l, k-i)}^I(\bar{a}\bar{p}_1 \cdots \bar{p}_{i-1}\bar{e}) \equiv_{r(l, k-i)}^{\text{bij}} N_{s(l, k-i)}^I(\bar{b}\bar{p}'_1 \cdots \bar{p}'_{i-1}g_i(\bar{e})).$$

This bijection is used in the following way to define f . For every $c \in \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a}))$,

- if c is a null, then it can be unequivocally identified with an instantiation $R(\bar{c}, c, \bar{n}) :- S(\bar{c}, \bar{e})$ of the rule of the program. Let $\bar{c}'\bar{e}' = g_i(\bar{c}\bar{e})$. Then c can be naturally and univocally associated with a null c' through the instantiation $R(\bar{c}', c', \bar{n}') :- S(\bar{c}', \bar{e}')$. Set c' to be $f(c)$ and $\bar{p}_i = \bar{c}\bar{e}$.
- if c is a constant, then $f(c) = g_i(c)$ and $\bar{p}_i = c$.

Claim A.14 The element $f(c)$ defined by the above strategy is in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$.

Proof: The proof of this is direct from the fact that $k' \geq m_{\mathbf{S}} \cdot d$. □

We show that the strategy presented above is a winning strategy for the duplicator by induction on the move $i \leq k$. For $i = 0$ the proof goes as follows. By contradiction assume that (\bar{a}, \bar{b}) is not a partial isomorphism in $\mathcal{F}_{\Pi}^n(I)$. Then without loss of generality, we can assume that there exists $T(\bar{a}')$ in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ such that $T(\bar{b}')$ is not in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, where $\bar{a}' \subseteq \bar{a}$, \bar{b}' is the corresponding subset of \bar{b} and $T \in \langle \mathbf{S}, R \rangle$. Then $T \neq R$ since tuple \bar{a} contains only constants. Furthermore, $T \notin \mathbf{S}$ since (\bar{a}, \bar{b}) is a partial isomorphism between $N_{d'}^I(\bar{a})$ and $N_{d'}^I(\bar{b})$, which leads to a contradiction.

Assume that $i - 1$ moves of the game have been played by following the strategy shown above. For the i -th move, $i \in [1, k]$, the spoiler chooses c_i in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$. Call $\bar{c} = c_1, \dots, c_{i-1}$ and $\bar{c}' = c'_1, \dots, c'_{i-1}$ to the elements played so far in the game on $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. Suppose first that $c_i \in \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})) \cap \text{dom}(I)$ and, hence, $c_i \in \text{dom}(N_{d'}^I(\bar{a}))$. Assume on the contrary that c'_i does not work as a winning duplicator response for the game on $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. Then without loss of generality, there is a tuple $T(\bar{e}, \bar{n}, c_i)$ in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$, such that $T(\bar{e}', \bar{n}', c'_i)$ is not in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, where $\bar{e} \subseteq \bar{a}\bar{c} \cap \text{dom}(I)$, $\bar{n} \subseteq \bar{c} \setminus \text{dom}(I)$, and \bar{e}' and \bar{n}' are their responses in $\bar{b}\bar{c}'$. If $T \in \mathbf{S}$ then \bar{n} is empty and $T(\bar{e}, c_i)$ is in $N_{d'}^I(\bar{a})$. Thus, given that $N_{d'}^I(\bar{a}) \equiv_{k'}^{\text{bij}} N_{d'}^I(\bar{b})$, we conclude that $T(\bar{e}', c'_i)$ is in $N_{d'}^I(\bar{b})$. This implies by the definition of

the strategy for the bijective game and by Claim A.14 that $T(\bar{e}', c'_i)$ is in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, which is a contradiction. If $T = R$, we know that $T(\bar{e}, \bar{n}, c_i) = R(\bar{e}, \bar{n}, c_i)$ comes from an instantiation $R(\bar{e}, \bar{n}, c_i) :- S(\bar{e}, c_i, \bar{e}_1)$ of the rule of the program. Since $|\bar{n}| \geq 1$, all the elements in \bar{e}_1 were already played in $N_{d'}^I(\bar{a})$. Let \bar{e}'_1 be the response to \bar{e}_1 in $N_{d'}^I(\bar{b})$. Then $S(\bar{e}', c'_i, \bar{e}'_1)$ is in $N_{d'}^I(\bar{b})$. By Claim A.14 we obtain that $c'_i \in \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b}))$, and by the definition of the strategy for the bijective game we obtain that $T(\bar{e}', \bar{n}', c'_i)$ is in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, which leads to a contradiction.

Assume now c_i is a null value. Then c_i comes from an instantiation $R(\bar{e}, c_i, \bar{n}) :- S(\bar{e}, \bar{e}_1)$ of the rule of the program, where $\bar{e}\bar{e}_1 \subseteq \text{dom}(I)$ and \bar{n} is a tuple of null values. To c_i we attach its natural response c'_i as previously described. Suppose on the contrary that c'_i does not work as a winning duplicator response for the game on $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$. Then $\bar{e} \subseteq \bar{a}\bar{c} \cap \text{dom}(I)$, $\bar{n} \subseteq \bar{c} \setminus \text{dom}(I)$, $R(\bar{e}, c_i, \bar{n})$ is in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{a})$ and $R(\bar{e}', c'_i, \bar{n}')$ is not in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, where \bar{e}' and \bar{n}' are the responses to \bar{e} and \bar{n} in $\bar{b}\bar{c}'$. Since $|\bar{n}| \geq 1$, all the elements in \bar{e}_1 were already played in $N_{d'}^I(\bar{a})$. Let \bar{e}'_1 be the response to \bar{e}_1 in $N_{d'}^I(\bar{b})$. Then $S(\bar{e}', \bar{e}'_1)$ is in $N_{d'}^I(\bar{b})$. By Claim A.14 we know that $c'_i \in \text{dom}(N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b}))$, and, hence, by the definition of the strategy of the game we obtain that $R(\bar{e}', c'_i, \bar{n}')$ is in $N_d^{\mathcal{F}_{\Pi}^n(I)}(\bar{b})$, which leads to a contradiction. \square

Lemma A.15 *Let $N_{s_d(l)}^I(\bar{a}) \equiv_{r_k(l)}^{\text{bij}} N_{s_d(l)}^I(\bar{b})$. Then there is a 1-1 mapping $f : \text{dom}(I)^l \rightarrow \text{dom}(I)^l$ such that, for every $\bar{c} \in \text{dom}(I)^l$, $N_d^I(\bar{a}\bar{c}) \equiv_k^{\text{bij}} N_d^I(\bar{b}f(\bar{c}))$.*

Proof: Let I_1, I_2 be two instances of \mathbf{S} , $\bar{a} \in \text{dom}(I_1)^l$ and $\bar{b} \in \text{dom}(I_2)^l$. By $(I_1, \bar{a}) \xleftrightarrow{d}^k (I_2, \bar{b})$ we express the fact that there is a bijection $f : \text{dom}(I_1) \rightarrow \text{dom}(I_2)$ such that, for every $c \in \text{dom}(I_1)$, $N_d^{I_1}(\bar{a}c) \equiv_k^{\text{bij}} N_d^{I_2}(\bar{b}f(c))$. We will need the following claim in order to prove the lemma.

Claim A.16 *For every $d, k \geq 0$, if $N_{3d+1}^{I_1}(\bar{a}) \equiv_{k'}^{\text{bij}} N_{3d+1}^{I_2}(\bar{b})$ and $I_1 \xleftrightarrow{d}^k I_2$, then $(I_1, \bar{a}) \xleftrightarrow{d}^k (I_2, \bar{b})$, where $k' = k + 2 \lfloor \log(2d+1) \rfloor + m_{\mathbf{S}} + 1$.*

Proof of Claim A.16: We need to define a bijection $f : \text{dom}(I_1) \rightarrow \text{dom}(I_2)$ such that $N_d^{I_1}(\bar{a}c) \equiv_k^{\text{bij}} N_d^{I_2}(\bar{b}f(c))$. Let us denote by g the bijection determined by the duplicator in the first round of the game between $N_{3d+1}^{I_1}(\bar{a})$ and $N_{3d+1}^{I_2}(\bar{b})$. Then the image under g of an arbitrary element at distance at most $2d+1$ from \bar{a} in I_1 is at distance at most $2d+1$ from \bar{b} in I_2 ; otherwise the spoiler could play “distance” from \bar{a} , contradicting that $N_{3d+1}^{I_1}(\bar{a}) \equiv_{k'}^{\text{bij}} N_{3d+1}^{I_2}(\bar{b})$. Using a similar argument we deduce that for each c in $N_{2d+1}^{I_1}(\bar{a})$, $N_d^{I_1}(\bar{a}c) \equiv_k^{\text{bij}} N_d^{I_2}(\bar{b}g(c))$ and $N_d^{I_1}(c) \equiv_k^{\text{bij}} N_d^{I_2}(g(c))$.

Since $I_1 \xleftrightarrow{d}^k I_2$ and for every c in $N_{2d+1}^{I_1}(\bar{a})$, $N_d^{I_1}(c) \equiv_k^{\text{bij}} N_d^{I_2}(g(c))$, there exists a bijection $h : \text{dom}(I_1) \rightarrow \text{dom}(I_2)$ such that, for every c in $(\text{dom}(I_1) \setminus N_{2d+1}^{I_1}(\bar{a}))$, $N_d^{I_1}(c) \equiv_k^{\text{bij}} N_d^{I_2}(h(c))$. Notice that for an arbitrary c in $(\text{dom}(I_1) \setminus N_{2d+1}^{I_1}(\bar{a}))$, $N_d^{I_1}(c)$ is disjoint from $N_d^{I_1}(\bar{a})$ and $N_d^{I_2}(h(c))$ is disjoint from $N_d^{I_2}(\bar{b})$, implying that $N_d^{I_1}(\bar{a}c) \equiv_k^{\text{bij}} N_d^{I_2}(\bar{b}h(c))$.

We now define $f : \text{dom}(I_1) \rightarrow \text{dom}(I_2)$ by

$$f(c) = \begin{cases} g(c) & \text{if } c \text{ in } N_{2d+1}^{I_1}(\bar{a}) \\ h(c) & \text{if } c \text{ not in } N_{2d+1}^{I_1}(\bar{a}) \end{cases}$$

Clearly f is a bijection. Moreover, for each $c \in \text{dom}(I_1)$, $N_d^{I_1}(\bar{a}c) \equiv_k^{\text{bij}} N_d^{I_2}(\bar{b}f(c))$. \square

Now we continue the proof of the lemma. This goes by induction on l . For $l = 0$ there is nothing to prove. For $l + 1$ we know that $s_d(l+1) = 3 \cdot s_d(l) + 1$ and $r_k(l+1) = r_k(l) + 2 \lfloor \log(2 \cdot s_d(l) + 1) \rfloor + m_{\mathbf{S}} + 1$. Assume that it holds for l and that

$$N_{3s_d(l)+1}^I(\bar{a}) \equiv_{r_k(l)+1}^{\text{bij}} N_{3s_d(l)+1}^I(\bar{b}).$$

Since $I \xleftrightarrow{s_d(l)}^{r_k(l)} I$, we obtain from Claim A.16 we obtain that $(I, \bar{a}) \xleftrightarrow{s_d(l)}^{r_k(l)} (I, \bar{b})$. By definition there is a bijection $g : \text{dom}(I) \rightarrow \text{dom}(I)$ such that, for every $c \in \text{dom}(I)$, $N_{s_d(l)}^I(\bar{a}c) \equiv_{r_k(l)}^{\text{bij}} N_{s_d(l)}^I(\bar{b}g(c))$. By induction hypothesis, for

every $c \in \text{dom}(I)$ there is a bijection $h : \text{dom}(I)^l \rightarrow \text{dom}(I)^l$ such that for every $\bar{e} \in \text{dom}(I)^l$,

$$N_d^I(\bar{a}c\bar{e}) \equiv_k^{\text{bij}} N_d^I(\bar{b}g(c)h(\bar{e})).$$

Then define $f : \text{dom}(I)^{l+1} \rightarrow \text{dom}(I)^{l+1}$ as follows: given $\bar{c} = c\bar{e} \in \text{dom}(I)^{l+1}$, $f(\bar{c}) = g(c)h(\bar{e})$. Clearly f is a bijection and $N_d^I(\bar{a}\bar{c}) \equiv_k^{\text{bij}} N_d^I(\bar{b}f(\bar{c}))$. \square

We now prove that $\mathcal{F}_{\text{core}}$ is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence. In order to do this we need to prove the following proposition.

Proposition A.17 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting and $\varphi(\bar{x}, \bar{j})$ an $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ formula over \mathbf{T} . Then there exists an $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ formula $\varphi'(\bar{x}, \bar{j})$ such that for every instance I of \mathbf{S} , every $\bar{a} \in \text{dom}(I)^m \cap \text{dom}(\mathcal{F}_{\text{univ}}(I))^m$ and every tuple of natural numbers \bar{j}_0 , $\mathcal{F}_{\text{univ}}(I) \models \varphi'(\bar{a}, \bar{j}_0)$ iff $\mathcal{F}_{\text{core}}(I) \models \varphi(\bar{a}, \bar{j}_0)$.*

Proof: Assume without loss of generality that $\bar{j} = \bar{j}_1 \cdots \bar{j}_p$ and $\varphi(\bar{x}, \bar{j}, j_1, \dots, j_p)$ is of the form

$$(\neg)\exists j_1 y_1 (\neg)\exists j_2 y_2 \dots (\neg)\exists j_p y_p \psi(\bar{x}, y_1, \dots, y_p, \bar{v}),$$

where ψ is a quantifier-free formula. It is not hard to check that $\varphi'(\bar{x}, \bar{v}, j_1, \dots, j_p)$ defined as

$$\text{Core}(\bar{x}) \wedge (\neg)\exists j_1 y_1 (\text{Core}(y_1) \wedge (\neg)\exists j_2 y_2 (\text{Core}(y_1, y_2) \wedge (\dots \wedge (\neg)\exists j_p y_p (\text{Core}(y_1, y_2, \dots, y_p) \wedge \psi(\bar{x}, y_1 \dots y_p, \bar{v})) \dots))),$$

where predicate Core is defined as in the proof of Theorem 4.8, satisfies the condition of the proposition. \square

From this we derive that $\mathcal{F}_{\text{core}}$ is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence. The proof is similar to the proof of Theorem 3.9, but we consider bijective games instead of Ehrenfeucht-Fraïssé games. Notice that Theorem 4.5 continues being valid when \mathcal{F} is a transformation that is locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence. This shows that every target query that is FO_{aggr}-rewritable over $\mathcal{F}_{\text{univ}}$ or $\mathcal{F}_{\text{core}}$ is locally source-dependent.

An anomaly of the usual semantics

We note that there are some reasons not to be completely satisfied with the standard certain answers semantics. For example, it satisfies the following property.

Proposition A.18 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a data exchange setting. Then for every Boolean query Q over \mathbf{T} , either $\text{certain}(Q, I)$ is false for all instances I of \mathbf{S} , or $\text{certain}(\neg Q, I)$ is false for all instances I of \mathbf{S} .*

Proof: Let Q be a Boolean query over \mathbf{T} , and assume that there exists an instance I_0 of \mathbf{S} such that $\text{certain}(Q, I_0) = \text{true}$. Then we show that for every instance I of \mathbf{S} , $\text{certain}(\neg Q, I) = \text{false}$.

Let I be an instance of \mathbf{S} and J a solution for I . Then given a solution J_0 for I_0 , the instance J' defined as $J'(R) = J(R) \cup J_0(R)$, for every $R \in \mathbf{T}$, is a solution for both I and I_0 . Since $\text{certain}(Q, I_0) = \text{true}$, $Q(J')$ is true and, therefore, there is a solution of I not satisfying $\neg Q$. We conclude that $\text{certain}(\neg Q, I) = \text{false}$. \square

Next we show that the universal solution semantics avoids the problem shown in the previous Proposition, that is, there exists a Boolean query Q such that $\text{u-certain}(Q, I_1) = \text{true}$ and $\text{u-certain}(\neg Q, I_2) = \text{true}$, for some instances I_1 and I_2 .

Given a data exchange setting with $\mathbf{S} = \langle P(\cdot), R(\cdot) \rangle$, $\mathbf{T} = \langle P'(\cdot), R'(\cdot) \rangle$ and $\Sigma_{st} = \{P(x) \rightarrow P'(x), R(x) \rightarrow R'(x)\}$, let Q be a Boolean query over \mathbf{T} defined as $\exists x (P'(x) \wedge R'(x))$. Define instances I_1, I_2 of \mathbf{S} as $\{P(a), R(a)\}$ and $\{P(a), R(b)\}$, respectively. Then both $\text{u-certain}(Q, I_1)$ and $\text{u-certain}(\neg Q, I_2)$ are true (if J is a universal solution for I , there is a homomorphism $h : J \rightarrow \mathcal{F}_{\text{univ}}(I) = \{P'(a), R'(b)\}$ and, hence, for every null value c in J it could not be the case that $P'(c)$ and $R'(c)$ are in J).

Proof of Theorem 5.3.

As in the proof of Theorem 5.2, we obtain this result by noticing that the following restatement of Theorem 4.5 is true:

Theorem A.19 *Let $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ be a FO_{aggr} -data exchange setting, and Q a query over \mathbf{T} . Assume that Q is FO_{aggr} -rewritable over a transformation \mathcal{F} under the universal solution semantics, where \mathcal{F} is either locally consistent or locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence. Then Q is locally source-dependent under the universal solution semantics.*

Proof: We prove the theorem for the case of locally consistent transformations. The proof for transformations that are locally consistent under $\mathcal{L}_{\infty\omega}^*(\mathbf{C})$ -equivalence is similar.

Let Q' be a FO_{aggr} rewriting of Q over \mathcal{F} under the universal solution semantics, that is, an m -ary FO_{aggr} query over \mathbf{T} such that for every instance I of \mathbf{S} , $\underline{\text{u-certain}}(Q, I) = Q'(\mathcal{F}(I))$. Assume that \mathcal{F} is locally consistent. In [24] it was shown that there exists a constant d' such that for every instance J of \mathbf{T} and m -tuples \bar{a}, \bar{b} in J , if $N_{d'}^J(\bar{a}) \cong N_{d'}^J(\bar{b})$, then $\bar{a} \in Q'(J)$ if and only if $\bar{b} \in Q'(J)$. Given that \mathcal{F} is locally consistent, there exists $d \geq 0$ such that for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} in I , if $N_d^I(\bar{a}) \cong N_d^I(\bar{b})$, then (1) $\bar{a} \in \text{dom}(\mathcal{F}(I))$ if and only if $\bar{b} \in \text{dom}(\mathcal{F}(I))$ and (2) $N_{d'}^{\mathcal{F}(I)}(\bar{a}) \cong N_{d'}^{\mathcal{F}(I)}(\bar{b})$. From this we conclude that Q is locally source-dependent under the universal solution semantics since for every instance I of \mathbf{S} and m -tuples \bar{a}, \bar{b} in I ,

$$\begin{aligned} N_d^I(\bar{a}) \cong N_d^I(\bar{b}) &\Rightarrow N_{d'}^{\mathcal{F}(I)}(\bar{a}) \cong N_{d'}^{\mathcal{F}(I)}(\bar{b}) \\ &\Rightarrow \bar{a} \in Q'(\mathcal{F}(I)) \text{ iff } \bar{b} \in Q'(\mathcal{F}(I)) \\ &\Rightarrow \bar{a} \in \underline{\text{u-certain}}(Q, I) \text{ iff } \bar{b} \in \underline{\text{u-certain}}(Q, I). \end{aligned}$$

□

Proof of Proposition 5.4.

Part 1)

For the proof, we introduce a notion of locality based on Hanf's condition [14, 10]. Given two structures \mathbf{A} and \mathbf{B} of the same vocabulary, we write $\mathbf{A} \stackrel{\text{u}}{\leftrightarrow}_d \mathbf{B}$ if there is a bijection $f : A \rightarrow B$ between their universes such that $N_d^{\mathbf{A}}(a) \cong N_d^{\mathbf{B}}(f(a))$ for every $a \in A$. In particular, $\mathbf{A} \stackrel{\text{u}}{\leftrightarrow}_d \mathbf{B}$ implies that $|A| = |B|$.

Given a data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$ and a Boolean query Q over the target, we say that Q is *Hanf-locally source-dependent* if there is a number d such that for any two source instances I and I' , it is the case that $I \stackrel{\text{u}}{\leftrightarrow}_d I'$ implies

$$\underline{\text{certain}}(Q, I) = \underline{\text{certain}}(Q, I').$$

Equivalently we define the notion of locally source-dependent under the universal solution semantics. We need the following lemma.

Lemma A.20 *Given a LAV data exchange setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, let Q be a Boolean query over the target schema that is FO_{aggr} -rewritable over the core, or over the canonical universal solution. Then Q is Hanf-locally source-dependent. This holds for both the usual and the universal solutions semantics.*

Proof of Lemma A.20. We start by showing the following claim. For every positive integer d there exists another positive integer d' such that

$$I_1 \stackrel{\text{u}}{\leftrightarrow}_{d'} I_2 \Rightarrow \mathcal{F}_{\text{univ}}(I_1) \stackrel{\text{u}}{\leftrightarrow}_d \mathcal{F}_{\text{univ}}(I_2). \quad (8)$$

To show this, let Π be the data exchange program given by the setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$. Then, by Lemma A.1, for some fixed number c depending on the data exchange setting, we have an equivalent data exchange program Π' in which all rules r

are of the form $R(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})$, where φ is c -bounded, and each predicate appears at most once as the head of a rule. Now, to prove (8), it suffices to show that for every d , there exists d' such that

$$I_1 \stackrel{d'}{\rightleftharpoons} I_2 \quad \Rightarrow \quad \mathcal{F}_r(I_1) \stackrel{d}{\rightleftharpoons} \mathcal{F}_r(I_2), \quad (9)$$

where $\mathcal{F}_r(I)$ is the result of applying a single rule r of the form $R(\bar{x}, \bar{y}) :- \varphi(\bar{x}, \bar{z})$. Then (8) follows by induction on the number of rules in Π' .

Let $d > 0$ be given. Assume that \bar{x} is not empty. Then every null produced by the rule r is at distance one from a constant. Since $\varphi(\bar{x}, \bar{z})$ is an FO or FO_{aggr} formula, there exists a number k_0 , that depends on the quantifier rank of φ , such that $(I_1, \bar{a}_1, \bar{b}_1) \stackrel{k_0}{\rightleftharpoons} (I_2, \bar{a}_2, \bar{b}_2)$ implies $I_1 \models \varphi(\bar{a}_1, \bar{b}_1)$ iff $I_2 \models \varphi(\bar{a}_2, \bar{b}_2)$ [17]. Then, according to [24], there exists a number k_1 such that $(I_1, \bar{a}_1) \stackrel{k_1}{\rightleftharpoons} (I_2, \bar{a}_2)$ implies that there exists a bijection $f : \text{dom}(I_1)^m \rightarrow \text{dom}(I_2)^m$ such that $(I_1, \bar{a}_1, \bar{b}) \stackrel{k_0}{\rightleftharpoons} (I_2, \bar{a}_2, f(\bar{b}))$, where m is the length of \bar{z} . We can assume without loss of generality that $k_1 > d$ (if it is not, simply replace it by $d + 1$). Finally, there exists a number k_2 such that $I_1 \stackrel{k_2}{\rightleftharpoons} I_2$ imply the existence of a bijection $g : \text{dom}(I_1)^p \rightarrow \text{dom}(I_2)^p$ such that that $(I_1, \bar{a}_1) \stackrel{k_1}{\rightleftharpoons} (I_2, g(\bar{a}_2))$, where p is the length of \bar{x} [24]. Summing up, if $I_1 \stackrel{k_2}{\rightleftharpoons} I_2$ for every \bar{a}_1 in $\text{dom}(I_1)^p$, $N_{k_1}^{I_1}(\bar{a}_1) \cong N_{k_1}^{I_2}(g(\bar{a}_1))$ and

$$|\{\bar{b}_1 \mid I_1 \models \varphi(\bar{a}_1, \bar{b}_1)\}| = |\{\bar{b}_2 \mid I_2 \models \varphi(g(\bar{a}_1), \bar{b}_2)\}|. \quad (10)$$

From the proof of Lemma 3.4, we know that there exists a number k_3 , that depends on d and c only, such that $N_{k_3}^I(\bar{a}) \cong N_{k_3}^I(\bar{b})$ implies that in $\mathcal{F}_r(I)$, the $d + 1$ -neighborhoods of \bar{a} and \bar{b} are isomorphic. We now take d' to be $\max\{k_2, k_3\}$.

Let $I_1 \stackrel{d'}{\rightleftharpoons} I_2$, that is, for some bijection $h : \text{dom}(I_1) \rightarrow \text{dom}(I_2)$, we have $N_{d'}^{I_1}(a) \cong N_{d'}^{I_2}(h(a))$. Since $d' \geq k_3$, we see that the same mapping h shows that when restricted to constants, $N_{d+1}^{\mathcal{F}_r(I_1)}(a) \cong N_{d+1}^{\mathcal{F}_r(I_2)}(h(a))$.

Next, consider a null v in $\mathcal{F}_r(I_1)$. Then this null is generated by two tuples \bar{a}_1, \bar{b}_1 such that $I_1 \models \varphi(\bar{a}_1, \bar{b}_1)$; in particular, in $\mathcal{F}_r(I)$ it is at distance 1 from \bar{a}_1 . Let g be the bijection that guarantees (10) (which exists, since $d' \geq k_2$). We thus can map all the nulls generated by \bar{a}_1 to all the nulls generated by $g(\bar{a}_1)$, and (10) ensures that this map is a bijection. Furthermore, since $d + 1$ -neighborhoods of \bar{a}_1 and $g(\bar{a}_1)$ are isomorphic, this map also preserves d -neighborhoods of the nulls, thereby proving $\mathcal{F}_r(I_1) \stackrel{d}{\rightleftharpoons} \mathcal{F}_r(I_2)$.

If \bar{x} is empty, then the rule generates only isolated nulls (not connected to constants) and the counting argument above suffices. This finished the proof of (8).

To show how Lemma A.20 follows from (8), let Q be a query that is FO_{aggr}-rewritable, by some query Q' , over the canonical universal solution. By [17], there is a number d such that $J \stackrel{d}{\rightleftharpoons} J'$ implies $Q'(J) = Q'(J')$. Let d' be given by (8), and let $I_1 \stackrel{d'}{\rightleftharpoons} I_2$. Then $\mathcal{F}_{\text{univ}}(I_1) \stackrel{d}{\rightleftharpoons} \mathcal{F}_{\text{univ}}(I_2)$ and

$$\underline{\text{certain}}(Q, I_1) = Q'(\mathcal{F}_{\text{univ}}(I_1)) = Q'(\mathcal{F}_{\text{univ}}(I_2)) = \underline{\text{certain}}(Q, I_2).$$

For queries rewritable over core, let Q' an FO_{aggr} query such that $\underline{\text{certain}}(Q, I)$ is true iff $Q'(\mathcal{F}_{\text{core}}(I))$ is true. Let k be the quantifier rank of Q' . From Proposition A.17, we conclude that there is an FO_{aggr} query Q'' such that $Q''(\mathcal{F}_{\text{univ}}(I)) = Q'(\mathcal{F}_{\text{core}}(I))$. Then the previous proof applies. Finally, the proof applies verbatim to the universal solutions semantics. This proves the lemma. \square

We now use Lemma A.20 to conclude the proof of Part 1 as follows.

Consider the LAV (and GAV) setting $(\mathbf{S}, \mathbf{T}, \Sigma_{st})$, where $\mathbf{S} = \langle E \rangle$, $\mathbf{T} = \langle E', D \rangle$ and Σ_{st} consists of the following source-to-target dependencies

$$E(x, y) \rightarrow E'(x, y), \quad E(x, y) \rightarrow D(x), \quad E(x, y) \rightarrow D(y).$$

Furthermore, define a query Q as

$$\exists x \exists y (E'(x, y) \wedge \neg D(x) \wedge \neg D(y)) \rightarrow \exists u \exists v [E'(u, v) \wedge \neg D(u) \wedge \neg D(v) \wedge \forall z (E'(z, u) \rightarrow D(z))].$$

We will prove that Q is not FO_{aggr}-rewritable over \mathcal{F} , where \mathcal{F} is $\mathcal{F}_{\text{univ}}$ or $\mathcal{F}_{\text{core}}$, under the universal solution semantics.

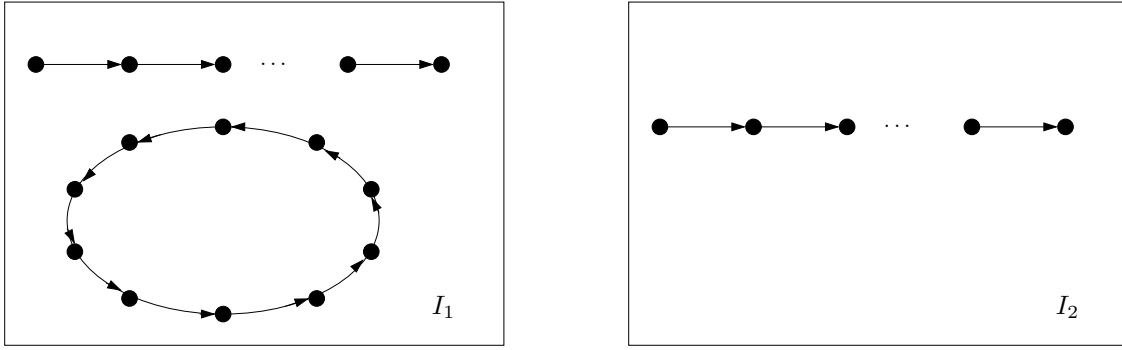


Figure 5: Instances I_1 and I_2 .

Suppose Q is rewritable. By Lemma A.20, find a number d such that $I_1 \stackrel{\text{u-certain}}{\leftrightarrow}_d I_2$ implies $\text{u-certain}(Q, I_1) = \text{u-certain}(Q, I_2)$. Let I_1 and I_2 be two source instances such that I_1 is a disjoint union of a directed cycle of length $2d + 2$ and a successor relation of length $2d + 2$, and I_2 is a successor relation of length $4d + 4$, see Figure 5. Clearly $I_1 \stackrel{\text{u-certain}}{\leftrightarrow}_d I_2$, implying $\text{u-certain}(Q, I_1) = \text{u-certain}(Q, I_2)$. However, we shall prove that $\text{u-certain}(Q, I_1) \neq \text{u-certain}(Q, I_2)$.

Let $J_i = \mathcal{F}_{\text{univ}}(I_i) = \mathcal{F}_{\text{core}}(I_i)$, $i = 1, 2$. First we show that $\text{u-certain}(Q, I_1) = \text{false}$. Consider an instance J'_1 such that $J'_1(D) = J'_1(E')$ and $J'_1(E')$ is equal to $J_1(E')$ plus a (directed) cycle of null values with the same cardinality as the cycle in J_1 . Then J'_1 is a universal solution for I_1 since the function sending each null value in J'_1 to a different element in the cycle of the constants is a homomorphism from J'_1 to J_1 . Moreover, given that all the null values in J'_1 are in the cycle, it is not hard to see that $Q(J'_1) = \text{false}$ and, hence, $\text{u-certain}(Q, I_1) = \text{false}$.

Next, we prove that $\text{u-certain}(Q, I_2) = \text{true}$. Assume on the contrary that $\text{u-certain}(Q, I_2) = \text{false}$. Hence, for some universal solution J'_2 for I_2 that contains at least two nulls in the relation E' ,

$$J'_2 \models \forall u[(\neg D(u) \wedge \exists v(E'(u, v) \wedge \neg D(v))) \rightarrow \exists z(E'(z, u) \wedge \neg D(z))].$$

Thus, J'_2 contains at least one cycle of null values. Therefore, there is no homomorphism from J'_2 to J_2 , contradicting that J'_2 is a universal solution. This shows that Q is not FO_{aggr} -rewritable over \mathcal{F} under the universal solutions semantics.

At the same time, it is not hard to see that under the usual semantics, $\text{certain}(Q, I) = \text{false}$ for every source instance I . Therefore, Q is FO_{aggr} -rewritable over \mathcal{F} under the usual semantics.

Part 2)

It was shown in [8] that there is a conjunctive query Q with one inequality that is not FO_{aggr} -rewritable over \mathcal{F} under the usual semantics. Moreover, it was shown in [9] that existential queries are FO -rewritable in the core under the universal solution semantics and, hence, Q is FO_{aggr} -rewritable over the core under this semantics. Given that Theorem 4.8 holds for the universal solution semantics as well (this is a corollary of the proof of Theorem 4.8, since this does not rely on the underlying semantics), we conclude that Q is FO_{aggr} -rewritable over the canonical universal solution under the universal solution semantics.