On the Data Complexity of Consistent Query Answering*

Balder ten Cate
UC Santa Cruz
btencate@cs.ucsc.edu

Gaëlle Fontaine
UC Santa Cruz
gaelle@cs.ucsc.edu

Phokion G. Kolaitis
UC Santa Cruz &
IBM Research—Almaden
kolaitis@cs.ucsc.edu

ABSTRACT

The framework of database repairs is a principled approach to managing inconsistency in databases. In particular, the consistent answers of a query on an inconsistent database provide sound semantics and the guarantee that the values obtained are those returned by the query on every repair of the given inconsistent database. In this paper, we carry out a systematic investigation of the data complexity of the consistent answers of conjunctive queries for set-based repairs and with respect to classes of constraints that, in recent years, have been extensively studied in the context of data exchange and data integration. Our results, which range from polynomial-time computability to undecidability, complement or improve on earlier work, and provide a fairly comprehensive picture of the data complexity of consistent query answering. We also address the problem of finding a "representative" or "useful" repair of an inconsistent database. To this effect, we introduce the notion of a universal repair, as well as relaxations of it, and then apply it to the investigation of the data complexity of consistent query answering.

Categories and Subject Descriptors
H.2.4 [**Systems**]: Relational databases

General Terms
Algorithms, Theory

Keywords
Inconsistent databases, constraints, repairs, consistent answers

1. INTRODUCTION

An inconsistent database is a database that fails to satisfy one or more integrity constraints that the data are hand are supposed to obey. Inconsistency in databases arises in a variety of applications, including data integration and data warehousing, where the task is to bring together data distributed over different sources that may obey mutually incompatible constraints. In practice, inconsistency is handled mainly via data cleaning, which means that the inconsistent database is transformed, through deletions or additions, to a consistent one that is then used for query answering or for warehousing purposes. This process, however, forces arbitrary choices to be made since, in general, there is a multitude of ways in which an inconsistent databases can be transformed to a consistent one. Arenas, Bertossi and Chomicki [4] introduced a principled approach to the management of inconsistency by formulating the notions of a repair of an inconsistent database and of consistent query answering. Intuitively, a *repair* of an inconsistent database I is a consistent database J that differs from I in a "minimal" way. Furthermore, the *consistent answers* of a query q on an inconsistent database I are defined to be the intersection $\bigcap\{q(J) : J \text{ is a repair of } I\}$. Thus, the inconsistencies in the database are kept, but are handled at query time by considering all repairs and returning the tuples that are guaranteed to be in the result of the query on every repair.

Two algorithmic problems concerning repairs of inconsistent databases naturally arise. The first is, of course, the problem of computing the consistent answers of a query over an inconsistent database. The second is the *repair checking* problem, which can be thought of as the model-checking problem for repairs: given two database instances I and J, is J a repair of I? Since the publication of [4] in 1999, these two problems have been extensively explored for different types of repairs (set-based repairs, cardinality-based repairs, attribute-based repairs) and for different types of constraints. As regards types of constraints, the earlier work on repair checking and consistent query answering focused on functional dependencies, inclusion dependencies, and denial constraints (see the overviews [6, 10]). More recently, broader classes of constraints, such as tuple-generating dependencies (tgds) and equality-generating dependencies (egds), have also been considered in the study of repair checking and consistent query answering. As is well known, these classes of constraints were originally investigated in the context of classical dependency theory, but in the past decade have found numerous uses in the context of data integration and data exchange. For some of these broader classes of constraints, the repair-checking problem has been studied in [2] and [19], and the consistent query answering problem in [26] and [3].

In this paper, we systematically explore the *data complexity* of consistent query answering for sets of tgds and egds. This means that for every fixed set Σ of tgds and egds and for every fixed conjunctive query q, we consider the complexity of the following algorithmic problem: given a database instance I, compute the consistent answers of q on I w.r.t. Σ. Concerning the types of repairs, we consider set-based repairs, that is, subset-repairs, superset-repairs, and $\oplus$-repairs (symmetric difference repairs).

Our main results about the data complexity of consistent query answering, together with previously known results, are summarized in Table 1. In this table, by an entry such as "C/C-comp." we mean that for every fixed set of dependencies in the class considered and

*This research was partially supported by NSF Grant IIS-0905276

ICDT '12 Berlin, Germany

Dependencies	$\star$-repair checking, where $\star \in \{\oplus, \text{subset}, \text{superset}\}$	superset-CQA	subset-CQA	$\oplus$-CQA
LAV tgds	PTIME†	PTIME†	PTIME†	PTIME†
GAV tgds[1]	PTIME [26]	PTIME [13] (implicit)	coNP[26] /coNP-comp.†	coNP[26]/coNP-comp.†
weakly acyclic tgds[1]	coNP/coNP-comp. [2, 19]	PTIME [13] (implicit)	Π_2^p/ Π_2^p-comp.†	Π_2^p/ Π_2^p-comp.†
arbitrary tgds[1]	coNP/coNP-comp. [2, 19]	undecidable [22] (implicit)	Π_2^p/ Π_2^p-comp.†	undecidable†

[1] Indicates that all results hold also in the presence of egds. †Indicates new result obtained in this paper.

Table 1: Data complexity of the repair checking problem and the consistent query answering problem for conjunctive queries.

for every fixed conjunctive query, the problem is in C and that there is a set of dependencies in the class and a conjunctive query for which the problem is C-complete.

One finding of our investigation is that the class of LAV (local-as-view) tgds exhibits tractable algorithmic behavior as regards the data complexity of consistent query answering with respect to all three types of set-based repairs considered here; this extends earlier results about inclusion dependencies and subset-repairs [11]. Another finding is that there is a set (in fact, a singleton set) of GAV (global-as-view) tgds for which the data complexity of the consistent query answering with respect to both subset-repairs and $\oplus$-repairs is coNP-complete; this strengthens a coNP-completeness result for GAV tgds and functional dependencies with respect to $\oplus$-constraints in [26]. We also show that there are sets of weakly acyclic sets of tgds for which the data complexity of the consistent query answering problem is Π_2^p-complete with respect to both subset-repairs and $\oplus$-repairs; earlier, Π_2^p-completeness results had been obtained for the data complexity of consistent query answering for sets of functional dependencies and universal constraints [26] with respect to $\oplus$-repairs. Finally, we show that the assumption of weak acyclicity is of the essence for the decidability of the consistent query answering problem. Specifically, we show that there is a fixed set of tgds and a fixed conjunctive query for which the consistent query answering problem is undecidable with respect to $\oplus$-repairs; furthermore, a similar undecidability result holds for superset-repairs. Previously, it was known that the consistent query answering problem was undecidable in combined complexity for conjunctive queries and for sets of inclusion dependencies and functional dependencies with respect to superset-repairs [25]. It was also known that the consistent query answering problem was undecidable in combined complexity for unions of conjunctive queries and for sets of inclusion dependencies and functional dependencies with respect to $\oplus$-repairs [9]. Finally, it was known that there is a fixed set of universal constraints and a fixed universal query for which consistent query answering is undecidable with respect to $\oplus$-repairs [3].

In addition to the data complexity of consistent query answering, we also addressed the following question: For which types of dependencies is it the case that, given a database instance, there is an efficient way to compute a "representative" and "useful" repair?

We formalized and answered this question by introducing the notion of a *universal repair* and the notion of an *n-universal repair*, where n is a positive integer. These notions are analogous to and, in fact, are motivated from the notion of a universal solution in data exchange [13]. Furthermore, the notion of universal repair is closely related to the notion of *nucleus* in [27], since a universal repair is nucleus that is also a repair. Informally, a universal repair is a repair such that the consistent answers of an arbitrary conjunctive query can be computed by essentially evaluating the query on the universal repair. Similarly, an n-universal repair has the same property but only for conjunctive queries with at most n atoms. We study the existence of universal repairs and of n-universal repairs, as well as structural properties of such repairs and the complexity of computing them. We show that, if Σ is a set of LAV tgds, then every database instance has a unique universal $\oplus$-repair that is also a universal subset-repair and can be computed in polynomial time. Furthermore, while not every database instance has a universal superset-repair, we show that for every $n \geq 1$, an n-universal superset-repair exists and can be computed in polynomial time. If Σ is a weakly acyclic set of tgds and egds, things are the other way around: every database instance has a universal superset-repair that can be computed in polynomial time, but not every instance has a universal, or even a 1-universal, $\oplus$-repair; furthermore, similar limitations hold true for subset-repairs.

2. BASIC NOTIONS

A *schema* $\mathbf{R}$ is a finite sequence $(R_1, \ldots, R_k)$ of relation symbols, each of a fixed arity. An *instance* I over $\mathbf{R}$ is a sequence $(R_1^I, \ldots, R_k^I)$, where each R_i^I is a relation of the same arity as R_i. For notational simplicity, we shall write R_i to denote both the relation symbol and the relation R_i^I that interprets it. A *fact* of an instance I (over $\mathbf{R}$) is an expression $R_i(v_1, \ldots, v_m)$, where R_i is one of the relations of I and $v_1, \ldots, v_m$ are values such that $(v_1, \ldots, v_m) \in R_i$. Every instance can be identified with the set of its facts. We assume that all instances I considered are finite, which means that every relation R_i of I is finite, for $1 \leq i \leq k$.

DEFINITION 2.1 (DEPENDENCIES). *A* tuple-generating dependency (tgd) *is a first-order sentence of the form*

$$\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \exists \mathbf{y}\, \psi(\mathbf{x}, \mathbf{y})),$$

where ϕ, ψ are conjunctions of atomic formulas, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ are tuples of variables, and every universally quantified variable x_i occurs in ϕ.

An equality-generating dependency (egd) *is a first-order sentence of the form*

$$\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow x_k = x_\ell),$$

where ϕ is a conjunction of atomic formulas, $\mathbf{x} = (x_1, \ldots, x_n)$ is a tuple of variables, $1 \leq k, \ell \leq n$, and each universally quantified variable x_i occurs in ϕ.

By the term dependency*, we will mean a tgd or an egd. Also, by a* set of dependencies*, we will mean a finite set of dependencies.*

DEFINITION 2.2. *A* local-as-view *or, simply,* LAV *tgd is a tgd $\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \exists \mathbf{y}\, \psi(\mathbf{x}, \mathbf{y}))$ in which ϕ is a single atomic formula.*

A global-as-view *or, simply,* GAV *tgd is a tgd $\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \psi(\mathbf{x}'))$ in which ψ is a single atomic formula such that the variables in $\mathbf{x}'$ are among the variables of $\mathbf{x}$.*

For example, the *copy* tgd $\forall x \forall y(E(x, y) \rightarrow F(x, y))$ is both a LAV tgd and a GAV tgd. In contrast,

$$\forall x \forall y(E(x, y) \rightarrow \exists z(F(x, z) \wedge F(z, y)))$$

is a LAV tgd that is not a GAV tgd, while

$$\forall x \forall y \forall z (E(x,z) \wedge E(z,y) \rightarrow F(x,y))$$

is a GAV tgd that is not a LAV tgd. Note that every inclusion dependency is a LAV tgd. Furthermore, every tgd with no existential quantifiers in its right-hand side (such tgds are called *full*) is logically equivalent to a set of GAV tgds.

From now on and for the sake of readability, we will often drop the universal quantifiers when writing dependencies.

As mentioned in the Introduction, tgds play an important role in data exchange and data integration, where they are used to specify the relationship between a source (local) schema and a target (global) schema or to express constraints in a target (global) schema. Moreover, it is known that *weakly acyclic* sets of tgds have good algorithmic properties as regards data exchange that, in general, are not possessed by arbitrary sets of tgds.

DEFINITION 2.3 (WEAK ACYCLICITY [12, 13]). *Let Σ be a set of tgds and egds over a schema $\mathbf{S}$.*

- *The* dependency graph *of Σ is the following directed graph.*

  The nodes are the pairs (R,i), where $R \in \mathbf{S}$ is a relation of some arity k, and $1 \leq i \leq k$. We call such pairs positions*.*

  There is an edge from position (R,i) to position (S,j) if Σ contains a tgd $\forall \mathbf{x}(\phi \rightarrow \exists \mathbf{y}\, \psi)$ such that some variable from $\mathbf{x}$ occurs in position (R,i) in ϕ and in position (S,j) in ψ.

  There is a special edge *from position (R,i) to position (S,j) if Σ contains a tgd $\forall \mathbf{x}(\phi \rightarrow \exists \mathbf{y}\, \psi)$ such that (i) some variable from $\mathbf{x}$ occurs in position (R,i) in ϕ and also occurs in ψ, and (ii) some variable from $\mathbf{y}$ occurs in position (S,j) in ψ.*
- *We say that Σ is* weakly acyclic *if the dependency graph contains no cycle going through a special edge.*
- *We say that a tgd θ is* weakly acyclic *if so is the set $\{\theta\}$.*

Every set of GAV tgds is weakly acyclic, since the dependency graph has no special edges. It is also easy to see that every acyclic set of inclusion dependencies is weakly acyclic. However, the set $\Sigma = \{D(x,y) \rightarrow M(y), M(y) \rightarrow \exists x D(x,y)\}$ is a cyclic, yet weakly acyclic set of inclusion dependencies. Finally, the inclusion dependency $R(x,y) \rightarrow \exists z R(y,z)$ is not weakly acyclic because the dependency graph contains a self-loop on position $R.2$.

Intuitively, a *repair* of an instance I with respect to a set of dependencies Σ is an instance J that satisfies Σ and "differs minimally" from I. Here, we will consider three different set-theoretic types of repairs. If I and J are instances, we will write $I \oplus J$ to denote the *symmetric difference* $(I \setminus J) \cup (J \setminus I)$ of I and J, that is, $I \oplus J$ is the set of all facts that either belong to I and not to J, or belong to J and not to I.

DEFINITION 2.4 (REPAIRS). *Let I, J be instances, and Σ a set of dependencies. We say that J is an* $\oplus$-repair *of I w.r.t. Σ if $J \models \Sigma$ and there is no instance J' such that $J' \models \Sigma$ and $I \oplus J' \subsetneq I \oplus J$.*

If, in addition, $J \subseteq I$ (or, $I \subseteq J$), then J is called a subset-repair *(respectively, a* superset-repair*) of I w.r.t. Σ.*

Equivalently, J is a subset-repair of I w.r.t. Σ if $J \subseteq I$, $J \models \Sigma$ and there is no instance $J' \subseteq I$ such that $J' \models \Sigma$ and $J \subsetneq J'$. Likewise, J is a superset-repair of I w.r.t. Σ if $I \subseteq J$, $J \models \Sigma$, and there is no instance J' such that $I \subseteq J'$, $J' \models \Sigma$ and $J' \subsetneq J$.

For example, let $\Sigma = \{P(x) \rightarrow Q(x)\}$, $I = \{P(a), P(b)\}$, $J_1 = \{P(a), P(b), Q(a), Q(b)\}$, $J_2 = \emptyset$, and $J_3 = \{P(a), Q(a)\}$. All three instances J_1, J_2, J_3 are $\oplus$-repairs of I. However, only J_1 is a superset-repair of I, and only J_2 is a subset-repair of I. Furthermore, J_1 is the only superset-repair of I, while J_2 is the only subset-repair of I.

For a different example, let $\Sigma = \{R(x,y) \rightarrow \exists z R(y,z)\}$ and $I = \{R(1,2)\}$. It is easy to see that, for every $n \geq 1$, the cycle

$$J_n = \{R(i,i+1) \mid 1 \leq i < n\} \cup \{R(n,1)\}$$

is a superset-repair (hence, also a $\oplus$-repair) of I w.r.t. Σ. Thus, I has infinitely many, and arbitrarily large, superset-repairs w.r.t. Σ.

Next, we give the definitions of the main algorithmic problems in the study of repairs.

DEFINITION 2.5 (REPAIR CHECKING). *Let Σ be a set of dependencies, and let $\star \in \{\oplus, subset, superset\}$. The* $\star$-repair checking problem *w.r.t. Σ asks: given instances I and I', check whether I' is a $\star$-repair of I w.r.t. Σ.*

DEFINITION 2.6 (CONSISTENT ANSWERS). *Let Σ be a set of dependencies, q a conjunctive query, I an instance, and $\star \in \{\oplus, subset, superset\}$.*

- *The* $\star$-consistent answers *of q on I w.r.t. Σ, denoted by $\star$-Con(q, I, Σ), is the intersection*

$$\bigcap \{q(J) \mid J \text{ is a } \star\text{-repair of } I \text{ w.r.t. } \Sigma\}.$$

- *The* $\star$-consistent query answering problem *of q w.r.t. Σ asks: given an instance I, compute the $\star$-consistent answer of q on I with respect to Σ.*

Several remarks are in order now. First, we wish to emphasize that only finite instances and only finite repairs are considered in this paper. It is known that there exist sets of dependencies Σ, conjunctive queries q, and instances I such that the consistent answer to q on I w.r.t. Σ would yield a different result if infinite repairs would be considered as well (see [25]). Second, if Σ consists of tgds and egds, there is always a subset repair and, hence, also an $\oplus$-repair. Indeed, the empty instance satisfies Σ, hence, for every instance I that fails to satisfy Σ, there is a maximal (w.r.t. $\subseteq$) instance J such that $J \subseteq I$ and $J \models \Sigma$. Finally, it is easy to see that, for all $\star \in \{\oplus, \text{subset}, \text{superset}\}$, it is the case that the $\star$-repair checking problem is in coNP. Moreover, the subset-consistent query answering is always in Π_2^p. For example, to check that J is not an $\oplus$-repair of I, one has to guess an instance J' of size at most $|I| + |J|$ and verify that $J' \models \Sigma$ and $I \oplus J' \subset I \oplus J$.

Note that the notion of *superset-consistent answers* is closely related to the notion of *certain answers* in incomplete information and in data exchange. To make the connection with data exchange precise, recall first that if $\mathcal{M}$ is a schema mapping, q is a query over the target schema of $\mathcal{M}$, and I is an instance over the source schema of $\mathcal{M}$, then the certain answers of q on I w.r.t. $\mathcal{M}$, denoted by certain$(q, I, \mathcal{M})$, are defined as follows:

$$\text{certain}(q, I, \mathcal{M}) = \bigcap \{q(J) | J \text{ is solution for } I \text{ w.r.t. } \mathcal{M}\}.$$

The connection between superset-consistent answers and certain answers is given by the next proposition.

PROPOSITION 2.7. *Let Σ is a set of dependencies and let $\mathcal{M}_\Sigma$ be the schema mapping defined as follows:*

(i) The source-to-target tgds of $\mathcal{M}_\Sigma$ are copy tgds of the form $R'(\mathbf{x}) \rightarrow R(\mathbf{x})$, where R is a relation symbol of the schema of Σ and R' is a new relation symbol whose arity is that of R.

(ii) The target tgds of $\mathcal{M}_\Sigma$ are the tgds in Σ.

Then for every conjunctive query q and every instance I,

$$\text{superset-Con}(q, I', \Sigma) = \text{certain}(q, I, \mathcal{M}_\Sigma),$$

where I' is the copy of the instance I obtained by changing the name of every relation R of I to R'.

The preceding proposition is proved by combining the monotonicity of conjunctive queries with the fact that every solution J for I' w.r.t. $\mathcal{M}_\Sigma$ contains a superset-repair of I. We will make use of this connection in some of our complexity results for the superset-consistent query answering problem.

3. UNIVERSAL REPAIRS

By analogy to the notion of a *universal solution* in data exchange [13], we will now introduce the notion of a *universal repair*. Since it will turn out that universal repairs often do not exist, we also introduce a weaker notion, namely, that of a n-universal repair, $n \geq 1$.

DEFINITION 3.1. *Let Σ be a set of dependencies, I an instance, and $\star \in \{\oplus, subset, superset\}$.*

- *A* universal $\star$-repair *of I w.r.t. Σ is a $\star$-repair J of I w.r.t. Σ such that if q is a conjunctive query, then*

$$\star\text{-Con}(q, I, \Sigma) = q(J)_\downarrow,$$

  where $q(J)_\downarrow$ is the set of tuples in $q(J)$ containing only values from the active domain of I.

- *For $n \geq 1$, an* n-universal $\star$-repair *of I w.r.t. Σ is a $\star$-repair J of I w.r.t. Σ such that if q is conjunctive query with at most n atoms, then*

$$\star\text{-Con}(q, I, \Sigma) = q(J)_\downarrow.$$

Clearly, a universal repair is also a n-universal repair, for every $n \geq 1$. The notion of a universal repair is closely related to the notion of a nucleus [27]. More precisely, a universal repair is a nucleus that is also a repair.

The next proposition, which follows immediately from the definitions and the fact that the data complexity of conjunctive queries is in PTIME, justifies the introduction of the notions of universal repairs and n-universal repairs.

PROPOSITION 3.2. *Assume that Σ is a set of dependencies and that n is a positive integer such that there is a polynomial-time algorithm that, given a instance I, it returns an n-universal $\star$-repair of I, where $\star \in \{\oplus, subset, superset\}$. Then the data complexity of the $\star$-consistent query answering problem for conjunctive queries with at most n atoms is in* PTIME.

Several examples are in order now.

EXAMPLE 3.3. Let $\Sigma = \{R(x, y) \rightarrow \exists z R(y, z)\}$.
Consider the instance $I = \{R(1, 2), R(2, 2), R(1, 3), R(3, 4)\}$. Then the instance $J = \{R(1, 2), R(2, 2)\}$ is a universal subset-repair of I w.r.t. Σ, because it is the only subset-repair of I.

Next, consider the instance $I = \{R(1, 2)\}$. Then I has no universal superset-repair; indeed, this is so because every superset-repair of I must contain a cycle of some length n_0 (recall that all repairs must be finite), and therefore cannot be universal since, as seen earlier, not every superset-repair contains a cycle of the same length n_0. However, for every $n \geq 1$, there is a n-universal superset-repair of I. Specifically, such a repair is the cycle

$$J_n = \{R(i, i+1) | 1 \leq i < n\} \cup \{R(n, 1)\}.$$

EXAMPLE 3.4. Let $\Sigma = \{P(x) \wedge Q(x) \rightarrow R(x)\}$ and $I = \{P(a), Q(a)\}$. The instance $J = \{P(a), Q(a), R(a)\}$ is a universal superset-repair of I w.r.t. Σ, because it is the only superset-repair of I. However, I has no universal subset-repair, since its subset-repairs are $\{P(a)\}$ and $\{Q(a)\}$. In fact, I does not even have a 1-universal subset-repair, since the queries $P(x)$ and $Q(x)$ return different values on the two subset-repairs of I.

EXAMPLE 3.5. Let $\Sigma = \{P(x) \wedge Q(y) \rightarrow x = y\}$ and $I = \{P(a), Q(b)\}$. First, I has no superset-repair w.r.t. Σ, hence it has no 1-universal superset-repair. Furthermore, it has neither a 1-universal subset-repair, nor a 1-universal $\oplus$-repair w.r.t. Σ.

The next proposition describes some basic properties of the different types of universal repairs, and the relationship between them.

PROPOSITION 3.6. *Let Σ be a set of dependencies.*

1. *Every instance has at most one universal superset-repair, up to isomorphism.*
2. *An instance has a universal subset-repair if and only if it has exactly one subset-repair. Consequently, every instance has at most one universal subset-repair.*
3. *Every universal $\oplus$-repair of an instance I is a universal subset-repair of I. Consequently, every instance has at most one universal $\oplus$-repair. In contrast, a universal subset-repair need not be a universal $\oplus$-repair.*

PROOF. 1. Let J_1 and J_2 be universal superset-repairs of I. Let q_1 and q_2 be the canonical conjunctive queries of J_1 and J_2, where the values from the active domain of I are treated as free variables of the queries, and the values outside of the active domain of I are treated as existentially quantified variables of the queries. By universality, $q_1(J_1)_\downarrow = q_1(J_2)_\downarrow$ and $q_2(J_1)_\downarrow = q_2(J_2)_\downarrow$. This implies that there are homomorphisms $h_1 : J_1 \rightarrow J_2$ and $h_2 : J_2 \rightarrow J_1$ such that $h_1(a) = a$ and $h_2(a) = a$ for all values a from the active domain of I. Consequently, J_1 and J_2 are homomorphically equivalent, when the values from the active domain of I are treated as constant. Furthermore, since J_1 and J_2 are repairs, it follows that no instance J' such that $I \subseteq J' \subsetneq J_1$ or $I \subseteq J' \subsetneq J_2$ satisfies Σ. Hence, J_1 and J_2 are cores. But it is a well known fact that any two homomorphically equivalent core instances are isomorphic (cf. [14]).

2. If an instance I has exactly one subset-repair J, then, trivially, J is a universal subset-repair. If, on the other hand, there are (at least) two different subset-repairs J_1 and J_2, then, by the definition of repairs, neither is a subset of the other, and hence, neither can be universal (each J_i, $i = 1, 2$, contains a fact, say $R(\mathbf{a})$, that is not included in all subset-repairs; therefore, J_i does not correctly compute the consistent answers for the query $R(\mathbf{x})$).

3. Let J be a universal $\oplus$-repair of I w.r.t. $\mathcal{M}$. Let J' be a subset-repair of I (recall that every instance has a subset-repair). Since J is a universal $\oplus$-repair, every fact $R(\mathbf{a})$ of I that belongs to J also belongs also to J' (for, if not, then evaluating the query $q(\mathbf{x}) = R(\mathbf{x})$ in J would not yield the consistent answers). Consequently, $J' \oplus I \subseteq J \oplus I$, and therefore, by the definition of $\oplus$-repairs, $J = J'$. Hence J is a subset-repair.

Not every universal subset-repair is a universal $\oplus$-repair. To see this, let $\Sigma = \{P(x) \rightarrow \exists y\, Q(y), P(x) \wedge Q(x) \rightarrow R(x)\}$, let $I = \{P(a), Q(a)\}$, and let $J = \{Q(a)\}$. Then J is a universal subset-repair of I, but not a universal $\oplus$-repair, as may be seen by considering the $\oplus$-repair $J' = \{P(a), Q(b)\}$ and the query $q(x) = Q(x)$. $\square$

Note that parts 2 and 3 of Proposition 3.6 hold with "n-universal" in place of "universal", where n is an arbitrary positive integer. The

reason is that universality was applied only to queries with a single atom. Note also that the notion of a *universal superset-repair* is intimately linked to the notion of a *core universal solution* in data exchange [14], as seen in the proof of Proposition 3.6. We will further exploit this connection when we consider superset-repairs for weakly acyclic sets of tgds and egds.

4. LAV TGDS

The main result of this section is that if Σ is a set of LAV tgds and q is a conjunctive query, then there is a polynomial-time algorithm for computing the $\star$-consistent answers of q on a given instance I w.r.t. Σ, where $\star \in \{\oplus, \text{subset}, \text{superset}\}$. We also show that there is a polynomial-time algorithm for the $\star$-repair checking problem w.r.t. Σ, where $\star \in \{\oplus, \text{subset}, \text{superset}\}$. In the special case in which Σ is a set of inclusion dependencies, the tractability of the subset-repair checking problem and the subset-consistent query answering problem for conjunctive queries was established in [11]. The tractability of the subset and the $\oplus$-repair checking problem for a weakly acyclic set of LAV tgds was shown in [2] and was subsequently extended to the broader class of semi-LAV sets of tgds [19] (which is still a subclass of the class of weakly acyclic sets of tgds). Finally, the $\oplus$-consistent query problem for sets of inclusion dependencies and universal constraints was studied in [8], where disjunctive logic programming was used to obtain the $\oplus$-consistent answers, but no complexity results were established.

In obtaining our tractability results, we will use extensively the notions of universal repair and of n-universal repair, $n \geq 1$, for sets of LAV tgds.

We say that a set of dependencies Σ is *closed under union* if for all instances I_1, I_2 such that $I_1 \models \Sigma$ and $I_2 \models \Sigma$, we have that $I_1 \cup I_2 \models \Sigma$. It is well known that every set of LAV tgds is closed under union (e.g., see [2]). The next result shows that, in a certain sense, this property is characteristic of LAV tgds. For completeness, we also include a proof that sets of LAV tgds are closed under union.

THEOREM 4.1. *Every set of LAV tgds is closed under union. Furthermore, if Σ is a set of tgds that is weakly acyclic and closed under union, then Σ is logically equivalent to a set of LAV tgds.*

PROOF. Let Σ be a set of LAV tgds. Suppose that I_1 and I_2 are instances such that both satisfy Σ. Let

$$\forall \mathbf{x}(R(\mathbf{x}) \rightarrow \exists \mathbf{y}\, \psi(\mathbf{x}, \mathbf{y}))$$

be a tgd in Σ. Then, for every tuple of values $\mathbf{a}$ such that $R(\mathbf{a})$ holds in $I_1 \cup I_2$, we have that $R(\mathbf{a})$ already holds in I_1 or I_2, and hence $\exists \mathbf{y}\, \psi(\mathbf{a}, \mathbf{y})$ holds in I_1 or in I_2, which implies that it also holds in $I_1 \cup I_2$. This shows that $I_1 \cup I_2 \models \Sigma$.

We now prove that the converse holds for weakly acyclic sets of tgds. Let Σ be a weakly acyclic set of tgds that is closed under union. Since the schema of Σ consists of finitely many relation symbols, there are only finitely many possible facts up to renaming of their members. Since Σ is weakly acyclic, we have that if f is a fact, then, by chasing $\{f\}$ with Σ, we obtain a finite instance I_f that contains f and has the property that for every instance J containing f and satisfying Σ, there is a homomorphism from I_f to J (in effect, I_f is a universal solution of $\{f\}$ w.r.t. a schema mapping with copy source-to-target tgds and with Σ as its set of target tgds - see [13]). Without loss of generality, we may assume that, for distinct facts f and f', the active domains of the instances I_f and $I_{f'}$ intersect only on values occurring both in f and in f'. Let σ_f be the LAV tgd whose left-hand side is f and whose right-hand side is the canonical query of I_f (the values from the instance $\{f\}$ are treated as universally quantified variables of σ_f). Let Σ' be the set of all LAV tgds σ_f, as f varies over all possible facts.

We now show that Σ is logically equivalent to Σ'. First, Σ logically implies Σ'. Indeed, this follows from the weak acyclicity of Σ, the monotonicity of the chase, and the aforementioned properties of the chase. Conversely, assume that $I \models \Sigma'$, where $I = \{f_1, \ldots, f_n\}$. Let $J = I_{f_1} \cup \ldots \cup I_{f_n}$. Since Σ is closed under union and each instance I_{f_i} satisfies Σ, we have that $J \models \Sigma$. Furthermore, by weak acyclicity and the aforementioned properties of the chase, there is a homomorphism $h : J \rightarrow I$ that is the identity on I. In other words, I is a retract of J. Since the truth of tgds is preserved when passing from an instance to any one of its retracts (cf. [18]), we have that $I \models \Sigma$. $\square$

We leave it as open problem whether or not the weak acyclicity hypothesis is indispensable in establishing Theorem 4.1. Closure under union implies that every instance has a universal $\oplus$-repair. In fact, we have the following:

PROPOSITION 4.2. *Let Σ be a set of dependencies. The following statements are equivalent:*

1. *Σ is closed under union.*
2. *Every instance has a universal $\oplus$-repair w.r.t. Σ.*
3. *Every instance has a universal subset-repair w.r.t. Σ.*

PROOF. [1$\Rightarrow$2] Let Σ be a set of tgds that is closed under union, and let I be an instance. Since the empty instance satisfies Σ, I has at least one subset-repair w.r.t. Σ. In fact, I has a unique subset-repair[1]. For, if J_1 and J_2 are subset-repairs of I, then $J_1 \cup J_2$ is also a subinstance of I satisfying Σ, and hence, by the definition of repairs, $J_1 = J_1 \cup J_2 = J_2$. Clearly, the unique subset-repair J of I is a universal subset-repair of I. Furthermore, we claim that J is a universal $\oplus$-repair of I. Let q be a conjunctive query. We have to show that

$$\oplus\text{-Con}(q, I, \Sigma) = q(J)_{\downarrow}.$$

From the definition of $\oplus$-Con(q, I, Σ) and since J is a $\oplus$-repair of I, it follows that $\oplus$-Con$(q, I, \Sigma) \subseteq q(J)_{\downarrow}$. For the other inclusion, let K be a $\oplus$-repair of I. We have to show that $q(J)_{\downarrow} \subseteq q(K)$. Since J and K are $\oplus$-repairs, $J \vDash \Sigma$ and $K \vDash \Sigma$. By closure under union, Σ is also true in $J \cup K$. As $J \subseteq I$, we have $I \oplus (J \cup K) \subseteq I \oplus K$. Since K is a $\oplus$-repair of I, this can only happen if $K \cup J = K$. That is, $J \subseteq K$. It follows that $q(J)_{\downarrow} \subseteq q(K)$. Hence, J is a universal $\oplus$-repair of I.

[2$\Rightarrow$ 3] Follows immediately from Proposition 3.6.

[3$\Rightarrow$ 1] Let I_1, I_2 be instances satisfying Σ. Towards a contradiction, suppose that $I_1 \cup I_2$ does not satisfy Σ. Let J be the universal subset-repair of $I_1 \cup I_2$. Then J must omit some fact of $I_1 \cup I_2$. Without loss of generality, we may assume that J omits a fact of I_1. Since I_1 satisfies Σ, there must be a subset-repair of $I_1 \cup I_2$ that contains all facts in I_1. But this subset-repair must then be incomparable to J, which means that J is not the only subset-repair of I, and hence, by Proposition 3.6, J is not a universal subset-repair of I, a contradiction. $\square$

COROLLARY 4.3. *If Σ is a set of LAV tgds, then every instance I has a unique subset repair, which is also the unique universal subset-repair and the unique universal $\oplus$-repair of I w.r.t. Σ.*

THEOREM 4.4. *Let Σ be a set of LAV tgds. There is a polynomial-time algorithm that, given an instance I, computes the unique universal subset-repair of I w.r.t. Σ (which is also the unique universal $\oplus$-repair of J w.r.t. Σ).*

[1]For inclusion dependencies, this fact was pointed out in [11].

PROOF. Let I be an instance. By Corollary 4.3, I has a unique subset repair I_0, which is also both the unique universal subset-repair and the unique universal $\oplus$-repair of I. We will show how to compute I_0 in polynomial time. We start with a definition. Let τ be a LAV tgd of the form

$$\forall \mathbf{x}\, R(\mathbf{x}) \rightarrow \exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y}).$$

We say that a fact t *falsifies* the LAV tgd τ in an instance J if t is in J, it is of the form $S(\mathbf{a})$, and $\exists \mathbf{y} \psi(\mathbf{a}, \mathbf{y})$ is false in J.

The algorithm works as follows. It starts with the instance I. As long as the current instance contains a fact $S(\mathbf{a})$ that falsifies some tgd in Σ, it removes it from the current instance.

Clearly, the algorithm runs in time polynomial in the size of I. It can be shown by induction that, at any time of the run of the algorithm, the current instance is a superset of the unique subset-repair I_0 of I. Hence, if J is the instance obtained at the end of the run of the algorithm, we have that $I_0 \subseteq J$. Moreover, since there is no tuple falsifying a tgd in Σ in J, we have that Σ is true in J. Since $J \models \Sigma$, $I_0 \subseteq J \subseteq I$, and I_0 is a subset-repair of I, we conclude that I_0 is equal to J. $\square$

The situation for superset-repairs is very different. Example 3.3 shows that an instance may have infinitely many superset-repairs of arbitrarily large size, and it may have no universal superset-repairs. The same example shows that superset-repairs may not exist. Nevertheless, we will show that if Σ is a set of LAV tgds then, for every instance I and every $n \geq 1$, there is an n-universal superset-repair for I w.r.t. Σ that can be computed in polynomial time in the size of I. For this, we need the following lemma.

LEMMA 4.5. *Let Σ be a set of LAV tgds. There is a polynomial time algorithm that, given instances I and J with $I \subseteq J$ and $J \models \Sigma$, computes a superset-repair K of I w.r.t. Σ such that $K \subseteq J$.*

PROOF. Figure 1 depicts an algorithm that, given I and J with $I \subseteq J$ and $J \models \Sigma$, either verifies that J is a repair of I or computes a set K such that $I \subseteq K \subsetneq J$ and $K \models \Sigma$. By applying repeatedly the algorithm until it outputs a repair, we obtain the result.

Input: Two instances I and J such that $I \subseteq J$ and $J \models \Sigma$

Output: Either "J is a repair" or an instance K such that $I \subseteq K \subsetneq J$ and $K \models \Sigma$

For each fact $R(\mathbf{a})$ in $J \backslash I$,
  compute the subset-repair L of $J \backslash \{R(\mathbf{a})\}$
  If $I \subseteq L$, **Then Return** L
  End If
End For
Return "J is a repair"

Figure 1: Algorithm for the superset-repair checking problem with respect to a set of LAV tgds

By Theorem 4.4, computing the subset-repair of an instance is in polynomial time. It follows that the algorithm in Figure 1 runs in time polynomial in the sizes of I and J.

We prove the correctness of the algorithm. Suppose first that the algorithm outputs an instance K. We have to show that $I \subseteq K \subsetneq J$ and $K \models \Sigma$. By construction, there exists a fact $R(\mathbf{a})$ in $J \backslash I$ such that K is the subset-repair of $J \backslash \{R(\mathbf{a})\}$. This implies that $K \models \Sigma$ and $K \subsetneq J$. Moreover, K contains I, as the algorithm returned K.

Next, assume that J is not a superset-repair. We have to prove that the algorithm does not output "J is a repair". Since J is not a superset-repair, there exists an instance J_0 such that $I \subseteq J_0 \subsetneq J$ and $J_0 \models \Sigma$. Take $R(\mathbf{a})$ in $J \backslash J_0$. Let L be the subset-repair of $J \backslash \{R(\mathbf{a})\}$.

Consider the instance $L_0 := L \cup J_0$. By Theorem 4.1, $L_0 \models \Sigma$. Moreover, $L_0 \subseteq J \backslash \{R(\mathbf{a})\}$ and $J \oplus L_0 \subseteq J \oplus K$. Since K is a subset-repair of J, this can only happen if $L_0 = K$. That is, $J_0 \subseteq L$. Together with $I \subseteq J_0$, we obtain $I \subseteq L$. Recall that L is the subset-repair of $J \backslash \{R(\mathbf{a})\}$. This means that, when the algorithm enters the loop and picks $R(\mathbf{a})$, it stops running and returns the instance L. In particular, the algorithm does not output "J is a repair". $\square$

THEOREM 4.6. *Let Σ be a set of LAV tgds, and let $n \geq 1$. There is a polynomial time algorithm that, given an instance I, computes an n-universal superset-repair of I w.r.t. Σ.*

PROOF. We first show how to reduce the problem to the case where Σ consists of inclusion dependencies. For every LAV tgd

$$\forall \mathbf{x} \phi(\mathbf{x}) \rightarrow \exists \mathbf{y} \bigwedge_{i=1}^{m} \psi_i(\mathbf{x}, \mathbf{y})$$

in Σ, where each ψ_i is an atomic formula, we introduce a new relation R, whose arity is equal to the number of variables in $\mathbf{y}$, and we replace the LAV tgd by the inclusion dependencies

$$\forall \mathbf{x}(\phi(\mathbf{x}) \rightarrow \exists \mathbf{y}\, R(\mathbf{y})) \quad \text{and} \quad \forall \mathbf{y}(R(\mathbf{y}) \rightarrow \psi_i(\mathbf{y})),\ 1 \leq i \leq m.$$

Let Σ' be the set of inclusion dependencies obtained by applying this transformation to each LAV tgd in Σ. It is not hard to see that every repair of I w.r.t. Σ induces a repair of I w.r.t. Σ', and vice versa. Hence, if we prove the desired result for Σ', then the result for Σ will follow. In [25], it was shown that in the case of inclusion dependencies, for every instance I and every $n \geq 1$, one can construct an instance J such that $J \models \Sigma$ and for every conjunctive query q of size $\leq n$, the following statements are equivalent:

1. $J \models q$.
2. For every K such that $J \subseteq K$ and $K \models \Sigma$, we have that $K \models q$.

Inspection of the algorithm in [25] shows that it runs in polynomial time for fixed Σ and n.

By Lemma 4.5, we can compute in polynomial time a superset-repair J_0 of I such that $J_0 \subseteq J$. If q is a conjunctive query, then $J_0 \models q$ implies $J \models q$. With the properties of J, this implies that for every conjunctive query q of size $\leq n$, we have that if $J_0 \models q$, then for all K such that $K \subseteq J$ and $K \models \Sigma$, it is the case that $K \models q$. Thus, J_0 is a n-universal superset-repair of I w.r.t. Σ. $\square$

The main result of this section is now an immediate consequence of Proposition 3.2, Theorem 4.4, and Theorem 4.6.

THEOREM 4.7. *For every set Σ of LAV tgds, for every conjunctive query q, and for $\star \in \{\oplus, superset, subset\}$, the $\star$-consistent query answering problem w.r.t. Σ is solvable in polynomial time.*

The repair-checking problem. We now consider the repair checking problem for sets of LAV tgds. First, we show how the $\oplus$-repair checking problem can be reduced to the subset-repair checking problem and the superset-repair checking problem.

LEMMA 4.8. *Let Σ be a set of LAV tgds, and let I and J be two instances. Then the following statements are equivalent:*

1. *J is a $\oplus$-repair of I w.r.t. Σ.*
2. *J is a superset-repair of $I \cap J$ w.r.t. Σ and J is a subset-repair of $I \cup J$ w.r.t. Σ.*

PROOF. Let Σ be a set of LAV tgds, and let I and J be two instances. For the direction $(1) \Rightarrow (2)$, assume that J is a $\oplus$-repair

of I. First, we prove that J is a superset-repair of $I \cap J$. Let K_0 be an instance such that $K_0 \models \Sigma$ and $I \cap J \subseteq K_0 \subseteq J$. We have to show that $K_0 = J$. As $I \cap J \subseteq K_0 \subseteq J$, we also have $I \oplus K_0 \subseteq I \oplus J$. Since $K_0 \models \Sigma$ and J is a $\oplus$-repair of I, this can only happen if $K_0 = J$.

Next we prove that J is a subset-repair of $I \cup J$. Let K_1 be an instance such that $K_1 \models \Sigma$ and $J \subseteq K_1 \subseteq I \cup J$. We have to show that $J = K_1$. Since $J \subseteq K_1 \subseteq I \cup J$, we have $I \oplus K_1 \subseteq I \oplus J$. Putting this together with the facts that $K_1 \models \Sigma$ and J is a $\oplus$-repair of I, we obtain $K_1 = J$.

For the direction $(2) \Rightarrow (1)$, assume that J is a superset-repair of $I \cap J$ w.r.t. Σ and that J is a subset-repair of $I \cap J$ w.r.t. Σ. Let K be an instance such that $K \models \Sigma$ and $I \oplus K \subseteq I \oplus J$. We have to show that $K = J$. Since $I \oplus K \subseteq I \oplus J$, we have

$$J \cap I \subseteq K \cap I \text{ and } K \backslash I \subseteq J \backslash I.$$

As $K \backslash I \subseteq J \backslash I$, we have $K \cup J \subseteq I \cup J$. Moreover, since $K \models \Sigma$ and $J \models \Sigma$, it follows from Theorem 4.1 that $K \cup J \models \Sigma$. Recall that J is a subset-repair of $I \cup J$. Since $K \cup J \models \Sigma$ and $J \subseteq K \cup J \subseteq I \cup J$, this implies that $J = K \cup J$. Hence, $K \subseteq J$.

Recall that $J \cap I \subseteq K \cap I$. Putting everything together, we have $I \cap J \subseteq K \subseteq J$. Since $K \models \Sigma$ and J is a repair of $I \cap J$, this implies $K = J$. □

THEOREM 4.9. *Let Σ be a set of LAV tgds, and let $\star \in \{\oplus, \textit{superset}, \textit{subset}\}$. The $\star$-repair checking problem w.r.t. Σ is solvable in polynomial time.*

PROOF. By Lemma 4.8, it suffices to give polynomial-time algorithms for the superset-repair checking problem and the subset-repair checking problem. The fact that the subset-repair checking problem is solvable in polynomial time follows directly from Theorem 4.4 (since it suffices to check for equality with the universal subset-repair). Finally, the fact that the superset-repair checking problem is solvable in polynomial time is a direct consequence of Lemma 4.5 □

Adding egds. The tractability results concerning LAV tgds are optimal, in the sense that if we consider sets of LAV tgds and egds, then most of them do not remain true. We discuss first the repair-checking problem for sets of LAV tgds and egds. In [11], it was shown that there is a set consisting of a cyclic inclusion dependency and a functional dependency for which the subset-repair checking problem is coNP-complete. In [2], the intractability of subset-repair checking (and $\oplus$-repair checking) was shown to hold for the union of an acyclic set of inclusion dependencies with a set of egds. However, the superset-repair checking problem can still be solved in polynomial time. This follows from the next lemma combined with Theorem 4.9.

LEMMA 4.10. *Let Σ be a set of tgds and egds, and let I and J be two instances such that $I \subseteq J$ and $J \models \Sigma$. Then J is a superset-repair of I w.r.t. Σ if and only if J is a superset-repair of I w.r.t. the set Σ_1 consisting of all tgds in Σ.*

PROOF. Assume that $\Sigma = \Sigma_1 \cup \Sigma_2$, where Σ_1 is a set of tgds and Σ_2 is a set of egds. Let I and J be two instances such that $I \subseteq J$ and $J \models \Sigma_1 \cup \Sigma_2$. It is clear that if J is a superset-repair of I w.r.t. Σ_1, then J is a superset-repair of I w.r.t. $\Sigma_1 \cup \Sigma_2$.

For the other direction, assume that J is a superset-repair of I w.r.t. $\Sigma_1 \cup \Sigma_2$. Towards a contradiction, assume that J is not a superset-repair of I w.r.t. Σ_1. Hence, there is an instance J' such that $I \subseteq J' \subsetneq J$ and $J' \models \Sigma_1$.

Since Σ_2 is a set of egds, $J \models \Sigma_2$ and $J' \subseteq J$, we also have that $J' \models \Sigma_2$. So, J' is a superset of I such that $J' \subseteq J$ and $J' \models \Sigma_1 \cup \Sigma_2$. This contradicts the fact that J is a superset-repair of I w.r.t. $\Sigma = \Sigma_1 \cup \Sigma_2$. □

Next, we comment on the consistent query answering problem for sets of LAV tgds and egds. Theorem 4.7 in [11] shows that there is a set Σ of inclusion dependencies and functional dependencies, and a conjunctive query q such that computing the subset-consistent answers of q w.r.t. Σ is a Π_2^p-complete problem. Furthermore, as regards the combined complexity of consistent query answering, it follows from Theorem 5 in [25] that the following problem is undecidable: given a set Σ of LAV tgds and egds, a conjunctive query q and an instance I, compute the superset-consistent answers of q on I w.r.t. Σ. We leave it as an open problem whether or not this undecidability result still holds for some fixed conjunctive query q and some fixed set Σ of LAV tgds and egds. It is also unknown whether or not the data complexity of $\oplus$-consistent answers for conjunctive queries is decidable. As mentioned in the Introduction, the $\oplus$-consistent answering problem for unions of conjunctive queries was shown to be undecidable in combined complexity [9].

Finally, if we consider LAV tgds together with egds, there might not always exist a universal (or even 1-universal) subset-repair, $\oplus$-repair or superset-repair, as seen in Example 3.5.

5. GAV TGDS

In this section, we investigate the existence and efficient computability of universal repairs for GAV tgds, we review known complexity results for repair checking and consistent query answering for GAV tgds, and we provide new matching lower bounds. Furthermore, toward the end of the section, we will briefly consider the addition of egds, and will also discuss semi-LAV sets of tgds; as mentioned earlier, semi-LAV sets of tgds were introduced in [19].

Recall that in the case of LAV tgds, every instance has a unique subset-repair, which is also the unique universal $\oplus$-repair. This is far from true for for GAV tgds. First of all, Example 3.4 shows that there is a set Σ of GAV tgds and an instance I such that I has no 1-universal subset repair w.r.t. Σ; therefore, Proposition 3.6 implies that I also has no 1-universal $\oplus$-repair w.r.t. Σ. In addition, even if an instance happens to have a universal subset-repair, this may not be a universal $\oplus$-repair, as revealed by the next example.

EXAMPLE 5.1. Consider the set

$$\Sigma = \{P(x) \wedge Q(y) \rightarrow M(y), M(x) \rightarrow P(x)\}$$

and the instance $I = \{M(a), Q(b)\}$. The instance $J = \{Q(b)\}$ is then the only subset-repair of I, hence it is also a universal subset-repair of I. However, J is not a universal $\oplus$-repair of I, as can be seen by considering the $\oplus$-repair $J' = \{M(a), P(a)\}$ and the query $Q(x)$. This also shows that, unlike the case of LAV tgds, the subset-consistent answers of a conjunctive query w.r.t. a class of GAV tgds do not always coincide with the $\oplus$-consistent answers of the same query w.r.t. the same class of GAV tgds.

Unlike the case for subset-repairs and $\oplus$-repairs, we will show that, for sets of GAV tgds, every instance has a universal superset-repair. In fact, every instance has a *unique* superset-repair; moreover, the latter property is characteristic for GAV tgds (this will be analogous to the characterization of LAV tgds given in the previous section). To state and proof this characterization, we need a definition and a lemma. We say that a set Σ of dependencies is *closed under intersection* if for all instances I_1, I_2 such that $I_1 \models \Sigma$ and $I_2 \models \Sigma$, we have that $I_1 \cap I_2 \models \Sigma$. Closure under intersection

will turn out to be equivalent to the existence of a unique superset-repair, and this characterizes the GAV tgds among all tgds.

LEMMA 5.2. *Let Σ be a set of tgds and let I be an instance that has a unique superset-repair J w.r.t. Σ. If I' is an instance and J' is a superset-repair of I' w.r.t. Σ, then every homomorphism $h : I \to I'$ can be extended to a homomorphism $h' : J \to J'$.*

PROOF. Here, adom(K) will denote the active domain of an instance K, while rng(h) will denote the range of the function h. Let J'_h be the following instance:

- $adom(J'_h) = adom(I) \cup (\text{adom}(J') \setminus \text{rng}(h))$.
- For every fact $R(b_1, \ldots, b_n)$ of J', we populate J'_h with all possible facts that can be obtained by replacing each value b_i by a value $a_i \in dom(I)$ such that $h(a_i) = b_i$, if $b_i \in \text{rng}(h)$, or leaving b_i untouched, otherwise.

Since $h(I) \subseteq I' \subseteq J'$, we have that $I \subseteq J'_h$. Let h' be a function defined on $adom(J'_h)$ in the following way: if $a \in dom(I)$, then $h(a) = h(a)$; if $a \notin dom(I)$ (but $a \in \text{adom}(J') \setminus \text{rng}(h)$), then $h'(a) = a$. From the construction of J'_h and the definition of h', it follows that h' is a homomorphism from J'_h to J'. To complete the proof, it suffices to show that $J'_h \models \Sigma$. Indeed, once we have shown this, it will follow that $J \subseteq J'_h$, because $I \subseteq J'_h$ and J is the unique superset-repair of I,; consequently, the restriction of h' on $adom(J)$ is a homomorphism from J to J'.

To show that $J'_h \models \Sigma$, consider an arbitrary tgd from Σ, say,

$$\phi(\mathbf{x}) \to \exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y}).$$

Suppose that J'_h satisfies $\phi(\mathbf{a})$ for some tuple of values $\mathbf{a}$. Then, by the construction of J'_h, we have that J' satisfies $\phi(h'(\mathbf{a}))$, and hence, it also satisfies $\psi(h'(\mathbf{a}), \mathbf{b})$ for some values $\mathbf{b} = b_1, \ldots, b_m$. Let $\mathbf{a}' = a'_1, \ldots, a'_m$ be a tuple of values such that $h'(a'_i) = b_i$, for $1 \leq i \leq m$. Then, by construction, J'_h satisfies $\psi(\mathbf{a}, \mathbf{a}')$. □

THEOREM 5.3. *Let Σ be a set of tgds. Then the following statements are equivalent:*

1. *Every instance has a unique (universal) superset-repair w.r.t. Σ.*
2. *Σ is closed under intersection.*
3. *Σ is logically equivalent to a set of GAV tgds.*

PROOF. The proof goes round-robin.

[1⇒3] Assume that every instance I has a unique superset-repair J. It follows that J can only contain values that come from the active domain of I (otherwise, by taking an isomorphic copy in which the values outside of the active domain of I are replaced by fresh values, we would refute the uniqueness of the repair). Consider a tgd $\tau \in \Sigma$ of the form $\phi(\mathbf{x}) \to \exists \mathbf{y}\, \psi(\mathbf{x}, \mathbf{y})$, where ψ is a conjunction of atomic formulas. Let I_ϕ be the canonical instance of $\phi(\mathbf{x})$, and let J_ϕ be the unique superset-repair of I_ϕ. Since J_ϕ is a repair of I_ϕ, it satisfies the tgd τ. It follows that $\exists \mathbf{y}\, \psi(\mathbf{x}, \mathbf{y})$ is satisfied in J_ϕ under the natural assignment. Since J_ϕ contains only values that are from the active domain of I_ϕ, this means that J_ϕ satisfies $\psi(\mathbf{x}, \mathbf{x'})$ for some tuple of values $\mathbf{x}'$ from the active domain of I_ϕ. Let $\widehat{\tau}$ be the dependency

$$\phi(\mathbf{x}) \to \psi(\mathbf{x}, \mathbf{x}'),$$

which can be equivalently written as a finite set of GAV tgds, and let $\widehat{\Sigma} = \{\widehat{\tau} \mid \tau \in \Sigma\}$. We claim that Σ and $\widehat{\Sigma}$ are logically equivalent. The fact that Σ' logically implies Σ follows directly from the construction of Σ'. For the other direction, suppose that $I \models \Sigma$ and $I \models \phi(\mathbf{a})$ for some tuple of values $\mathbf{a}$. Let I_ϕ be again the canonical instance of $\phi(\mathbf{x})$. Then the function $h : \mathbf{x} \mapsto \mathbf{a}$ is a homomorphism from I_ϕ to I. Since J_ϕ is the unique superset-repair of I_ϕ w.r.t. Σ, Lemma 5.2 implies that h extends to a homomorphism from J_ϕ to every superset-repair of I. Since I satisfies Σ, it is its own repair, which means that I satisfies $\psi(\mathbf{a}, h(\mathbf{a}))$. Therefore, $I \models \widehat{\tau}$.

[3⇒ 2] Let Σ be a set of GAV tgds. Let $I_1 \models \Sigma$ and $I_2 \models \Sigma$, and suppose that, for some GAV tgd $t \in \Sigma$ of the form

$$\phi(x_1, \ldots, x_n) \to R(x_{i_1}, \ldots, x_{i_k})$$

it holds that $I_1 \cap I_2 \models \phi(a_1, \ldots, a_n)$ for some values $a_1, \ldots, a_n$. Then both I_1 and I_2 satisfy $\phi(a_1, \ldots, a_n)$; since I_1 and I_2 satisfy Σ, we have that I_1 and I_2 satisfy $R(a_{i_1}, \ldots, a_{i_k})$; thus, the fact $R(a_{i_1}, \ldots, a_{i_k})$ belongs to the intersection of I_1 and I_2.

[2⇒ 1] Let Σ be a set of tgds that is closed under intersection. Let I be an instance and let J be the instance consisting of all possible facts over the active domain of I. Then $I \subseteq J$ and $J \models \Sigma$, hence a superset-repair of I must exist. Suppose that I has two distinct superset-repairs J_1 and J_2. By closure under intersection, $J_1 \cap J_2 \models \Sigma$, hence the properties of superset-repairs imply that $J_1 = J_1 \cap J_2 = J_2$, a contradiction. □

The classical chase procedure from dependency theory (see [1]) provides a method for efficiently computing the unique superset-repair of an instance w.r.t. a set of GAV tgds.

THEOREM 5.4. *Let Σ be a set of GAV tgds. There is a polynomial time algorithm that, given an instance I, computes the unique (universal) superset-repair of I w.r.t. Σ.*

We now move to consistent query answering. Let Σ be a set of GAV tgds and let q be a conjunctive query. The preceding Theorem 5.4 implies immediately that the superset-consistent query answering problem for q w.r.t. Σ is solvable in polynomial time. Staworko and Chomicki [26] showed that both the subset-consistent query answering problem and the $\oplus$-consistent query answering problem for q w.r.t. Σ are in CONP. This will also follow from Theorem 6.2 in the next section, which implies that the size of every $\oplus$-repair of an instance I w.r.t. a set of GAV tgds is bounded by a polynomial in the size of I. In [26], it is also shown that there is a conjunctive query q and a set Σ consisting of GAV tgds and egds such that the $\oplus$-consistent query answering problem for q w.r.t. Σ is CONP-complete. Our next result improves on this lower bound by showing that it can hold even for a set consisting of a single GAV tgd.

THEOREM 5.5. *There is a set Σ consisting of a single GAV tgd and a conjunctive query q such that both the subset-consistent query answering problem and the $\oplus$-consistent query answering problem for q w.r.t. Σ are* CONP*-complete.*

PROOF (HINT). We produce a GAV tgd τ, a conjunctive query q, and a polynomial-time reduction from the complement of POSITIVE 1-IN-3-SAT to the problem of finding the $\oplus$-consistent answers of q w.r.t. $\{\tau\}$. Recall that POSITIVE 1-IN-3-SAT is the following NP-complete problem [17]: given a Boolean formula ϕ in conjunctive normal form and such that each clause is a disjunction of the form $(x_1 \vee x_2 \vee x_3)$ of three positive literals, is there a truth assignment that makes true exactly one variable in every clause?

Let τ be the following GAV tgd:

$$\forall x, u, u'(P(x, u) \wedge P(x, u') \to E(u, u'))$$

and let q be the following conjunctive query:

$$\exists x_1, x_2, x_3, u_1, u_2, u_3 (R(x_1, x_2, x_3) \wedge P(x_1, u_1) \wedge P(x_2, u_2) \wedge P(x_3, u_3) \wedge S(u_1, u_2, u_3)).$$

The intuition behind the relation symbols is as follows: P encodes truth values for the variables in a given Boolean formula,

while E is used to simulate equality. The tgd τ expresses that each variable is assigned at most one truth value.

The relation R encodes every clause $(x_1 \vee x_2 \vee x_3)$ occurring in a formula in conjunctive normal form such that each clause is a disjunction of three positive literals. The relation S will consist of the triples in $\{0,1\}^3 \backslash \{(1,0,0),(0,1,0),(0,0,1)\}$. The conjunctive query q expresses that there are a clause $(x_1 \vee x_2 \vee x_3)$ and truth values assigned to x_1, x_2 and x_3 such that it is not the case that exactly one variable is true in the clause $(x_1 \vee x_2 \vee x_3)$.

Given a Boolean formula ϕ in conjunctive normal form such that each clause is a disjunction of three positive literals, we can construct an instance I such that

$$\oplus\text{-Con}(q, I, \Sigma) = \mathit{true} \quad \text{iff} \quad \phi \notin \text{POSITIVE 1-IN-3-SAT}. \quad \square$$

Extensions: egds and semi-LAV. All complexity upper bounds described in this section hold for the more general case of sets of GAV tgds and egds. Note that if Σ is a set of GAV tgds and egds, then an instance I may not have any superset-repair w.r.t. Σ. Still, if a superset-repair exists, then it is unique, and it can be computed in polynomial time using the chase procedure. The existence of a superset-repair can be tested in polynomial time as well.

In [19], the class of *semi-LAV* sets of dependencies was introduced; it contains properly the class of all sets of GAV tgds, as well as the class of all weakly acyclic sets of LAV tgds. It was shown in [19] that the $\oplus$-repair checking problem for semi-LAV sets of tgds is still solvable in polynomial time (however, this is no longer true if egds are allowed [2]). For $\star$-consistent query answering, where $\star \in \{\text{subset}, \text{superset}, \oplus\}$, the complexity for semi-LAV sets of tgds is the same as the complexity for sets of GAV tgds. The lower bounds naturally transfer, and the upper bounds follow from the fact that repair checking is in polynomial time, together with Theorem 6.1 and Theorem 6.2 in the next section, since every semi-LAV set of tgds is, by definition, weakly acyclic.

6. WEAKLY ACYCLIC SETS OF TGDS (AND EGDS)

In this section, we study the consistent query answering problem for conjunctive queries and weakly acyclic sets of tgds and egds. We begin by considering universal superset-repairs. Recall that, according to Proposition 3.6, every instance has at most one universal superset-repair, up to isomorphism.

THEOREM 6.1. *Let Σ be a weakly acyclic set of tgds and egds. If an instance has a superset-repair w.r.t. Σ, then it has a universal superset-repair. Moreover, there is a polynomial time algorithm that, given an instance I, tests whether it has a superset-repair w.r.t. Σ and if so, computes the (unique up to isomorphism) universal superset-repair of I.*

PROOF. We rely on known results from data exchange. We say that an instance J is a *solution* for an instance I w.r.t. Σ if $I \subseteq J$ and J satisfies Σ. Thus, a superset-repair of I is a solution of I such that no strict subset is a solution for I. In [18], it was shown that if Σ is a fixed set of tgds and egds, then there is a polynomial-time algorithm that, given an instance I, tests whether it has a solution w.r.t. Σ, and if so, computes a solution J satisfying the following additional properties:

(i) For each solution J' of I, there is a homomorphism $h : J \to J'$ that is the identity on values from the active domain of I.

(ii) For each solution J'' satisfying the above condition (i), we have that $J \subseteq J''$.

This solution J is is known as the "core universal solution" of I [14], and is unique up to isomorphism.

Let J be the core universal solution of I w.r.t Σ. Since $I \subseteq J$ and J satisfies Σ, it remains to show that (a) J is a superset-repair of I, i.e., there is no $J' \subsetneq J$ that contains I and satisfies Σ, and (b) J is a *universal* superset-repair of I. The first item follows immediately from the above properties (i) and (ii): any such J' would satisfy condition (i), thereby contradicting the fact that J satisfies condition (ii) above. The second item follows from the fact that J satisfies condition (i) and from the preservation of conjunctive queries under homomorphisms. $\square$

Recall from Example 3.4 that, in general, an instance might not have a 1-universal subset-repair (hence, a 1-universal $\oplus$-repair) w.r.t. a set of weakly acyclic tgds.

Next, we move to the consistent query answering problem. One of the key observations for obtaining an upper bound for the complexity of the $\oplus$-consistent answers is that the size of every $\oplus$-repair of an instance I is polynomial in the size of I. The proof relies on the *solution-aware chase*, which was introduced in the context of peer data exchange [26].

THEOREM 6.2. *Let Σ be a set of weakly acyclic tgds and egds. There is a polynomial $p(x)$ such that for all instances I and for all $\oplus$-repairs J of I with respect to Σ, the size of J is bounded by $p(|I|)$, where $|I|$ is the size of I.*

PROOF. Let Σ be a set of weakly acyclic tgds and egds. The proof relies on the algorithm described in the proof of Lemma 3.4 in [15]. Specifically, the proof of that lemma yields a polynomial $p(x)$ and an algorithm that works as follows. Given instances K and L such that $K \subseteq L$ and $L \models \Sigma$, the algorithm computes an instance $f(K, L)$ such that $K \subseteq f(K, L) \subseteq L$, $f(K, L) \models \Sigma$, and the size of $f(K, L)$ is bounded by $p(|K|)$.

Fix a repair I_0 of I. We show that the size of I_0 is bounded by $p(|I|)$. We can use the algorithm described in Lemma 3.4 in [15] to produce an instance $J := f(I \cap I_0, I_0)$ satisfying the following. The size of J is bounded by $p(|I \cap I_0|)$, $I \cap I_0 \subseteq J \subseteq I_0$ and $J \models \Sigma$. In particular, the size of J is bounded by $p(|I|)$.

In order to show that the size of I_0 is bounded by $p(|I|)$, it is sufficient to show that $I_0 = J$. Since I_0 is a $\oplus$-repair of I, this is equivalent to proving that $J \models \Sigma$ and $I \oplus J \subseteq I \oplus I_0$.

We already know that $J \models \Sigma$. It remains to show that $I \oplus J \subseteq I \oplus I_0$. Let $R(\mathbf{a})$ be a fact in $I \oplus J$. We prove that $R(\mathbf{a})$ belongs to $I \oplus I_0$. Suppose first that $R(\mathbf{a})$ belongs to $J \backslash I$. We know that $J \subseteq I_0$. Hence, if $R(\mathbf{a})$ belongs to $J \backslash I$, then $R(\mathbf{a})$ belongs to $I_0 \backslash I$. This implies that $R(\mathbf{a})$ belongs to $I \oplus I_0$. Next, assume that $R(\mathbf{a})$ belongs to $I \backslash J$. We have to show that $R(\mathbf{a}) \in I \oplus I_0$. Since $R(\mathbf{a})$ belongs to I, this means that we have to prove that $R(\mathbf{a} \notin I_0$. Suppose for contradiction that $R(\mathbf{a})$ belongs to I_0. Then $R(\mathbf{a})$ belongs to $I \cap I_0$. Recall that $I \cap I_0 \subseteq J$. Putting this together with $R(\mathbf{a} \in I \cap I_0$, we obtain that $R(\mathbf{a})$ belongs to J, which is a contradiction. $\square$

Our main result concerning consistent query answering w.r.t. a weakly acyclic set of tgds is a Π_2^p lower bound both for subset-repairs and for $\oplus$-repairs. A Π_2^p lower bound had been obtained earlier for the problem of finding the $\oplus$-consistent answers w.r.t. a set of functional dependencies and universal constraints [26].

THEOREM 6.3. *Let Σ be a weakly acyclic set of tgds and egds and let q be a conjunctive query.*

1. *The superset-consistent query answering problem for q w.r.t. Σ is in* PTIME.

2. *The $\oplus$-consistent (subset-consistent) query answering problem for q w.r.t. Σ is in Π_2^p.*
3. *There is a set Σ of weakly acyclic tgds and a conjunctive query q such that both the $\oplus$-consistent and the subset-consistent query answering problems for q w.r.t. Σ are Π_2^p-complete.*

PROOF (HINT). Part 1 is an an immediate consequence of Theorem 6.1. It also follows from Proposition 2.7 and results in [13] concerning the tractability of the certain answers of conjunctive queries in data exchange w.r.t. weakly acyclic sets of tgds and egds. Part 2 follows from Theorem 6.2 and the fact that both the $\oplus$-repair and the subset-repair checking problems with respect to any set of tgds and egds is in CONP. For Part 3, we give a weakly acyclic set Σ of tgds and a Boolean conjunctive query q with the following properties: for every quantified Boolean formula ϕ of the form $\forall p_1 \ldots \forall p_n \exists q_1 \ldots \exists q_k \psi$, where ψ is a conjunction of clauses containing 3 literals, we can construct in polynomial time an instance I_ϕ so that the following statements are equivalent:

1. ϕ is true ;
2. subset-Con$(q, I_\phi, \Sigma) = \top$;
3. $\oplus$-Con$(q, I_\phi, \Sigma) = \top$.

Evaluating such formulas is known to be a Π_2^p-complete problem [24]. We introduce the following tgds:

$$\begin{aligned}
\phi_1 &= Q(q,v) \rightarrow \exists s A(s),\\
\phi_2 &= A(s) \wedge Q'(q) \rightarrow \exists v Q(q,v),\\
\phi_3 &= Q(q,v) \wedge Q(q,v') \rightarrow E(v,v'),\\
\phi_4 &= P(p,v) \wedge P(p,v') \rightarrow E(v,v'),\\
\phi_{a_1a_2a_3}(X_1,X_2,X_3) &= R_{a_1a_2a_3}(x_1,x_2,x_3) \wedge A(s) \wedge\\
&\quad X_1'(x_1) \wedge X_2'(x_2) \wedge X_3'(x_3) \rightarrow\\
&\quad \exists v_1,v_2,v_3(X_1(x_1,v_1) \wedge X_2(x_2,v_2)\\
&\quad \wedge X_3(x_3,v_3) \wedge T_{a_1a_2a_3}(v_1,v_2,v_3)),
\end{aligned}$$

where $a_i \in \{0,1\}$ and $X_i \in \{P,Q\}$, for $i = 1,2,3$. We now let

$$\begin{aligned}
\Sigma &= \{\phi_1,\phi_2,\phi_3,\phi_4\} \cup\\
&\quad \{\phi_{a_1a_2a_3}(X_1,X_2,X_3) \mid a_i \in \{0,1\}, X_i \in \{P,Q\}\}.
\end{aligned}$$

We also let q be the query $\exists s A(s)$. Note that Σ is a weakly acyclic set. Indeed, the only special edges are from positions in Q', P' and R; however, there is no incoming edge to a position in Q', P' or R.

The intuition behind the relation symbols is as follows. The relation P' encodes the universally quantified variables $p_1,\ldots,p_n$ of ϕ, while Q' encodes the existentially quantified variables $q_1,\ldots,q_k$ of ϕ. Furthermore, P encodes truth values for $p_1,\ldots,p_n$, while Q encodes truth values for $q_1,\ldots,q_k$. We use the symbol E to simulate equality; it will consist of the tuples $(0,0)$ and $(1,1)$. The tgds ϕ_3 and ϕ_4 express that each variable gets assigned at most one truth value.

If x is a variable, we define the inverse sign of x, denoted by $is(x)$, as 0. If x is the negation of a variable, we define $is(x)$ as 1. The relation $R_{a_1a_2a_3}$ encodes the clauses of the form $x_1 \vee x_2 \vee x_3$ such that $is(x_i) = a_i$. The relation $T_{a_1a_2a_3}$ consists of those truth assignments that make true the clauses of the form $x_1 \vee x_2 \vee x_3$, where $is(x_i) = a_i$.

The symbol A acts as guard. It is activated (that is, becomes non-empty) as soon as one variable in $\{q_1,\ldots,q_k\}$ gets assigned a truth value (see the tgd ϕ_1). Once A has been activated, it makes sure that each variable in $\{q_1,\ldots,q_k\}$ is assigned a truth value (as expressed by the tgd ϕ_2). Hence, either no variable in $\{q_1,\ldots,q_k\}$ gets assigned a truth value or all variables in $\{q_1,\ldots,q_k\}$ get assigned a truth value. Moreover, when A is non-empty, the tgds of the form $\phi_{a_1a_2a_3}(X_1,X_2,X_3)$ express that the truth assignment is such that each clause of the form $x_1 \vee x_2 \vee x_3$ occurring in ϕ is true under the truth assignment. □

7. ARBITRARY TGDS (AND EGDS)

In this section, we obtain results about arbitrary sets of tgds that contrast sharply with our earlier results about weakly acyclic sets of tgds. We begin by pointing out that universal repairs are not of help in the study of arbitrary sets of tgds. First, as seen in Example 3.3, there is a (non-weakly acyclic) set of LAV tgds Σ and an instance I such that I has superset-repairs w.r.t. Σ (hence also $\oplus$-repairs w.r.t. Σ) of arbitrarily large sizes that cannot be bounded by any function in the size of the original instance. This contrasts with the state of affairs for weakly acyclic sets of tgds and egds in Theorem 6.2 Furthermore, recall Theorem 4.6, which states that if Σ is any set of LAV tgds, then every instance has an n-universal superset-repair w.r.t. Σ, for $n \geq 1$ (even though a universal superset-repair may not exist). We now observe that if Σ consists of arbitrary tgds, then even a 1-universal superset-repair may not exist.

PROPOSITION 7.1. *There is a set of tgds Σ and an instance I, such that I has no 1-universal superset-repair with respect to Σ.*

PROOF. Consider the following two tgds:

$$\begin{aligned}
P(x) &\rightarrow \exists y\, R(x,y) \wedge P(y)\\
R(x,y) \wedge R(y,z) \wedge R(z,u) \wedge R(u,v) &\rightarrow \exists w\, R(x,w) \wedge R(w,u)\\
R(x,y) \wedge R(y,x) &\rightarrow C_2(x)\\
R(x,y) \wedge R(y,z) \wedge R(z,x) &\rightarrow C_3(x)
\end{aligned}$$

Consider the instance $I = \{P(a)\}$. By the first tgd, every superset-repair of I must contain a cycle. By the second tgd, there must be a cycle of length 1, 2, or 3. Hence, by the final three tgds, each superset-repair satisfies either $\exists x\ C_2(x)$ or $\exists x\ C_3(x)$. However, there is a superset-repair that does not satisfy the query $\exists x\ C_2(x)$, and a superset-repair that does not satisfy the query $\exists x\, C_3(x)$. This shows that I has no 1-universal superset-repair. □

We present now the results concerning the consistent query answering problem. We begin with the subset-consistent answering and the superset-consistent answering problems.

THEOREM 7.2. *The following statements are true.*

1. *If Σ is a set of tgds and q is a conjunctive query, then the subset-consistent query answering problem for q w.r.t. Σ is in Π_2^p.*
2. *There is a set of tgds Σ and a conjunctive query q such that the subset-consistent query answering problem for q w.r.t. Σ is Π_2^p-complete.*
3. *There is a set Σ of tgds and a conjunctive query q such that the superset-consistent query answering problem for q w.r.t. Σ is undecidable.*

PROOF. We already mentioned the Π_2^p upper bound of subset-consistent query answering in Section 2; the lower bound was shown in Theorem 6.3, even for a weakly acyclic set of tgds.

As regards superset-repairs, it was shown in [22] that there is a (non-weakly acyclic) set Σ of tgds and egds such that the following problem is undecidable: given an instance I, is there an instance $J \models \Sigma$ such that $I \subseteq J$? Since every such instance J contains a superset-repair for I, we have that checking whether a given instance has a superset-repair with respect to Σ is an undecidable problem as well. By taking $q = \exists x\ P(x)$, where P is a fresh relation, it follows that the superset-consistent query answering problem is also undecidable. The egds of Σ can be eliminated by a standard construction: we add another binary relation E to

the schema and extend Σ with GAV tgds stating that E is a congruence, i.e., an equivalence relation such that every other relation in the schema is closed under substitution of equals by equals with respect to the equivalence relation E. Alternatively, a proof of this undecidability result can be obtained by adapting the proof of Theorem 7.3 given later on in this section. □

We now come to the main result, which asserts that there is a set Σ of tgds and a conjunctive query q such that computing the $\oplus$-consistent answers of q w.r.t. Σ is an undecidable problem. As mentioned in the Introduction, this improves a result in [3] asserting that there is a set of universal first-order sentences and a universal query (in fact, the negation of a conjunctive query) such that computing the $\oplus$-consistent answers is an undecidable problem.

THEOREM 7.3. *There is a set Σ of tgds and a conjunctive query q such that the $\oplus$-consistent query answering problem for q w.r.t. Σ is undecidable.*

The remainder of this section contains an outline of the proof of Theorem 7.3. We start with the following lemma. By a *Boolean combination of Boolean conjunctive queries*, we mean an expression built from Boolean conjunctive queries using $\neg$, $\vee$ and $\wedge$.

LEMMA 7.4. *Let Σ be a set of tgds and q a Boolean combination of Boolean conjunctive queries over a schema $\mathcal{S}$. There is a set Σ' of tgds and a conjunctive query q' over a schema $\mathcal{S}' \supseteq \mathcal{S}$ such that for every instance I over $\mathcal{S}$, we can compute in polynomial time an instance I' over $\mathcal{S}'$ satisfying*

$$\oplus\text{-Con}(q, I, \Sigma) = \oplus\text{-Con}(q', I', \Sigma').$$

Note that Lemma 7.4 remains true if we consider subset-repairs, instead of $\oplus$-repairs; however, it does not hold for superset-repairs.

Theorem 7.3 will be proved via a reduction from the halting problem for two-register machines. In describing two-register machines, we follow the definition given in [5].

A *two-register machine* (2RM) is similar to Turing machines, except that, instead of a tape, it has two registers r_1, r_2. Each register contains a natural number. A 2RM is programmed by a numbered sequence $\alpha_0, \ldots, \alpha_l$ of instructions. Each instruction α_i is either an addition or a subtraction. An addition has the form $+(rg, j)$, with rg a register number and $j \leq l$ an instruction number. Its semantics is: add one to register rg and move to instruction α_j. A subtraction has the form $-(rg, j, k)$ with rg a register number and $j, k \leq l$ instruction numbers. Its semantics is: if content of register rg is zero, then move to instruction α_j; otherwise, subtract one from register rg and move to instruction α_k.

An *instantaneous description* (ID) of a 2RM M is a triple (s, t, i) with $i \leq l$ an instruction number and $m, n \geq 0$ natural numbers representing the content of the registers r_1 and r_2. The unique *successor of an ID* (s, t, i) is the ID (s', t', i') such that

- if $\alpha_i = +(1, j)$, then $s' = s + 1$, $t' = t$, $i' = j$;
- if $\alpha_i = +(2, j)$, then $s' = s$, $t' = t + 1$, $i' = j$;
- if $\alpha_i = -(1, j, k)$ and $s \neq 0$, then $s' = s - 1$, $t' = t$, $i' = j$;
- if $\alpha_i = -(1, j, k)$ and $s = 0$, then $s' = s$, $t' = t$, $i' = k$;
- if $\alpha_i = -(2, j, k)$ and $t \neq 0$, then $s' = s$, $t' = t - 1$,$i' = j$;
- if $\alpha_i = -(2, j, k)$ and $t = 0$, then $s' = s$, $t' = t$, $i' = k$.

The ID $(0, 0, l)$ is called *final*. If $M = (\alpha_0, \ldots, \alpha_l)$ is a 2RM, then the *run* of M is the sequence $(D_i)_{i \geq 1}$ of ID's such that $D_0 = (0, 0, 0)$ and D_{i+1} is the successor of D_i, for all $i \geq 1$. The run is *halting* if it contains the ID $(0, 0, l)$. The *halting problem* for 2RM is to determine whether the run of a given 2RM is halting.

THEOREM 7.5. [7] *The halting problem for two-register machines is undecidable.*

We now proceed with the proof of Theorem 7.3.

PROOF OF THEOREM 7.3 (SKETCH). We shall give a reduction from the halting problem for 2RMs. For this, we define a set Σ' of tgds and a conjunctive query q' such that for every 2RM M, we can compute an instance I' such that

$$\text{the run of } M \text{ is halting} \iff \oplus\text{-Con}(q', I', \Sigma') \neq \top. \quad (1)$$

The idea is that each $\oplus$-repair of I^M encodes the run of a subset of the instructions in M (due to the symmetric difference semantics, we cannot prevent instructions from being dropped). The run of M will be halting if and only if some $\oplus$-repair of I^M contains the final ID of M. Hence, if q is a conjunctive query expressing that the final ID occurs, then the run of M is halting if and only if q is true in some $\oplus$-repair of I. That is, the run of M is halting if and only if $\oplus\text{-Con}(\neg q, I^M, \Sigma) \neq \top$. The problem is that $\neg q$ is not a conjunctive query, and this is where we use Lemma 7.4.

By Lemma 7.4, in order to find Σ' and q' satisfying (1), it suffices to find a set Σ of tgds and a Boolean combination q of Boolean conjunctive queries such that for all 2RM M, we can compute an instance I satisfying

$$\text{the run of } M \text{ is halting} \iff \oplus\text{-Con}(q, I, \Sigma) \neq \top. \quad (2)$$

Let Σ be the set of the following tgds:

$$\begin{aligned}
\phi_1^+ &= S(s,t,i) \wedge R_{1,+}(i,j) \rightarrow \exists s' \mathit{Succ}(s,s') \\
\psi_1^+ &= S(s,t,i) \wedge R_{1,+}(i,j) \wedge \mathit{Succ}(s,s') \rightarrow S(s',t,j) \\
\phi_2^+ &= S(s,t,i) \wedge R_{2,+}(i,j) \rightarrow \exists t' \mathit{Succ}(t,t') \\
\psi_2^+ &= S(s,t,i) \wedge R_{2,+}(i,j) \wedge \mathit{Succ}(t,t') \rightarrow S(s,t',j) \\
\phi_1^- &= S(s,t,i) \wedge R_{1,-}(i,j,k) \wedge \mathit{Succ}(s',s) \rightarrow S(s',t,j) \\
\phi_1^= &= S(s,t,i) \wedge R_{1,-}(i,j,k) \wedge \mathit{Zero}(s) \rightarrow S(s,t,k) \\
\phi_2^- &= S(s,t,i) \wedge R_{2,-}(i,j,k) \wedge \mathit{Succ}(t',t) \rightarrow S(s,t',j) \\
\phi_2^= &= S(s,t,i) \wedge R_{2,-}(i,j,k) \wedge \mathit{Zero}(t) \rightarrow S(s,t,k) \\
\phi_1^{\mathit{Anc}} &= \mathit{Anc}(u,s) \wedge \mathit{Succ}(s,s') \rightarrow \mathit{Anc}(u,s') \\
\phi_2^{\mathit{Anc}} &= \mathit{Succ}(s,s') \rightarrow \mathit{Anc}(s,s') \\
\phi_f &= \mathit{Final}(j) \wedge S(s,t,j) \wedge \mathit{Zero}(s) \wedge \mathit{Zero}(t) \rightarrow \exists s F(s)
\end{aligned}$$

Let q be the query $\exists s \mathit{Anc}(s,s) \vee \neg(\exists s F(s))$. The intuition behind the relation symbols is as follows. We use S to encode IDs: the fact $S(s,t,i)$ corresponds to the ID where s is the content of the first register, t is the content of the second register and i is an instruction number. We use the relation Succ to encode the successor relation on an initial segment of the natural numbers, while Anc encodes the strict natural ordering on that initial segment. The relation Anc is the transitive closure of Succ, as expressed by the tgds ϕ_1^{Anc} and ϕ_2^{Anc}. We simulate the initial and the final instructions of the machine with the relations Zero and Final. The relation Zero will consist of the singleton 0, while Final will consist of the singleton l (where $l + 1$ is the number of instructions of the machine).

The relations $R_{1,+}$, $R_{1,-}$, $R_{2,+}$ and $R_{2,-}$ encode the instructions of the machines. A fact $R_{n,+}(i,j)$ expresses that the instruction α_i is equal to $+(n,j)$. Similarly, a fact $R_{n,-}(i,j,k)$ expresses that the instruction α_i is the subtraction $-(n,j,k)$.

The tgd ϕ_1^+ expresses that if the ID is (s,t,i) and instruction α_i needs to add one to the first register, then the initial segment of the natural numbers that we consider should at least contain the successor of s. The tgd ψ_1^+ expresses that if the ID is (s,t,i), the instruction α_i is $+(1,j)$, and s' is equal to $s+1$, then we move to the ID (s',t,j). The tgds ϕ_2^+ and ψ_2^+ have similar meanings.

The tgd ϕ_1^- expresses that if the ID is (s, t, i), the instruction α_i is $-(1, j, k)$, and s' is equal to $s - 1$ (in particular, $s \neq 0$), then we move to the ID (s', t, j). The tgd $\phi_1^=$ expresses that if the ID is (s, t, i), the instruction α_i is $+(1, j)$ and s is equal to 0, then we move to the ID (s, t, k).

Finally, F acts as a guard. It becomes non-empty, when the instance contains an ID of the form $(0, 0, l)$. This is expressed by the tgd ϕ_f. The query q is false in an instance if and only if the guard F is non-empty and the ancestor relation does not contain any tuple of the form (s, s). Intuitively, this means that the instance contains an ID of the form $(0, 0, l)$ and that the successor relation does not contain any cycle. □

8. CONCLUDING REMARKS

In this paper, we carried out a fairly comprehensive investigation of the data complexity of the consistent query answering problem for conjunctive queries with respect to classes of constraints that have played a major role in data exchange and data integration. In the process, we brought into front stage and used extensively the notions of universal repairs and n-universal repairs, $n \geq 1$.

One problem left open by our investigation is the data complexity of the superset-consistent answers and of the $\oplus$-consistent answers for conjunctive queries w.r.t. sets of inclusion dependencies and equality-generating dependencies. As mentioned earlier, this problem has been shown to be undecidable in combined complexity [25]. It would also be interesting to investigate algorithmic aspects concerning the existence of universal repairs. Specifically, what can one say about the complexity of the following decision problem: given a set Σ of dependencies and an instance I, does I have a $\star$-universal repair w.r.t. Σ, where $\star \in \{\oplus$, subset, superset$\}$? And similarly for n-universal repairs, $n \geq 1$. Note that results in [21, 20] already imply that there is a set Σ of tgds for which the problem of checking whether a given instance has a universal superset-repair is undecidable.

Finally, the results presented give rise to challenging complete classification problems that, if resolved, may take the form of a dichotomy or a trichotomy theorems. For example, is it true that for every set Σ of GAV tgds and every conjunctive query q, the $\oplus$-consistent answers of q are either in PTIME or coNP-complete? Even for just key constraints only partial results towards such a dichotomy theorem have been obtained so far [16, 23, 28, 29].

Acknowledgments We thank the anonymous reviewers for their comments and pointers to the literature.

9. REFERENCES

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] F. N. Afrati and P. G. Kolaitis. Repair checking in inconsistent databases: algorithms and complexity. In *ICDT*, pages 31–41, 2009.

[3] M. Arenas and L. E. Bertossi. On the decidability of consistent query answering. In *AMW*, 2010.

[4] M. Arenas, L. E. Bertossi, and J. Chomicki. Consistent query answers in inconsistent databases. In *PODS*, pages 68–79, 1999.

[5] M. Benedikt, W. Fan, and F. Geerts. XPath satisfiability in the presence of dtds. *J. ACM*, 55(2), 2008.

[6] L. E. Bertossi. Consistent query answering in databases. *SIGMOD Record*, 35(2):68–76, 2006.

[7] E. Börger, E. Grädel, and Y. Gurevich. *The Classical Decision Problem*. Perspectives in Mathematical Logic. Springer, 1997.

[8] L. Bravo and L. E. Bertossi. Consistent query answering under inclusion dependencies. In *CASCON*, pages 202–216, 2004.

[9] A. Calì, D. Lembo, and R. Rosati. On the decidability and complexity of query answering over inconsistent and incomplete databases. In *PODS*, pages 260–271, 2003.

[10] J. Chomicki. Consistent query answering: Five easy pieces. In *ICDT*, pages 1–17, 2007.

[11] J. Chomicki and J. Marcinkowski. Minimal-change integrity maintenance using tuple deletions. *Inf. Comput.*, 197(1-2):90–121, 2005.

[12] A. Deutsch and V. Tannen. Reformulation of XML queries and constraints. In *ICDT*, pages 225–241, 2003.

[13] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theor. Comput. Sci.*, 336(1):89–124, 2005.

[14] R. Fagin, P. G. Kolaitis, and L. Popa. Data Exchange: Getting to the Core. *TODS*, 30(1):174–210, 2005.

[15] A. Fuxman, P. G. Kolaitis, R. J. Miller, and W. C. Tan. Peer data exchange. *ACM Trans. Database Syst.*, 31(4):1454–1498, 2006.

[16] A. Fuxman and R. J. Miller. First-order query rewriting for inconsistent databases. *J. Comput. Syst. Sci.*, 73(4):610–635, 2007.

[17] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[18] G. Gottlob and A. Nash. Efficient core computation in data exchange. *Journal of the ACM*, 55(2):1–49, 2008.

[19] G. Grahne and A. Onet. Data correspondence, exchange and repair. In *ICDT*, pages 219–230, 2010.

[20] A. Hernich. *Foundations of Query Answering in Relational Data Exchange*. PhD thesis, Goethe-Universität Frankfurt am Main, 2010.

[21] A. Hernich and N. Schweikardt. CWA-solutions for data exchange settings with target dependencies. In *PODS*, pages 113–122, 2007.

[22] P. G. Kolaitis, J. Panttaja, and W.-C. Tan. The complexity of data exchange. In *Proceedings of PODS'06*, pages 30–39, 2006.

[23] P. G. Kolaitis and E. Pema. A dichotomy in the complexity of consistent query answering for queries with two atoms. *Inf. Process. Lett.*, 112(3):77–85, 2012.

[24] C. M. Papadimitriou. *Computational complexity*. Addison-Wesley, Reading, Massachusetts, 1994.

[25] R. Rosati. On the finite controllability of conjunctive query answering in databases under open-world assumption. *J. Comput. Syst. Sci.*, 77(3):572–594, 2011.

[26] S. Staworko and J. Chomicki. Consistent query answers in the presence of universal constraints. *Information Systems*, 35(1):1–22, 2010.

[27] J. Wijsen. Database repairing using updates. *ACM Trans. Database Syst.*, 30:722–768, September 2005.

[28] J. Wijsen. On the first-order expressibility of computing certain answers to conjunctive queries over uncertain databases. In *PODS*, pages 179–190, 2010.

[29] J. Wijsen. A remark on the complexity of consistent conjunctive query answering under primary key violations. *Inf. Process. Lett.*, 110(21):950–955, 2010.