Characterizing Schema Mappings via Data Examples

Bogdan Alexe
UC Santa Cruz
abogdan@cs.ucsc.edu

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

Wang-Chiew Tan
UC Santa Cruz
wctan@cs.ucsc.edu

ABSTRACT

Schema mappings are high-level specifications that describe the relationship between two database schemas; they are considered to be the essential building blocks in data exchange and data integration, and have been the object of extensive research investigations. Since in real-life applications schema mappings can be quite complex, it is important to develop methods and tools for understanding, explaining, and refining schema mappings. A promising approach to this effect is to use "good" data examples that illustrate the schema mapping at hand.

We develop a foundation for the systematic investigation of data examples and obtain a number of results on both the capabilities and the limitations of data examples in explaining and understanding schema mappings. We focus on schema mappings specified by source-to-target tuple generating dependencies (s-t tgds) and investigate the following problem: which classes of s-t tgds can be "uniquely characterized" by a finite set of data examples? Our investigation begins by considering finite sets of positive and negative examples, which are arguably the most natural choice of data examples. However, we show that they are not powerful enough to yield interesting unique characterizations. We then consider finite sets of universal examples, where a universal example is a pair consisting of a source instance and a universal solution for that source instance. We unveil a tight connection between unique characterizations via universal examples and the existence of Armstrong bases (a relaxation of the classical notion of Armstrong databases). On the positive side, we show that every schema mapping specified by LAV s-t tgds is uniquely characterized by a finite set of universal examples with respect to the class of LAV s-t tgds. Moreover, this positive result extends to the much broader classes of n-modular schema mappings, n a positive integer. Finally, we show that, on the negative side, there are schema mappings specified by GAV s-t tgds that are not uniquely characterized by any finite set of universal examples and negative examples with respect to the class of GAV s-t tgds (hence also with respect to the class of all s-t tgds).

PODS'10, June 6–11, 2010, Indianapolis, Indiana, USA.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases – Data Translation; H.2.4 [**Database Management**]: Systems – Relational Databases

General Terms

Algorithms, Languages, Theory

Keywords

Schema mappings, data examples, data exchange, data integration

1. Introduction and Summary of Results

Schema mappings are high-level specifications that describe the relationship between two database schemas. Schema mappings are considered to be the essential building blocks in data exchange [18] and data integration [19] systems. The work on schema mappings to date has branched into two different (and, to a large extent, independent) directions of research. The first direction is concerned with the structural and algorithmic properties of schema mappings as given objects, and with their uses in the execution of data exchange and data integration tasks. The second direction is concerned with the discovery of schema mappings between two schemas, which is one of the first crucial steps taken towards the exchange or integration of data across database schemas.

Since real-life schemas can be complex, the discovery of a schema mapping between two schemas can be a difficult task. Consequently, several commercial systems, such as Altova Mapforce, Microsoft BizTalk Mapper, and Stylus Studio, as well as IBM's research prototype Clio [13, 15], have been developed to facilitate the task of producing a schema mapping between two schemas. All these systems adopt a common architecture towards the completion of this task: First, a user interface that displays both schemas is used to facilitate the derivation of a set of correspondences between attributes of relations of the two schemas. The set of correspondences is usually derived automatically or semi-automatically with the help of a schema-matching engine. After this, a schema-mapping generation component derives a schema mapping between the source schema and the target schema that is consistent with the set of correspondences. Typically, there are multiple schema mappings that are consistent with a set of correspondences between a source schema and a target schema, and most commercial systems produce just one of them. Actually, different commercial systems may produce semantically different schema mappings even when they are presented with the same set of correspondences between a source and a target schema [3]. In research prototypes, such as Clio, multiple schema mappings that are consistent with the set of

correspondences are generated, and one of them is designated as the default schema mapping.

Regardless of whether just a single schema mapping or multiple schema mappings may be derived by these systems, there is a clear and pressing need to illustrate the exact semantics of the schema mappings that are generated. In fact, in real-life applications, schema mappings can be quite complex even if they are derived manually. A promising approach to this effect is to use "good" data examples that illustrate the schema mapping at hand. This approach is motivated from the long and venerable tradition of using test examples in understanding and debugging computer programs. The use of data examples for schema mappings was advocated in [26], where the concept of a mapping example was introduced together with a set of operators for manipulating such examples. More recently, data examples were used in [2] to aid designers in refining various aspects of schema-mapping specifications; furthermore, in [5], the "behavior" of schema mappings was illustrated in the form of routes from source to target data. Beyond schema mappings, the problem of generating "illustrative" examples for dataflow programs was recently investigated in [22].

In this paper, we develop a foundation for the systematic investigation of data examples for schema mappings and obtain a number of results that shed light on both the capabilities and the limitations of data examples in explaining and understanding schema mappings. We focus on schema mappings specified by source-to-target tuple generating dependencies (s-t tgds, in short), also known as GLAV (global-and-local-as-view) dependencies; this class of schema mappings comprises the most extensively studied schema mappings to date and contains, as important special cases, the classes of schema mappings specified by LAV (local-as-view) dependencies and by GAV (global-as-view) dependencies.

Let $\mathbf{S}$ be a source schema and $\mathbf{T}$ be a target schema. We consider schema mappings $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a finite set of s-t tgds. A data example is a pair (I, J) such that I is a source instance and J is a target instance. The central notion in our investigation is what it means to say that a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is uniquely characterized by a finite set $\mathcal{F}$ of data examples with respect to (w.r.t.) a class $\mathcal{C}$ of s-t tgds of interest. Informally, $\mathcal{M}$ is *uniquely characterized by $\mathcal{F}$ w.r.t. $\mathcal{C}$* if Σ is, up to logical equivalence, the only finite set Σ' of s-t tgds from $\mathcal{C}$ such that each data example in $\mathcal{F}$ has the "same relationship" with Σ as it has with Σ'. This concept is formalized by making precise the notion of the "relationship" between a data example and a set of s-t tgds. As we shall see, this notion can be made precise in different natural ways; furthermore, the different notions obtained give rise to different results concerning unique characterizations of schema mappings.

Our investigation of unique characterizations begins by considering finite sets $\mathcal{F}$ of data examples that are *positive examples* or *negative examples*, where a data example (I, J) is a positive example for a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ if $(I, J) \models \Sigma$, and it is a negative example if $(I, J) \not\models \Sigma$. Positive and negative examples are arguably the most natural types of data examples to consider; in fact, these types of examples are the main objects of study in the context of computational learning (e.g., see [17]). We show that if the source schema $\mathbf{S}$ and the target schema $\mathbf{T}$ contain only unary relation symbols, then every schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a finite set of s-t tgds, can be uniquely characterized by a finite set of positive and negative examples w.r.t. the class of all s-t tgds. This result appears to be a promising first step, but, unfortunately, it does not extend to schema mappings over source and target schemas that contain non-unary relation symbols. Indeed, we exhibit a LAV schema mapping over a source schema with one binary relation symbol and a target schema with one binary relation symbol that is *not* uniquely characterizable by any finite set of positive and negative examples w.r.t. the class of LAV s-t tgds (hence, also w.r.t. the class of all s-t tgds). Furthermore, we exhibit a GAV schema mapping for which a similar state of affairs holds with respect to the class of all GAV s-t tgds (hence, also w.r.t. the class of all s-t tgds).

In view of the failure of positive and negative examples to yield unique characterizations beyond the very limited case of schemas consisting of unary relation symbols only, we consider the notion of a *universal example*, where a data example (I, J) is a universal example for a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ if J is a universal solution for I with respect to $\mathcal{M}$. Universal solutions were introduced in [9] and shown to be the preferred solutions to materialize in data exchange because, among other reasons, they are the most general solutions and they represent (in a precise technical sense) the entire space of solutions for a given source instance. These properties of universal solutions suggest that universal examples are indeed a natural type of data example to consider as candidates for unique characterizations of schema mappings.

Before delineating the capabilities and limitations of universal examples, we unveil a very tight (and unexpected) connection between the existence of unique characterizations of schema mappings $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ via universal examples and the existence of an *Armstrong basis for* Σ, which is a relaxation of the classical notion of an Armstrong database for Σ. As is well known, an Armstrong database for Σ w.r.t. a class $\mathcal{C}$ of database dependencies is a database D that satisfies all the dependencies in $\mathcal{C}$ that are logical consequences of Σ, and no other dependencies in $\mathcal{C}$. Armstrong databases were extensively studied in the context of database dependency theory in the 1970s and 1980s (see [7] for a survey). Clearly, if Σ and Σ' are two sets of s-t tgds and if (I, J) is an Armstrong database for both Σ and Σ' w.r.t. a class $\mathcal{C}$ of s-t tgds containing Σ and Σ', then Σ is logically equivalent to Σ'. Thus, Armstrong databases are ideal data examples for unique characterizations of schema mappings. Nevertheless, it is rare that a schema mapping specified by s-t tgds possesses an Armstrong database. For this reason, we introduce and study the following relaxation of the notion of an Armstrong database. Let Σ be a set of s-t tgds and let $\mathcal{D} = \{(I_1, J_1), \ldots, (I_n, J_n)\}$ be a finite set of data examples. We say that $\mathcal{D}$ is an *Armstrong basis for Σ w.r.t. a class $\mathcal{C}$ of s-t tgds* if for every s-t tgd σ in $\mathcal{C}$, we have that Σ logically implies σ if and only if $(I_i, J_i) \models \sigma$, for every $i = 1, \ldots, n$. This is a strict relaxation of the notion of an Armstrong database because we show that there are LAV s-t tgds that have an Armstrong basis w.r.t. the class of all LAV s-t tgds, but not an Armstrong database. Also, it is quite easy to see that if $\mathcal{D}$ is an Armstrong basis for both Σ and Σ' w.r.t. a class $\mathcal{C}$ containing Σ and Σ', then Σ is logically equivalent to Σ'. Thus, when they exist, Armstrong bases readily yield unique characterizations of schema mappings. We show that a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a set of s-t tgds, is uniquely characterized by a finite set of universal examples w.r.t. a class $\mathcal{C}$ of s-t tgds containing Σ if and only if Σ possesses an Armstrong basis w.r.t. to $\mathcal{C}$. This result reinforces the "goodness" of universal examples and, at the same time, reveals an a priori unexpected connection between (a natural relaxation of) a key notion in database dependency theory and a key notion in data exchange.

The following question naturally arises. Which classes of schema mappings specified by s-t tgds possess unique characterizations via universal examples? Equivalently, which schema mappings specified by s-t tgds possess Armstrong bases? On the positive side, we show that every schema mapping specified by LAV s-t tgds is uniquely characterized by a finite set of universal examples w.r.t. the class of LAV s-t tgds. We then extend this positive result to

the class of n-modular schema mappings $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, $n \geq 1$, where $\mathcal{M}$ is n-modular if whenever (I, J) is a negative example for Σ, then there is a sub-instance I' of I of size at most n such that (I', J) is also a negative example for Σ. The notion of n-modularity was introduced in [24] and used to characterize schema-mapping languages in terms of their structural properties. Finally, on the negative side, we show that there are natural schema mappings specified by GAV s-t tgds that are *not* uniquely characterized by any finite set of universal examples and negative examples w.r.t. the class of GAV s-t tgds. The proof of this theorem makes use of sophisticated results from graph theory, namely, a generalization of Erdös' celebrated result [6] asserting the existence of graphs of arbitrarily large girth and chromatic number.

2. Preliminaries

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. We shall often write R_i to denote both the relation symbol and the relation R_i^I that interprets it. An *atom* (over $\mathbf{R}$) is a formula $P(x_1, \ldots, x_m)$, where P is a relation symbol in $\mathbf{R}$ and $x_1, \ldots, x_m$ are variables, not necessarily distinct. A *fact* of an instance I (over $\mathbf{R}$) is an expression $P^I(v_1, \ldots, v_m)$, where P is a relation symbol in $\mathbf{R}$ and $v_1, \ldots, v_m$ are values such that $(v_1, \ldots, v_m) \in P^I$. We assume that all instances I considered are finite, which means that every relation R_i^I is finite, for $1 \leq i \leq k$.

Schema Mappings. A *schema mapping* is a triple $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ consisting of a source schema $\mathbf{S}$, a target schema $\mathbf{T}$, and a set Σ of constraints. We say that $\mathcal{M}$ is *specified by* Σ. In general, the constraints in Σ are formulas in some logical formalism. Here, we will focus on schema mappings specified by source-to-target tuple-generating dependencies.

A *source-to-target tuple-generating dependency (s-t tgd)* is a first-order sentence φ of the form

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

where $\varphi(\mathbf{x})$ is a conjunction of atoms over $\mathbf{S}$, each variable in $\mathbf{x}$ occurs in at least one atom in $\varphi(\mathbf{x})$, and $\psi(\mathbf{x}, \mathbf{y})$ is a conjunction of atoms over $\mathbf{T}$ with variables in $\mathbf{x}$ and $\mathbf{y}$. For simplicity, we will often drop the universal quantifiers $\forall \mathbf{x}$ in the above formula. Another name for s-t tgds is *global-and-local-as-view* (GLAV) constraints (see [19]). They contain GAV and LAV constraints as important special cases.

A GAV *(global-as-view)* constraint is a s-t tgd in which the right-hand side is a single atom, i.e., it is of the form

$$\forall \mathbf{x}(\varphi(\mathbf{x}) \rightarrow P(\mathbf{x})),$$

where $P(\mathbf{x})$ is an atom over the target schema. A LAV *(local-as-view)* constraint is a s-t tgd in which the left-hand side is a single atom, i.e., it is of the form

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

where $Q(\mathbf{x})$ is an atom over the source schema.

Satisfaction and Logical Implication. The symbol $\models$ will be used to denote several different notions. If Σ is a set of first-order sentences and D is an instance, then $D \models \Sigma$ means that D satisfies every sentence in Σ. If σ is a first-order sentence, then $\Sigma \models \sigma$ denotes *logical implication*, i.e., it means that for every (finite or infinite) instance D such that $D \models \Sigma$, we have that $D \models \sigma$. If Σ' is a set of first-order sentences, then $\Sigma \models \Sigma'$ means that for every $\sigma' \in \Sigma'$, we have that $\Sigma \models \sigma'$. Finally, $\Sigma \equiv \Sigma'$ denotes that Σ and Σ' are *logically equivalent*, i.e., $\Sigma \models \Sigma'$ and $\Sigma' \models \Sigma$.

The notion of logical implication is defined using all (finite and infinite) instances. There is a companion notion of *logical implication in the finite*, denoted by $\models_{\mathrm{Fin}}$, where $\Sigma \models_{\mathrm{Fin}} \sigma$ means that for every finite instance D such that $D \models \Sigma$, we have that $D \models \sigma$. In general, $\models$ and $\models_{\mathrm{Fin}}$ are different notions (clearly, if $\Sigma \models \sigma$, then $\Sigma \models_{\mathrm{Fin}} \sigma$, but the converse need not be true). It is easy to see, however, that these two notions coincide on finite sets of s-t tgds. Specifically, assume that Σ is a finite set of s-t tgds and σ is a s-t tgd. Then $\Sigma \models \sigma$ if and only if $\Sigma \models_{\mathrm{Fin}} \sigma$. For the non-trivial direction, assume that $\Sigma \models_{\mathrm{Fin}} \sigma$ but $\Sigma \not\models \sigma$. Let (I, J) be such that $(I, J) \models \Sigma$, but $(I, J) \not\models \sigma$. Assume that σ is $\varphi(\mathbf{x}) \rightarrow \exists \mathbf{z} \psi(\mathbf{x}, \mathbf{z})$. Then there is a tuple $\mathbf{a}$ such that $I \models \varphi(\mathbf{a})$ and $J \models \forall \mathbf{z} \neg \psi(\mathbf{a}, \mathbf{z})$. Let I_0 be the sub-instance of I consisting of the facts $\varphi(\mathbf{a})$. Assume that Σ consists of the s-t tgds $\sigma_1, \ldots, \sigma_k$. Then there are finite sub-instances J_i of J such that $(I_0, J_i) \models \sigma_i$, $i = 1, \ldots, k$. Let J_0 be the union of all J_i, $1 \leq i \leq k$. Then (I_0, J_0) is a finite instance that satisfies Σ but not σ, which is a contradiction.

Solutions, Homomorphisms, and Universal Solutions. We now review some basic notions and results from [9]. We assume that we have a fixed infinite set Const of constants and a fixed infinite set Var of nulls that is disjoint from Const. We write $\mathrm{adom}(I)$ for the *active domain* of an instance I, that is, the set of all values occurring in I. All values occurring in a source instance I are assumed to be constants, i.e., $\mathrm{adom}(I) \subseteq$ Const. In contrast, target instances have values in Const $\cup$ Var. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping. If I is a source instance, then a *solution for I w.r.t. $\mathcal{M}$* is a target instance J such that $(I, J) \models \Sigma$. From a semantic point of view, a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ can be identified with the collection $\{(I, J) : I$ is a source instance and J is a solution for $I\}$.

Assume that K, K' are two instances over the target schema $\mathbf{T}$. A function h from Const$\cup$Var to Const$\cup$Var is a *homomorphism* from K to K' if for every $c \in$ Const, we have that $h(c) = c$, and for every relation symbol R in $\mathbf{T}$ and every tuple $(a_1, \ldots, a_n) \in R^K$, we have that $(h(a_1), \ldots, h(a_n)) \in R^{K'}$. We write $K \rightarrow K'$ to denote that there is a homomorphism from K to K'. The instances K and K' are said to be *homomorphically equivalent* if $K \rightarrow K'$ and $K' \rightarrow K$.

Given a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ and a source instance I, a *universal solution for I w.r.t. $\mathcal{M}$* is a solution J for I w.r.t. $\mathcal{M}$ such that for every solution J' for I w.r.t. $\mathcal{M}$, we have that $J \rightarrow J'$. Intuitively, universal solutions are the "most general" solutions among all solutions for I, hence the preferred solutions to materialize in data exchange. Clearly, if both J_1 and J_2 are universal solutions for I, then J_1 and J_2 are homomorphically equivalent.

Chase. The *chase procedure* is an algorithm that was originally designed to reason about database dependencies (see [1]), but it turned out to have numerous applications to data exchange and data integration. In particular, as shown in [9], if $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping specified by s-t tgds, then the chase procedure can be used to produce, given a source instance I, a universal solution $\mathit{chase}_{\mathcal{M}}(I)$ for I in time bounded by a polynomial in the size of I.

There are several variants of the chase procedure. Here, we will consider the simplest such variant, called the *naive* chase. Given a source instance I, the naive chase produces a universal solution $\mathit{chase}_{\mathcal{M}}(I)$ for I as follows. For every s-t tgd

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

in Σ and for every tuple $\mathbf{a}$ of constants from $\mathrm{adom}(I)$ such that $I \models \varphi(\mathbf{a})$, we add to $\mathit{chase}_{\mathcal{M}}(I)$ all facts in $\psi(\mathbf{a}, \mathbf{b})$, where $\mathbf{b}$ is a tuple of new nulls interpreting the existential quantified variables $\mathbf{y}$. Thus, nulls are created independently each time and without considering whether the right-hand side of the s-t tgd at hand could be satisfied using facts that involve nulls created earlier.

3. Positive and Negative Examples

Let $\mathbf{S}$ be a source schema and $\mathbf{T}$ a target schema. A *data example* is a pair (I, J) such that I is a source instance and J is a target instance. Assume now that $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping, where Σ is a finite set of s-t tgds. This is a finite syntactic description of a schema mapping. As mentioned in Section 2, from a semantic point of view, $\mathcal{M}$ can be identified with the infinite collection $\{(I, J) : (I, J) \models \Sigma\}$. Our main goal in this paper is to address the following question: can this infinite collection of data examples be "captured" by a finite set of data examples. In other words, does $\mathcal{M}$ have a finite semantic description in terms of data examples. We make this question precise by considering different "types" of data examples and stipulating that a finite set $\mathcal{F}$ of examples *uniquely characterizes* $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ *w.r.t. a class* $\mathcal{C}$ *of s-t tgds* if the following holds: for every finite set Σ' of s-t tgds from $\mathcal{C}$ such that each example in $\mathcal{F}$ has the same "type" w.r.t. Σ as it has w.r.t. Σ', we have that $\Sigma \equiv \Sigma'$. It should be noted that, in addition to the concept of logical equivalence ($\equiv$), two other notions of equivalence between schema mappings have been considered, namely *data-exchange equivalence* and *conjunctive-query equivalence* [10]. In general, these three notions of equivalence are distinct; however, they are known to coincide for s-t tgds [10]. Thus, the preceding concept of unique characterization of a schema mapping amounts to asserting that for every set Σ' of s-t tgds from $\mathcal{C}$ such that each example in $\mathcal{F}$ has the same "type" w.r.t. Σ as it has w.r.t. Σ', we have that Σ is data-exchange equivalent or conjunctive-query equivalent to Σ'.

We begin by considering positive and negative examples, two natural types of examples that have been widely used in computational learning [17].

DEFINITION 3.1. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping.

A *positive example* for $\mathcal{M}$ is a data example (I, J) such that $(I, J) \models \Sigma$.

A *negative example* for $\mathcal{M}$ is a data example (I, J) such that $(I, J) \not\models \Sigma$. □

DEFINITION 3.2. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping, let $\mathcal{P}$ and $\mathcal{N}$ be two finite sets of positive and, respectively, negative examples for $\mathcal{M}$, and let $\mathcal{C}$ be a class of s-t tgds.

We say that $\mathcal{M}$ is *uniquely characterized by* $\mathcal{P}$ *and* $\mathcal{N}$ *w.r.t.* $\mathcal{C}$ if for every finite set $\Sigma' \subseteq \mathcal{C}$ such that $\mathcal{P}$ and $\mathcal{N}$ are sets of positive and, respectively, negative examples for the schema mapping $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$, we have that Σ is logically equivalent to Σ' (in symbols, $\Sigma \equiv \Sigma'$). □

3.1 Unary Schemas and Unique Characterizations

A schema $\mathbf{R} = (R_1, \ldots, R_k)$ is said to be *unary* if every relation symbol R_i in $\mathbf{R}$ is unary (has arity 1). In this section, we show that if both the source and the target schemas are unary, then every schema mapping specified by a finite set of s-t tgds can be uniquely characterized by finite sets of positive and negative examples w.r.t. the class of all s-t tgds. The proof makes essential use of the following lemma.

LEMMA 3.3. *Let* $\mathbf{S}$ *be a unary source schema and* $\mathbf{T}$ *a unary target schema. Then, up to logical equivalence, there are finitely many schema mappings* $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ *such that* Σ *is a finite set of s-t tgds.*

PROOF. *(Hint)* Assume that $\mathbf{S}$ and $\mathbf{T}$ are unary schemas. We show that every finite set Σ of s-t tgds over $\mathbf{S}$ and $\mathbf{T}$ is logically equivalent to a finite set of s-t tgds in a certain *canonical form*, and that there are finitely many finite sets of s-t tgds in canonical form. This canonical form is defined as follows. We say that a s-t tgd is in *canonical form* if either it is a GAV s-t tgd of the form

$$\phi_1(x_1) \wedge \ldots \wedge \phi_k(x_k) \rightarrow T(x_j),$$

or a (non-GAV) s-t tgd of the form

$$\phi_1(x_1) \wedge \ldots \wedge \phi_k(x_k) \rightarrow \exists y\, \psi(y),$$

where (a) each formula $\phi_i(x_i)$ is a conjunction of distinct source relational atoms that share the same variable x_i; (b) if $i \neq l$, then the set of relational symbols in $\phi_i(x_i)$ is different from the set of relational symbols in $\phi_l(x_l)$; (c) T is a relation symbol in $\mathbf{T}$; and (d) $\psi(y)$ is a conjunction of distinct atoms over $\mathbf{T}$ that share the same variable y. Since $\mathbf{S}$ and $\mathbf{T}$ are unary schemas, it is easy to see that there are finitely many finite sets of s-t tgds in canonical form. Furthermore, using suitable rewrite rules, it can be shown that every finite set Σ of s-t tgds is logically equivalent to a finite set of s-t tgds in canonical form. □

In effect, the proof of the preceding Lemma 3.3 has the flavor of a quantifier-elimination result about the class of s-t tgds over unary source and target schemas. It is well known that first-order logic over unary schemas admits quantifier elimination (for example, see [16, page 66]). However, Lemma 3.3 is a more refined result that cannot be derived (at least in a straightforward way) from the quantifier-elimination result for first-order logic over unary schemas.

THEOREM 3.4. *Let* $\mathbf{S}$ *be a unary source schema and* $\mathbf{T}$ *a unary target schema. If* $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ *is a schema mapping such that* Σ *is a finite set of s-t tgds, then* $\mathcal{M}$ *can be uniquely characterized by finite sets of positive and negative examples with respect to the class of all s-t tgds.*

PROOF. By Lemma 3.3, there are, up to logical equivalence, finitely many schema mappings specified by finite sets of s-t tgds. Consequently, there are finitely many schema mappings $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ specified by a finite set of s-t tgds such that $\Sigma \not\equiv \Sigma'$. Let $\mathcal{M}_1 = (\mathbf{S}, \mathbf{T}, \Sigma_1), \ldots, \mathcal{M}_k = (\mathbf{S}, \mathbf{T}, \Sigma_k)$ be an exhaustive list of all (up to logical equivalence) such schema mappings. Therefore, for each $i \leq k$, we have that $\Sigma \not\models \Sigma_i$ or $\Sigma_i \not\models \Sigma$. We construct a finite set $\mathcal{P}$ of positive examples for $\mathcal{M}$ and a finite set $\mathcal{N}$ of negative examples for $\mathcal{M}$ as follows. Initially, both $\mathcal{P}$ and $\mathcal{N}$ are empty. For each $\mathcal{M}_i$, $i \leq k$, there are two cases to consider.

Case 1. $\Sigma \not\models \Sigma_i$. In this case, let (I, J) be a data example such that $(I, J) \models \Sigma$ and $(I, J) \not\models \Sigma_i$. We add (I, J) to $\mathcal{P}$.

Case 2. $\Sigma_i \not\models \Sigma$. In this case, let (I, J) be a data example such that $(I, J) \models \Sigma_i$ and $(I, J) \not\models \Sigma$. We add (I, J) to $\mathcal{N}$.

By construction, $\mathcal{P}$ is a finite set of positive examples for $\mathcal{M}$, and $\mathcal{N}$ is a finite set of negative examples for $\mathcal{M}$. Moreover, it easy to verify that the sets $\mathcal{P}$ and $\mathcal{N}$ uniquely characterize $\mathcal{M}$ w.r.t. the class of all s-t tgds. □

3.2 Limitations of Positive and Negative Examples

The preceding Theorem 3.4 does not extend to schema mappings over source and target schemas that contain non-unary relation symbols. As a matter of fact, if the source and the target schema contain binary relation symbols, then there is a schema mapping $\mathcal{M}$ specified by a LAV s-t tgd such that *no* finite sets of positive and negative examples uniquely characterize $\mathcal{M}$ w.r.t. to the class of all LAV s-t tgds. Furthermore, a similar result holds for GAV s-t tgds.

THEOREM 3.5. *Let* $\mathbf{S}$ *be a source schema consisting of a single binary relation symbol* P *and let* $\mathbf{T}$ *be a target schema consisting of a single binary relation symbol* P'.

1. *There is a schema mapping* $\mathcal{M}$ *specified by a single LAV s-t tgd such that* $\mathcal{M}$ *is not uniquely characterizable by any finite sets of positive and negative examples with respect to the class of all LAV s-t tgds.*
2. *There is a schema mapping* $\mathcal{M}'$ *specified by a single GAV s-t tgd such that* $\mathcal{M}'$ *is not uniquely characterizable by any finite sets of positive and negative examples with respect to the class of all GAV s-t tgds.*

PROOF. For the first part, let σ be the LAV s-t tgd

$$P(x, y) \rightarrow \exists z P'(z, z).$$

Assume that $\mathcal{P}$ and $\mathcal{N}$ are finite sets of positive and, respectively, for $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \{\sigma\})$. We will show that there is a LAV s-t tgd σ' such that $\sigma \not\equiv \sigma'$, yet $\mathcal{P}$ and $\mathcal{N}$ are sets of positive and, respectively, negative examples for σ'. Let n be a positive integer bigger than the maximum size of the active domains of the data examples in $\mathcal{N}$, and let σ' be the s-t tgd

$$P(x, y) \rightarrow \exists x_1 ... \exists x_n K_n,$$

where K_n asserts that $x_1, \ldots, x_n$ form an n-clique in P'. To see that $\sigma \not\equiv \sigma'$, let (I, J) be the data example such that $I = \{(1, 2)\}$ and J is a n-clique. Then $(I, J) \models \sigma'$, but $(I, J) \not\models \sigma$, hence $\sigma' \not\models \sigma$.

We now show that every $(I, J) \in \mathcal{P}$ satisfies σ'. Suppose $I \models P(a, b)$ for some (not necessarily distinct) values a and b. Since $(I, J) \models \sigma$, it follows that J must contain a fact $P'(c, c)$ for some value c. Hence, $J \models \exists x_1 ... \exists x_n K_n$ (by mapping every variable x_i, $1 \leq i \leq n$, to c). Next, we show that every member of $\mathcal{N}$ is a negative example for σ'. If $(I, J) \in \mathcal{N}$, then, $(I, J) \not\models \sigma$, there are values a and b such that $I \models P(a, b)$. Towards a contradiction, assume that there is a homomorphism $h : \exists x_1 ... \exists x_n K_n \rightarrow J$. Since n is greater than the number of distinct values in J, there must be two variables x_i and $x_j$such that $i \neq j$ and $h(x_i) = h(x_j)$. Hence, $P'(h(x_i), h(x_j)) \in J$ and so the mapping g, where $g(z) = h(x_i)$, is a homomorphism from $\exists z P'(z, z)$ to J. Thus, $(I, J) \models \sigma$, which is a contradiction.

The second part of this theorem follows from the proof of Theorem 6.4. □

4. Universal Examples and Armstrong Bases

The limitations of positive and negative examples suggest that a stronger type of data example should be considered. In this section, we introduce *universal examples* and show them to be intimately connected with *Armstrong bases*, a relaxation of the classical notion of an Armstrong database studied in the context of dependency theory a long time ago.

4.1 Universal Examples

DEFINITION 4.1. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping in which Σ is a finite set of s-t tgds. A data example (I, J) is a *universal example for* $\mathcal{M}$ if J is a universal solution for I w.r.t. $\mathcal{M}$. □

As discussed in Section 2, universal solutions are the preferred solutions to materialize in data exchange because (by way of having homomorphisms to every solution) they are the "most general" solutions. Furthermore, as shown in [9], universal solutions represent the entire space of solutions in the following sense. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping in which Σ is a set of s-t tgds, and let (I_1, J_1) and (I_2, J_2) be two universal examples for $\mathcal{M}$. Then the space of solutions for I_1 coincides with the space of solutions for I_2 if and only if J_1 and J_2 are homomorphically equivalent. These properties motivate universal examples as candidates for unique characterizations of schema mappings.

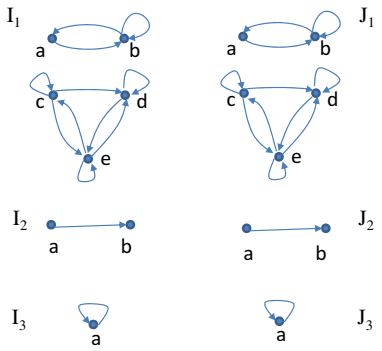

Figure 1: Three universal examples that uniquely characterize $E(x, y) \rightarrow F(x, y)$ w.r.t. GAV s-t tgds.

DEFINITION 4.2. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping in which Σ is a finite set of s-t tgds, let $\mathcal{U}$ be a finite set of universal examples for $\mathcal{M}$, and let $\mathcal{C}$ be a class of s-t tgds.

We say that $\mathcal{M}$ is *uniquely characterized by* $\mathcal{U}$ *w.r.t.* $\mathcal{C}$ if for every finite set $\Sigma' \subseteq \mathcal{C}$ such that $\mathcal{U}$ is a set of universal examples for the schema mapping $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$, we have that $\Sigma \equiv \Sigma'$. □

Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \{\sigma\})$ be the schema mapping where σ is the LAV s-t tgd $P(x, y) \rightarrow \exists z P'(z, z)$. As seen in the proof of Theorem 3.5, no finite sets of positive and negative examples uniquely characterize $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds. In contrast, there is a finite set of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds. Indeed, it is not hard to verify that the set $\{(I_1, J_1), (I_2, J_2)\}$ has this property, where $I_1 = \{P(a, b)\}$, $J_1 = \{P'(N, N)\}$, $I_2 = \{P(a, a)\}$, $J_2 = \{P'(N, N)\}$, and N is a null. Thus, universal examples go beyond what positive and negative examples can offer. Later on, however, we will see that universal examples have their own limitations. For now, we illustrate further the capabilities of universal examples by establishing unique characterizations for the binary *copy* s-t tgd, which is both a LAV and a GAV s-t tgd.

PROPOSITION 4.3. *Let* $\mathcal{M}$ *be the schema mapping specified by the binary copy s-t tgd* $E(x, y) \rightarrow F(x, y)$.

1. $\mathcal{M}$ *is uniquely characterizable by a finite set of universal examples w.r.t. the class of all LAV s-t tgds.*
2. $\mathcal{M}$ *is uniquely characterizable by a finite set of universal examples w.r.t. the class of all GAV s-t tgds.*

PROOF. *(Sketch)* For the first part, it can be shown that the set consisting of the universal examples $\{(I_1, J_1), (I_2, J_2)\}$, where $I_1 = \{E(a, b)\}$, $J_1 = \{F(a, b)\}$, $I_2 = \{E(a, a)\}$, $J_2 = \{F(a, a)\}$, uniquely characterizes $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds. Actually, as we will see later on, this will also follow from a general result to the effect that every schema mapping specified by a finite set of LAV s-t tgds is uniquely characterized by a finite set of universal examples w.r.t. the class of all LAV s-t tgds.

For the second part, let $\mathcal{U}$ be the set consisting of the three universal examples (I_1, J_1), (I_2, J_2), (I_3, J_3) depicted in Figure 1. With some work, it can be shown that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$

w.r.t. the class of all GAV s-t tgds. A detailed proof will be given in the full version of the paper; here we limit ourselves into providing an informal explanation. Let $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ be a schema mapping such that Σ' is a finite set of GAV s-t tgds and the examples in Figure 1 are universal examples for $\mathcal{M}'$. The first example (I_1, J_1) is used to show that $\Sigma \models \Sigma'$. Indeed, if $\Sigma \not\models \Sigma'$, then one can easily show that J_1 is not a solution for I_1 w.r.t. $\mathcal{M}'$. The remaining two examples (I_2, J_2) and (I_3, J_3) are used to show that $\Sigma' \models \Sigma$; this is based on the observation that every source fact that is copied over to the target by the copy tgd is isomorphic to I_2 or I_3. □

The results in the preceding Proposition 4.3 inevitably raise the question as to whether or not the schema mapping $\mathcal{M}$ specified by the binary copy s-t tgd can also be uniquely characterized via universal examples w.r.t. the class of *all* s-t tgds. In Section 6, we will show that such a unique characterization is not true for $\mathcal{M}$.

4.2 Armstrong Databases and Armstrong Bases

Database dependencies are integrity constraints, typically expressed as formulas in some fragment of first-order logic. The study of database dependencies was the focus of extensive research activity during the 1970s and the early 1980s (see [12] for a survey). A central problem in this area is the *implication problem* for dependencies, which is the problem of determining whether or not a given finite set of dependencies logically implies another given dependency. Armstrong databases turned out to be a useful tool in attacking this problem; they were introduced explicitly and studied in their own right by Fagin [8], but, in the case of functional dependencies, were implicit in Armstrong's earlier work [4].

DEFINITION 4.4. Let Σ and $\mathcal{C}$ be two sets of database dependencies over the same schema.

An *Armstrong database for Σ w.r.t. $\mathcal{C}$* is an instance D such that for every $\sigma \in \mathcal{C}$, we have $\Sigma \models \sigma$ if and only if $D \models \sigma$. In other words, an Armstrong database for Σ w.r.t. $\mathcal{C}$ is an instance that satisfies all the dependencies in $\mathcal{C}$ that are logically implied by Σ, and no other dependencies in $\mathcal{C}$.

An *Armstrong database for a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ w.r.t. $\mathcal{C}$* is an Armstrong database for Σ w.r.t. $\mathcal{C}$. □

A moment's reflection tells that Armstrong databases give rise to a new type of data examples for unique characterizations of schema mappings. Indeed, let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ and $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ be two schema mappings where Σ and Σ' are finite sets of s-t tgds, and let $\mathcal{C}$ be a class of s-t tgds containing Σ and Σ'. An immediate consequence of Definition 4.4 is that if a data example (I, J) is an Armstrong database for both $\mathcal{M}$ and $\mathcal{M}'$ w.r.t. $\mathcal{C}$, then $\Sigma \equiv \Sigma'$. Thus, the existence of an Armstrong database yields a unique characterization via a single data example.

In spite of their desirable properties, Armstrong databases need not exist, even for fairly simple sets of database dependencies (see, e.g., [11]). We now introduce a relaxation of the notion of an Armstrong database.

DEFINITION 4.5. Let Σ and $\mathcal{C}$ be two sets of database dependencies over the same schema.

An *Armstrong basis for Σ w.r.t. $\mathcal{C}$* is a finite set $\mathcal{D}$ of instances such that for every dependency $\sigma \in \mathcal{C}$, we have that $\Sigma \models \sigma$ if and only if $D \models \sigma$, for every instance $D \in \mathcal{D}$.

An *Armstrong basis for a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ w.r.t. $\mathcal{C}$* is an Armstrong basis for Σ w.r.t. $\mathcal{C}$. □

It is clear that the existence of an Armstrong database implies the existence of an Armstrong basis, since, if D is an Armstrong database for Σ w.r.t. $\mathcal{C}$, then the singleton $\{D\}$ is an Armstrong basis for Σ w.r.t. $\mathcal{C}$. The next result shows that the converse need not be true.

PROPOSITION 4.6. *Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping, where $\Sigma = \{P(x) \rightarrow P'(x),\ Q(x) \rightarrow Q'(x)\}$.*

1. *There does not exist an Armstrong database for $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds.*
2. *There is an Armstrong basis for $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds.*

PROOF. For the first part and towards a contradiction, suppose that (I, J) is an Armstrong database for $\mathcal{M}$ w.r.t. the class of LAV s-t tgds. Consider the set $\Sigma' = \{\sigma_1, \sigma_2\}$, where σ_1 is the LAV s-t tgd $P(x) \rightarrow \exists y Q'(y)$, and σ_2 is the LAV s-t tgd $Q(x) \rightarrow \exists y P'(y)$. Since $\Sigma \not\models \sigma_1$ and $\Sigma \not\models \sigma_2$, it follows that $(I, J) \not\models \sigma_1$ and $(I, J) \not\models \sigma_2$. This implies that $P(a) \in I$ and $Q(b) \in I$, for some values a and b. Since $(I, J) \models \Sigma$, we must have that $P'(a) \in J$ and $Q'(b) \in J$. But this means that $(I, J) \models \Sigma'$, which is a contradiction.

For the second part, let $\mathcal{D} = \{D_1, D_2\}$, where $D_1 = (\{P(a)\}, \{P'(a)\})$ and $D_2 = (\{Q(a)\}, \{Q'(a)\})$. We will show that $\mathcal{D}$ is an Armstrong basis for $\mathcal{M}$. Clearly, $D_1 \models \Sigma$ and $D_2 \models \Sigma$. Next, let σ be a LAV s-t tgd. We will show that if $\Sigma \not\models \sigma$, then $D_1 \not\models \sigma$ or $D_2 \not\models \sigma$. There are two cases to consider. First, suppose that σ is of the form $P(x) \rightarrow \exists \mathbf{y} \psi(x, \mathbf{y})$. Since $\Sigma \not\models \sigma$, there must exist a data example (I, J) such that $(I, J) \models \Sigma$ and $(I, J) \not\models \sigma$. Hence, $I \models P(b)$, for some value b, and $J \models \neg \exists \mathbf{y} \psi(b, \mathbf{y})$. Since $(I, J) \models \Sigma$, it follows that $P'(b) \in J$. Hence, $(\{P(b)\}, \{P'(b)\}) \not\models \sigma$. Therefore, $D_1 \not\models \sigma$, since D_1 is isomorphic to $(\{P(b)\}, \{P'(b)\})$. Finally, suppose that σ is of the form $Q(x) \rightarrow \exists \mathbf{y} \psi(x, \mathbf{y})$. A similar argument is used to show that $D_2 \not\models \sigma$. □

As far as we can tell from perusing the literature, the notion of an Armstrong basis is new; in particular, it has not been considered during the investigation of Armstrong databases. One plausible explanation for this is that much of the research on Armstrong databases focused on *unirelational* databases (i.e., on databases over a schema consisting of a single relation) and on *typed tgds* (see [8, 7] for the precise definition). It turns out that, in that context, an Armstrong database exists if and only if an Armstrong basis exists. The reason is that results in [8] imply that if $\mathcal{D} = \{D_1, \ldots, D_k\}$ is an Armstrong basis for a set Σ of typed tgds w.r.t. the set of all typed tgds over a unirelational schema, then the direct product $D_1 \times \cdots \times D_k$ is an Armstrong database for Σ.

Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ and $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ be two schema mappings where Σ and Σ' are finite sets of s-t tgds, and let $\mathcal{C}$ be a class of s-t tgds containing Σ and Σ'. From Definition 4.5, it follows easily that if $\mathcal{D} = \{D_1, \ldots, D_k\}$ is an Armstrong basis for both $\mathcal{M}$ and $\mathcal{M}'$ w.r.t. $\mathcal{C}$, then $\Sigma \equiv \Sigma'$. Thus, the existence of an Armstrong basis yields a unique characterization of the schema mapping via a finite set of data examples.

The next simple proposition gives a connection between unique characterizations via positive examples and Armstrong bases.

PROPOSITION 4.7. *Assume that $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping, where Σ is a finite set of s-t tgds, and $\mathcal{P}$ is a finite set of positive examples that uniquely characterizes $\mathcal{M}$ w.r.t. a class $\mathcal{C}$ of s-t tgds. Then $\mathcal{P}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$. The converse is not true in general.*

PROOF. To show that $\mathcal{P}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$, we need to show that for every $\sigma \in \mathcal{C}$, we have $\Sigma \models \sigma$ if and only if $D \models \sigma$ for every $D \in \mathcal{P}$. It is easy to see that if $\Sigma \models \sigma$, then $D \models \sigma$ for every $D \in \mathcal{P}$, since $\mathcal{P}$ is a set of positive examples for Σ. The converse direction follows immediately from the fact that $\mathcal{P}$ uniquely characterizes Σ. Assume that $D \models \sigma$ for every $D \in \mathcal{P}$. Hence, $\mathcal{P}$ consists of positive examples for σ. Since $\mathcal{P}$ uniquely

characterizes Σ w.r.t. $\mathcal{C}$, it follows that $\Sigma \equiv \sigma$ and, in particular, $\Sigma \models \sigma$.

As seen in Theorem 3.5, there is a schema mapping specified by a single LAV s-t tgd that is not uniquely characterizable by any finite set of positive and negative examples w.r.t. the class of all LAV s-t tgds. In Section 5, however, we shall show that every finite set of LAV s-t tgds has an Armstrong basis w.r.t. the class of all LAV s-t tgds. □

In the next section, we establish a necessary and sufficient condition for the existence of an Armstrong basis.

4.3 Armstrong Bases and Universal Examples

In this section, we show that the existence of an Armstrong basis is equivalent to unique characterizability by a finite set of universal examples. We begin with a lemma that will be used repeatedly in the proofs.

LEMMA 4.8. *Let $\mathcal{C}$ be a class of s-t tgds, let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping where Σ is a finite set of s-t tgds, and let $(I_1, J_1), \ldots, (I_k, J_k)$ be data examples.*

1. *Assume that $\{(I_1, J_1), \ldots, (I_k, J_k)\}$ is a set of universal examples for $\mathcal{M}$ that uniquely characterizes $\mathcal{M}$ w.r.t. $\mathcal{C}$, and let $J'_1, \ldots, J'_k$ be target instances such that J_i is homomorphically equivalent to J'_i, for all $i \leq k$. Then the set $\{(I_1, J'_1), \ldots, (I_k, J'_k)\}$ is a set of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. $\mathcal{C}$.*
2. *Assume that $\{(I_1, J_1), \ldots, (I_k, J_k)\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$, and let $J'_1, \ldots, J'_k$ be target instances such that J'_i is a solution for I_i w.r.t. $\mathcal{M}$ and $J'_i \rightarrow J_i$, for all $i \leq k$. Then the set $\{(I_1, J'_1), \ldots, (I_k, J'_k)\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$.*

   In particular, if $\{(I_1, J_1), \ldots, (I_k, J_k)\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$, and J'_i is a universal solution for I_i w.r.t. $\mathcal{M}$, for all $i \leq k$, then the set $\{(I_1, J'_1), \ldots, (I_k, J'_k)\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$.

PROOF. The proofs of both parts follow easily from the fact that s-t tgds are preserved under target homomorphisms, that is, if σ is a s-t tgd, (I, J) is a data example such that $(I, J) \models \sigma$, and K is a target instance such that $J \rightarrow K$, then $(I, K) \models \sigma$. We leave the details of the first part to the reader. For the second part, observe first that if $\Sigma \models \sigma$, then for every $i \leq k$, we have that $(I_i, J'_i) \models \sigma$ because $(I, J'_i) \models \Sigma$ (since J'_i is a solution for I_i w.r.t. $\mathcal{M}$). Assume now that $\Sigma \not\models \sigma$, for some $\sigma \in \mathcal{C}$. Then, by the definition of an Armstrong basis, there is some $i \leq k$ such that $(I_i, J_i) \not\models \sigma$. Since s-t tgds are preserved under target homomorphisms and since $J'_i \rightarrow J_i$, it follows that $(I_i, J'_i) \not\models \sigma$. □

THEOREM 4.9. *Assume that $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping, where Σ is a finite set of s-t tgds, and $\mathcal{U}$ is a finite set of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. to a class $\mathcal{C}$ of s-t tgds. Then $\mathcal{U}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$.*

PROOF. Assume that $\mathcal{U} = \{(I_1, J_1), ..., (I_k, J_k)\}$. For each $i \leq k$, let $\mathit{chase}_{\mathcal{M}}(I_i)$ be the universal solution for I_i w.r.t. $\mathcal{M}$ obtained applying the naive chase procedure to I_i. Since J_i is homomorphically equivalent to $\mathit{chase}_{\mathcal{M}}(I_i)$ for all $i \leq k$, by the first part of Lemma 4.8, it immediately follows that the set $\mathcal{U}'$= $\{(I_1, \mathit{chase}_{\mathcal{M}}(I_1)), \ldots, (I_k, \mathit{chase}_{\mathcal{M}}(I_k))\}$ is a set of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. $\mathcal{C}$. Next, we will show that $\mathcal{U}'$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$, which, by the second part of Lemma 4.8 will imply that $\mathcal{U}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$.

Clearly, $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \models \Sigma$, for every $i \leq k$, hence if $\sigma \in \mathcal{C}$ and $\Sigma \models \sigma$, then $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \models \sigma$, for every $i \leq k$. It remains to show that if $\sigma \in \mathcal{C}$ and $\Sigma \not\models \sigma$, then there is some $i \leq k$ such that $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \not\models \sigma$. Since $\Sigma \not\models \sigma$, it follows that $\Sigma \not\models \Sigma \cup \{\sigma\}$. Consequently, there is some $i \leq k$ such that $\mathit{chase}_{\mathcal{M}}(I_i)$ is not a universal solution for I_i w.r.t. to the schema mapping $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma \cup \{\sigma\})$. We claim that $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \not\models \sigma$. Towards a contradiction, suppose that $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \models \sigma$. Hence, $\mathit{chase}_{\mathcal{M}}(I_i)$ is a solution for I_i w.r.t. $\mathcal{M}'$. Consider the universal solution $\mathit{chase}_{\mathcal{M}'}(I_i)$ for I_i w.r.t. $\mathcal{M}'$ obtained by chasing I_i with $\Sigma \cup \{\sigma\}$. Then $\mathit{chase}_{\mathcal{M}'}(I_i) \rightarrow \mathit{chase}_{\mathcal{M}}(I_i)$. At the same time, by the construction of the result of the naive chase, we have that $\mathit{chase}_{\mathcal{M}}(I_i) \subseteq \mathit{chase}_{\mathcal{M}'}(I_i)$, hence $\mathit{chase}_{\mathcal{M}}(I_i) \rightarrow \mathit{chase}_{\mathcal{M}'}(I_i)$. It follows that $\mathit{chase}_{\mathcal{M}}(I_i)$ is homomorphically equivalent to $\mathit{chase}_{\mathcal{M}'}(I_i)$, hence $\mathit{chase}_{\mathcal{M}}(I_i)$ is a universal solution for I_i w.r.t. $\mathcal{M}'$, which is a contradiction. □

THEOREM 4.10. *Assume that $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping, where Σ is a finite set of s-t tgds, and $\mathcal{A}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. a class $\mathcal{C}$ of s-t tgds. Then there is a finite set $\mathcal{U}$ of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. $\mathcal{C}$.*

PROOF. Assume that $\mathcal{A} = \{(I_1, J_1), \ldots, (I_k, J_k)\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$. By Lemma 4.8, the set $\mathcal{U}_1 = \{(I_1, \mathit{chase}_{\mathcal{M}}(I_1)), \ldots, (I_k, \mathit{chase}_{\mathcal{M}}(I_k))\}$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$. Let $\mathcal{U}_2$ be the set of all pairs $(I, \mathit{chase}_{\mathcal{M}}(I))$ such that $|\mathrm{adom}(I)| \leq n$, where n is the maximum number of variables in the antecedents of s-t tgds in Σ. We will show that $\mathcal{U}_1 \cup \mathcal{U}_2$ is a set of universal examples for $\mathcal{M}$ that uniquely characterizes $\mathcal{M}$ w.r.t $\mathcal{C}$, i.e., if $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ is a schema mapping such that $\mathcal{U}_1 \cup \mathcal{U}_2$ is a set of universal examples for $\mathcal{M}'$, then $\Sigma \equiv \Sigma'$.

We first show that $\Sigma \models \Sigma'$. Let $\sigma' \in \Sigma'$. Since $\mathcal{U}_1$ is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$, if $\Sigma \not\models \sigma'$, there exists $1 \leq i \leq k$ such that $(I_i, \mathit{chase}_{\mathcal{M}}(I_i)) \not\models \sigma'$. This, however, contradicts our assumption that $\mathcal{U}_1 \cup \mathcal{U}_2$ is a set of universal examples for $\mathcal{M}'$.

Next, we show that $\Sigma' \models \Sigma$, that is, if $(I', J') \models \Sigma'$, then $(I', J') \models \Sigma$. Let $\sigma \in \Sigma$ be a s-t tgd of the form $\phi(\mathbf{x}) \rightarrow \exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y})$ and suppose that $I' \models \phi(\mathbf{a})$, which means that I' contains all the facts in $\phi(\mathbf{a})$. Since the number of distinct variables in the antecedent of σ is at most n, there must be a pair $(I, J) \in \mathcal{U}_2$ such that (the instance consisting of the facts in) $\phi(\mathbf{a})$ is isomorphic to I. If h is an isomorphism from I to $\phi(\mathbf{a})$, then $I \models \phi(h^{-1}(\mathbf{a}))$. Since h is an isomorphism from I to $\phi(\mathbf{a})$ and since I' contains all the facts in $\phi(\mathbf{a})$, it follows that h is a homomorphism from I to I'. From Theorem 3.9 of [24], we know that $\mathcal{M}'$ *reflects source homomorphisms*. By definition, this means that for all source instances K, K' and for all target instances L, L' such that L is a universal solution for K and L' is a solution for K', we have that every homomorphism $h : K \rightarrow K'$ extends to a homomorphism from L to L'. (Note that in this definition, we do not require the homomorphisms to be constant on $\mathrm{adom}(K)$.) Thus, since $\mathit{chase}_{\mathcal{M}}(I)$ is a universal solution for I w.r.t. $\mathcal{M}'$, and J' is a solution for I' w.r.t. $\mathcal{M}'$, h can be extended to a homomorphism $h' : \mathit{chase}_{\mathcal{M}}(I) \rightarrow J'$. Since $(I, \mathit{chase}_{\mathcal{M}}(I)) \models \sigma$ and $I \models \phi(h^{-1}(\mathbf{a}))$, we have that $\mathit{chase}_{\mathcal{M}}(I) \models \psi(h^{-1}(\mathbf{a}), \mathbf{b})$ for some $\mathbf{b}$, hence $J' \models \psi(h'(h^{-1}(\mathbf{a})), h'(\mathbf{b}))$. Thus, $J' \models \psi(\mathbf{a}, \mathbf{b}')$, for some values $\mathbf{b}'$, which was to be shown. □

By combining Theorems 4.9 and 4.10, we conclude that the existence of an Armstrong basis is equivalent to unique characterizability by universal examples.

COROLLARY 4.11. *Assume that $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is a schema mapping, where Σ is a finite set of s-t tgds, and $\mathcal{C}$ is a set of s-t tgds. Then the following statements are equivalent.*

1. *There is a finite set $\mathcal{U}$ of universal examples for $\mathcal{M}$ such that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$ w.r.t. $\mathcal{C}$.*
2. *There is an Armstrong basis for $\mathcal{M}$ w.r.t. $\mathcal{C}$.*

5. Characterizations via Universal Examples

In this section, we explore the capabilities of universal examples in yielding unique characterizations of schema mappings specified by s-t tgds.

THEOREM 5.1. *If $\mathcal{M}$ is a schema mapping specified by a finite set of LAV s-t tgds, then there is a finite set $\mathcal{U}$ of universal examples for $\mathcal{M}$ such that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$ w.r.t. the class of all LAV s-t tgds.*

PROOF. Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping, where Σ is a finite set of LAV s-t tgds. We will construct a finite set $\mathcal{U}$ of universal examples and will show that $\mathcal{M}$ is uniquely characterized by $\mathcal{U}$ w.r.t. the class of all LAV s-t tgds.

Suppose that the source schema $\mathbf{S}$ consists of the relation symbols $R_1, ..., R_s$. For each $i \leq s$, let r_i be the arity of R_i, and let k be the maximum of $r_1, \ldots, r_s$. Let D be a set of k distinct elements, say, $D = \{d_1, ..., d_k\}$. For each relation symbol R_i, $1 \leq i \leq s$, and each r_i-ary tuple $\mathbf{d}$ of elements from D, construct the data example $(\{R_i(\mathbf{d})\}, \mathit{chase}_{\mathcal{M}}(\{R_i(\mathbf{d})\}))$. Let $\mathcal{U}$ be the set of all data examples obtained via this construction. Clearly, every member of $\mathcal{U}$ is a universal example for $\mathcal{M}$. In what follows, we will show that $\mathcal{M}$ is uniquely characterized by $\mathcal{U}$ w.r.t. the class of all LAV s-t tgds. Let $\mathcal{M}' = (\mathbf{S}, \mathbf{T}, \Sigma')$ be a schema mapping, where Σ' is a finite set of LAV s-t tgds, and assume that every member of $\mathcal{U}$ is a universal example for $\mathcal{M}'$. We have to show that $\Sigma \equiv \Sigma'$.

We first show that $\Sigma \models \Sigma'$. Let (I, J) be a data example such that $(I, J) \models \Sigma$ and let $R_j(\mathbf{x}) \rightarrow \exists \mathbf{y} \phi(\mathbf{x}, \mathbf{y})$ be a LAV s-t tgd in Σ' such that $I \models R_j(\mathbf{a})$, for some tuple $\mathbf{a}$. We will show that $J \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$. Observe that, by the construction of $\mathcal{U}$, the singleton instance $\{R_j(\mathbf{a})\}$ must be isomorphic to a singleton instance $\{R_j(\mathbf{a}')\}$ used in the construction of the set $\mathcal{U}$. For notational simplicity, we will denote these singleton instances by $R_j(\mathbf{a})$ and $R_j(\mathbf{a}')$, respectively. It follows that the naive chase procedure on these instances produces isomorphic results, that is, we have that $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}'))$ is isomorphic to $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}))$ via an isomorphism that maps $\mathbf{a}'$ to $\mathbf{a}$. Also, by the construction of $\mathcal{U}$, we have that $(R_j(\mathbf{a}'), \mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}'))) \in \mathcal{U}$. Since $(R_j(\mathbf{a}'), \mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}')))$ is a universal example for $\mathcal{M}'$, it follows that $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}')) \models \exists \mathbf{y} \phi(\mathbf{a}', \mathbf{y})$, hence it must be that $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a})) \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$. On the other hand, since $(I, J) \models \Sigma$ and $R_j(\mathbf{a})$ is a sub-instance of I, we have that J is a solution for $R_j(\mathbf{a})$ w.r.t. $\mathcal{M}$. Since $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}))$ is a universal solution for $R_j(\mathbf{a})$ w.r.t. $\mathcal{M}$, this implies that $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a})) \rightarrow J$; consequently, $J \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$.

Next, we show that $\Sigma' \models \Sigma$. Suppose $(I, J) \models \Sigma'$. We will show that $(I, J) \models \Sigma$. Let $R_j(\mathbf{x}) \rightarrow \exists \mathbf{y} \phi(\mathbf{x}, \mathbf{y})$ be a LAV s-t tgd in Σ and assume that $I \models R_j(\mathbf{a})$ for some $\mathbf{a}$. We will show that $J \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$. As before, by the construction of $\mathcal{U}$, we have that $R_j(\mathbf{a})$ must be isomorphic to a source instance $R_j(\mathbf{a}')$ such that $(R_j(\mathbf{a}'), \mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}'))) \in \mathcal{U}$. Moreover, $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}'))$ is isomorphic to $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}))$ via an isomorphism that maps $\mathbf{a}'$ to $\mathbf{a}$. Since the pair $(R_j(\mathbf{a}'), \mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}')))$ is a universal example for $\mathcal{M}'$, it follows that $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a})))$ is a universal solution for $R_j(\mathbf{a})$ w.r.t. $\mathcal{M}'$. On the other hand, J is a solution for $R_j(\mathbf{a})$ w.r.t. M' (since J is a solution for I w.r.t. $\mathcal{M}'$), hence $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a})) \rightarrow J$. At the same time, since $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a}))$ is a universal solution for $R_j(\mathbf{a})$ w.r.t. $\mathcal{M}$, we know $\mathit{chase}_{\mathcal{M}}(R_j(\mathbf{a})) \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$; consequently, $J \models \exists \mathbf{y} \phi(\mathbf{a}, \mathbf{y})$. □

As an immediate consequence of Theorems 4.9 and 5.1, we obtain the following result that every LAV schema mapping has an Armstrong basis w.r.t. LAV s-t tgds.

COROLLARY 5.2. *If $\mathcal{M}$ is a schema mapping specified by a finite set of LAV s-t tgds, then $\mathcal{M}$ has an Armstrong basis w.r.t. the class of all LAV s-t tgds.*

Recall that, by Proposition 4.6, there is a schema mapping specified by two LAV s-t tgds that has no Armstrong database w.r.t. the class of all LAV s-t tgds. Thus, the preceding Corollary 5.2 cannot be strengthened to assert that every schema mapping specified by a finite set of LAV s-t tgds has an Armstrong database w.r.t. the class of all LAV s-t tgds.

Are there broader classes of schema mappings that have unique characterizations via universal examples? Equivalently, are there broader classes of schema mappings possessing Armstrong bases?

DEFINITION 5.3. ([24, Definition 2.6]) Let n be a positive integer. We say that a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a finite set of s-t tgds, is n-modular if for every data example (I, J) such that $(I, J) \not\models \Sigma$, there is a sub-instance I' of I such that $|\text{adom}(I')| \leq n$ and $(I', J) \not\models \Sigma$. □

The concept of n-modularity was introduced and studied in [24], where schema-mapping languages were characterized in terms of their structural properties. Intuitively, n-modularity means that every negative example has an "explanation" of size at most n. Every schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ specified by a finite set of s-t tgds is n-modular for some n; in fact, n can be taken to be the maximum number of variables occurring in the s-t tgds in Σ (see [24, Proposition 2.7]). Note, however, that, if $\mathbf{S}$ and $\mathbf{T}$ are non-unary schemas, then there is no fixed number k such that every schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ specified by a finite set Σ of s-t tgds is k-modular. To see this, let E be a binary source relation, let F be a binary target relation, and let σ_n be the GAV s-t tgd $\forall x \forall y (P_n(x, y) \rightarrow F(x, y))$, where $P_n(x, y)$ asserts that there is a path along E-edges of length n from x to y. Then, σ_n is $(n+1)$-modular, but not n-modular. In contrast, every schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ specified by a finite set of LAV s-t tgds is k-modular, where k is the maximum arity of the relation symbols in $\mathbf{S}$.

Our next result shows that Theorem 5.1 can be extended to the class of all n-modular schema mappings, n a positive integer. The proof, which is a generalization of the proof of Theorem 5.1, will be given in the full version of the paper.

THEOREM 5.4. *Let n be a positive integer and let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping, where Σ is a finite set of s-t tgds. If $\mathcal{M}$ is n-modular, then there is a finite set $\mathcal{U}$ of universal examples such that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$ w.r.t. the class of all m-modular schema mappings, $m \geq n$.*

Consequently, every n-modular schema mapping specified by a finite set of s-t tgds has an Armstrong basis w.r.t. the class of all m-modular schema mappings, $m \geq n$.

The preceding Theorem 5.4, has a number of applications, including the following one that covers many schema mappings occurring in practice.

DEFINITION 5.5. An s-t tgd $\phi(\mathbf{x}) \rightarrow \exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y})$ is said to be *self-join-free on the source* if none of the relation symbols in $\phi(\mathbf{x})$ is repeated. □

COROLLARY 5.6. *Let $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ be a schema mapping where Σ is a finite set of s-t tgds that are self-join-free on the source. Then there is a finite set $\mathcal{U}$ of universal examples such that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$ w.r.t. the class of all s-t tgds that are self-join-free on the source.*

PROOF. It is easy to see that if Σ consists of s-t tgds that are self-join-free on the source, then the schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ is n-modular, where n is the sum of the arities of all relation symbols in $\mathbf{S}$. Hence, by Theorem 5.4, there is a finite set $\mathcal{U}$ of universal examples such that $\mathcal{U}$ uniquely characterizes $\mathcal{M}$ w.r.t. the class of all n-modular schema mappings and, in particular, w.r.t. the class of s-t tgds that are self-join-free on the source. □

6. Limitations of Universal Examples

So far, we have shown that several important classes of schema mappings possess unique characterizations via universal examples. In this section, we establish that, although superior to positive and negative examples, universal examples have their own limitations.

By Proposition 4.3, the schema mapping $\mathcal{M}$ specified by the binary copy s-t tgd $E(x, y) \rightarrow F(x, y)$ can be uniquely characterized via universal examples w.r.t. to both the class of all LAV s-t tgds and the class of all GAV s-t tgds. Moreover, since the binary copy s-t tgd is 2-modular, Theorem 5.4 implies that, for every $m \geq 2$, $\mathcal{M}$ can be uniquely characterized via universal examples w.r.t. the class of all m-modular s-t tgds. The next proposition reveals that these results of $\mathcal{M}$ do not extend to a unique characterization of $\mathcal{M}$ via universal examples w.r.t. the class of *all* s-t tgds. Its proof illustrates the use of the connection between Armstrong databases and unique characterizations via universal examples.

PROPOSITION 6.1. *Let $\mathcal{M}$ be the schema mapping specified by the binary copy s-t tgd $E(x, y) \rightarrow F(x, y)$. Then there is no finite set of universal examples that uniquely characterizes $\mathcal{M}$ w.r.t. the class of all s-t tgds.*

PROOF. By Theorem 4.9, it suffices to show that $\mathcal{M}$ does not have an Armstrong basis w.r.t. the class of all s-t tgds. Towards a contradiction, assume that $\{(I_1, J_1), \ldots, (I_k, J_k)\}$ is such an Armstrong basis. Let n be a positive integer bigger than the maximum of $|\text{adom}(I_i)|$, for $1 \leq i \leq k$. Also, let c_i be the length of some cycle in I_i, if I_i contains at least one cycle; if I_i contains no cycle, then let $c_i = 1$. Take the product $m = n \cdot c_1 \cdot \ldots \cdot c_k$ of these quantities and consider the following s-t tgd σ':

$$\text{Path}_m(x_1, \ldots, x_{m+1}) \rightarrow \exists y_1 \ldots \exists y_m \text{Cycle}_m(y_1, \ldots, y_m),$$

where Path_m is a conjunction of E-atoms asserting that the variables $x_1, \ldots, x_{m+1}$ form a path of length m in E, and Cycle_m is a conjunction of F atoms asserting that the variables $y_1, \ldots, y_m$ form cycle of length m in F. Note that σ' is neither a LAV, nor a GAV s-t tgd. Clearly, $\Sigma \not\models \sigma'$. In what follows, we will show that $(I_i, J_i) \models \sigma'$, for all $i \leq k$, which will contradict the assumption that $(I_1, J_1), ..., (I_k, J_k)$ form an Armstrong basis for $\mathcal{M}$ w.r.t. the class of all s-t tgds. Indeed, take some (I_i, J_i), where $i \leq k$. If I_i contains no cycle, then $(I_i, J_i) \models \sigma'$ trivially, because $m > |\text{adom}(I_i)|$. If I_i contains a cycle, then J_i must contain all cycles of I_i (since (I_i, J_i) satisfies the binary copy s-t tgd). Now, I_i clearly contains a path of length m. Since J_i contains all cycles of I_i and since m is a multiple of the length of one of the cycles in I_i, it must be the case that J_i contains a cycle of length m; hence $(I_i, J_i) \models \sigma'$. □

We now address the question of whether or not schema mappings specified by GAV s-t tgds possess unique characterizations via universal examples w.r.t. the class of all GAV s-t tgds. Note again that the schema mapping $\mathcal{M}$ specified by the binary copy s-t tgd, which is a GAV s-t tgd, possesses such a characterization. Our main result in this section is that there are schema mappings specified by quite natural and simple-to-describe GAV s-t tgds for which this is not true, even if negative examples are also used. The proof of this result will make use of sophisticated machinery from graph theory that we describe next.

Back in 1959, Erdös [6] showed that there are graphs of arbitrarily large girth and chromatic number, where the girth of a graph is the size of its smallest cycle (cycles are assumed to have length at least 3), and the chromatic number of a graph is the minimum number of colors needed to color it. This result was proved via one of the first applications of the *probabilistic method*, that is, such graphs were not constructed explicitly but, instead, were shown to have a positive probability. Explicit constructions were given much later; in particular, there is a family of explicitly constructed graphs, known as *Ramanujan graphs*, that have arbitrarily large girth and chromatic number [20]. Later on, Nešetřil and Rödl [21] vastly generalized Erdös' result using the probabilistic method. Next, we describe this generalization following the exposition in [14, Chapter 3]. We begin with a definition.

DEFINITION 6.2. Let k be a positive integer. Two graphs G and H are said to be *k-equivalent* if for every graph K with at most k vertices, there is a homomorphism from G to K if and only if there is a homomorphism from H to K. □

In Definition 6.2, the notion of *homomorphism* is the standard one in graph theory: a homomorphism from a graph $G = (V_1, E_1)$ to a graph $H = (V_2, E_2)$ is a function h from V_1 to V_2 that maps edges in E_1 to edges in E_2, i.e., if $E_1(a, b)$ holds, then also $E_2(h(a), h(b))$ holds (thus, homomorphisms are not required to be constant on some nodes)

THEOREM 6.3. ([14, Theorem 3.15]) *Let k and m be two positive integers. Then every graph G has a k-equivalent graph H of girth at least m.*

The preceding Theorem 6.3 provides us with the ideal tool for establishing the main result of this section. Before stating the main result, we need to introduce one more concept.

Let $\mathbf{S}$ be a source schema consisting of a unary relation symbol P and a binary relation symbol E, and let $\mathbf{T}$ be a target schema consisting of a unary relation symbol R. If $G = (V_1, E_1)$ is a graph, then we write Q^G to denote the *canonical conjunctive query of G*, that is, Q^G is a Boolean conjunctive query asserting that there are $|V_1|$ nodes connected the same way as the nodes of G are. For example, if G is the complete graph K_n on n nodes, then Q^G is

$$\exists x_1 \ldots \exists x_n \bigwedge_{i \neq j} E(x_i, x_j).$$

Let $G = (V_1, E_1)$ be a graph and consider the first-order sentence

$$\forall x(P(x) \wedge Q^G \rightarrow R(x)),$$

where the variable x is different from all variables occurring in Q^G. This sentence is logically equivalent to a GAV s-t tgd σ_G obtained by pulling the existential quantifiers in Q^G to the front and turning them into universal quantifiers. In effect, σ_G is a unary copy s-t tgd with a "trigger". Specifically, assume that I is a source instance consisting of a unary relation P^I and a binary relation E^I. Then the relation P^I is copied to the target relation interpreting R, provided E^I satisfies the Boolean conjunctive query Q^G, that is, provided there is a homomorphism from E_1 to E^I. We will refer to the GAV s-t tgd σ_G as the *unary copy s-t tgd with trigger G*.

THEOREM 6.4. *Let $G = (V_1, E_1)$ be a graph containing a cycle and let $\mathcal{M}_G$ be the schema mapping specified by the unary copy s-t tgd with trigger G.*

1. *There are no finite sets of universal examples and negative examples that uniquely characterize $\mathcal{M}_G$ w.r.t. the class of all GAV s-t tgds.*
2. *There are no finite sets of positive examples and negative examples that uniquely characterize $\mathcal{M}_G$ w.r.t. the class of all GAV s-t tgds.*

PROOF. We will prove the first part; the proof of the second part is similar. Assume that $(I_1, J_1), \ldots, (I_s, J_s)$ are universal examples for $\mathcal{M}_G$ and $(I'_1, J'_1), \ldots, (I'_t, J'_t)$ are negative examples for $\mathcal{M}_G$. Let k be the maximum of $|\text{adom}(I_i)|$, $1 \leq i \leq s$, and of $|\text{adom}(I'_j)|$, $1 \leq j \leq t$, and let $m = \text{girth}(G) + 1$, where $\text{girth}(G)$ stands for the girth of G. By Theorem 6.3, there is a graph $H = (V_2, E_2)$ that is k-equivalent to G and has girth at least m. Let σ_H be the unary copy s-t tgd with trigger H, and let $\mathcal{M}_H$ be the schema mapping specified by σ_H. We claim that the following hold: (a) $\sigma_G \not\models \sigma_H$; (b) each (I_i, J_i) is a universal example for $\mathcal{M}_H$, $1 \leq i \leq s$; and (c) each (I'_j, J'_j) is a negative example for $\mathcal{M}_H$, $1 \leq j \leq t$.

We first show that $\sigma_G \not\models \sigma_H$. Let I be the source instance in which the unary relation P is interpreted by some non-empty set A and the binary relation is interpreted by the edge relation E_2 of H. Since $\text{girth}(H) > \text{girth}(G)$, there cannot be a homomorphism from G to H. This implies that (the antecedent of) σ_G is never "triggered" on I; consequently, $(I, \emptyset) \models \sigma_G$. In contrast, $(I, \emptyset) \not\models \sigma_H$, since σ_H is "triggered" on I, but the set A is not contained in $\emptyset$ (i.e., the emptyset).

Next, we show that each (I_i, J_i) is a universal example for σ_H. There are two cases to consider. In the first case, assume that there is an assignment h from the variables of σ_H to values in $\text{adom}(I_i)$ so that the antecedent of σ_H becomes true. In particular, h is a homomorphism from E_2 to the binary relation E^{I_i} of I_i. To show that J_i is a universal solution for I_i w.r.t. $\mathcal{M}_H$, we have to show that $R^{J_i} = P^{I_i}$, because in this case σ_H is "triggered" on I_i. Since $|\text{adom}(I_i)| \leq k$ and H is k-equivalent of G, it follows that there is a homomorphism g from E_1 to the binary relation E^{I_i} of I_i, Consequently, σ_G is "triggered" on I_i and, since J_i is a universal solution for I_i w.r.t. $\mathcal{M}_G$, we must have that $R^{J_i} = P^{I_i}$. In the second case, assume that there is no assignment from the variables of σ_H to values in $\text{adom}(I_i)$ so that the antecedent of σ_H becomes true. In particular, there is no homomorphism from E_2 to the binary relation E^{I_i} of I_i. In this case, to show that J_i is a universal solution for I_i w.r.t. $\mathcal{M}_H$, we must show that $R^{J_i} = \emptyset$. Since $|\text{adom}(I_i)| \leq k$ and H is k-equivalent of G, it follows that there is no homomorphism from E_1 to the binary relation E^{I_i} of I_i (observe that here we are using the other direction of k-equivalence). Consequently, σ_G is not "triggered" on I_i and, since J_i is a universal solution for I_i w.r.t. $\mathcal{M}_G$, we must have that $R^{J_i} = \emptyset$.

Finally, we show that each (I'_j, J'_j) is a negative example for $\mathcal{M}_H$. Since (I'_j, J'_j) is a negative example for $\mathcal{M}_G$, there is an assignment g from the variables of the antecedent of σ_G to I_i so that the following hold: (a) g is a homomorphism from E_1 to the binary relation $E^{I'_j}$ of I'_j; (b) there is a value a such $a \in P^{I'_j}$ and $a \notin R^{J'_j}$. Since $|\text{adom}(I'_j)| \leq k$ and H is k-equivalent of G, it follows that there is a homomorphism h from E_2 to $E^{I'_j}$. Hence, σ_H is "triggered" on I'_j and so $(I'_j, J'_j) \not\models \sigma_H$, because $a \in P^{I'_j}$ but $a \notin R^{J'_j}$. □

COROLLARY 6.5. *Let $G = (V_1, E_1)$ be a graph containing a cycle. The schema mapping $\mathcal{M}_G$ specified by the unary copy s-t tgd with trigger G has no Armstrong basis w.r.t. the class of all GAV s-t tgds.*

Our negative result about schema mappings specified by GAV s-t tgds raises the following natural question: is the unique characterizability via universal examples of GAV schema mappings w.r.t. the class of all GAV s-t tgds a decidable problem? More precisely, is there an algorithm that solves the following decision problem: given a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ specified by finite set Σ of GAV s-t tgds, does there exist a finite set of universal examples for $\mathcal{M}$ that uniquely characterizes $\mathcal{M}$ w.r.t. the class of all GAV s-t tgds? In a followup paper [25], it is shown that this problem is indeed decidable.

7. Concluding Remarks

Schema mappings specified by finite sets of s-t tgds are the most extensively studied and widely used schema mappings in data exchange and data integration. A schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a finite set of s-t tgds, constitutes a finite syntactic representation of the infinite space of all data examples (I, J) such that $(I, J) \models \Sigma$. In this paper, we addressed the following question: Can this infinite space of data examples be "captured" by a finite set of data examples? We formalized this question by considering notions of unique characterizations of schema mappings via a finite set of examples of a certain "type" (or of certain "types") w.r.t. a class of s-t tgds. We showed that, although very natural, positive and negative examples do not yield interesting unique characterizations. For this reason, we focused on universal examples as candidates for unique characterizations of schema mappings. We delineated the capabilities and limitations of universal examples, and, in the process, unveiled an a priori unexpected connection with the classical notion of an Armstrong database.

In this paper, we have considered positive and negative examples, and universal examples as natural candidates for unique characterizations of schema mappings. Naturally, the following question arises: Are there other "types" of data examples or combinations of such "types" of examples that yield interesting unique characterizations of rich classes of schema mappings?

It is worth pointing out that we regard the results reported here as the first step towards a broader program aiming to develop a methodology and a set of tools for understanding and refining schema mappings. Beyond unique characterizations, we plan to investigate weaker ways in which a finite set of data examples "captures" a schema mapping. In particular, given a finite set of data examples of various "types", is there a schema mapping that is "consistent" with the given data examples? This problem is analogous to problems in computational learning, where the goal is to find a concept that is compatible with a finite set of examples that are labeled positive or negative. It should also be noted that a framework and an accompanying cost model for discovering a schema mapping based on a single example were recently introduced and studied in [23].

In the long term, we envision the development of a system that would be capable of generating data examples that illustrate a schema mapping. Furthermore, after the data examples have been generated, a mapping designer would be allowed to modify the data examples at hand, and then the system would automatically fine-tune the existing schema mapping based on the modified data examples.

Acknowledgments. We would like to thank Balder ten Cate for numerous insightful comments on an earlier version of this paper. Alexe, Kolaitis, and Tan are supported by NSF grants IIS-0430994 and NSF grant IIS-0905276. Tan is also supported by a NSF CAREER award IIS-0347065.

8. References

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

[2] B. Alexe, L. Chiticariu, R. J. Miller, and W. C. Tan. Muse: Mapping Understanding and deSign by Example. In *International Conference on Data Engineering (ICDE)*, pages 10–19, 2008.

[3] B. Alexe, W. C. Tan, and Y. Velegrakis. STBenchmark: Towards a Benchmark for Mapping Systems. *Proceedings of the VLDB Endowment (PVLDB)*, 1(1):230–244, 2008.

[4] W. W. Armstrong. Dependency Structures of Data Base Relationships. In *IFIP Congress*, pages 580–583, 1974.

[5] L. Chiticariu and W. C. Tan. Debugging schema mappings with routes. In *International Conference on Very Large Data Bases (VLDB)*, pages 79–90, 2006.

[6] P. Erdös. Graph theory and probability. *Canadian J. of Mathematics*, 11:34–38, 1959.

[7] R. Fagin. Armstrong Databases. In *7th IBM Symposium on Mathematical Foundations of Computer Science*, 1982.

[8] R. Fagin. Horn Clauses and Database Dependencies. *Journal of the Association for Computing Machinery (JACM)*, 29(4):952–985, Oct. 1982.

[9] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data Exchange: Semantics and Query Answering. *Theoretical Computer Science (TCS)*, 336(1):89–124, 2005.

[10] R. Fagin, P. G. Kolaitis, A. Nash, and L. Popa. *(PODS)*, 2008. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 33–42, 2008.

[11] R. Fagin and M. Y. Vardi. Armstrong Databases for Functional and Inclusion Dependencies. *Inf. Process. Lett.*, 16(1):13–19, 1983.

[12] R. Fagin and M. Y. Vardi. The Theory of Data Dependencies - A Survey. In M. Anshel and W. Gewirtz, editors, *Proc. of Symposia in Applied Mathematics*, volume 34 - Mathematics of Information Processing, pages 19–71. American Mathematical Society, Providence, Rhode Island, 1986.

[13] L. M. Haas, M. A. Hernández, H. Ho, L. Popa, and M. Roth. Clio Grows Up: From Research Prototype to Industrial Tool. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 805–810, 2005.

[14] P. Hell and J. Nešetřil. *Graphs and Homomorphisms*. Oxford University Press, 2004.

[15] M. A. Hernández, R. J. Miller, L. Haas, L. Yan, C. T. H. Ho, and X. Tian. Clio: A Semi-Automatic Tool for Schema Mapping, *System Demonstration*. *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 30(2), May 2001.

[16] W. Hodges. *A Shorter Model Theory*. Cambridge University Press, 1997.

[17] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

[18] P. G. Kolaitis. Schema Mappings, Data Exchange, and Metadata Management. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 61–75, 2005.

[19] M. Lenzerini. Data Integration: A Theoretical Perspective. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 233–246, 2002.

[20] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[21] J. Nešetřil and V. Rödl. Chromatically optimal rigid graphs. *J. Comb. Theory, Ser. B*, 46(2):133–141, 1989.

[22] C. Olston, S. Chopra, and U. Srivastava. Generating example data for dataflow programs. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 245–256, 2009.

[23] P. Senellart and G. Gottlob. On the complexity of deriving schema mappings from database instances. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 23–32, 2008.

[24] B. ten Cate and P. G. Kolaitis. Structural Characterizations of Schema-Mapping Languages. In *International Conference on Database Theory (ICDT)*, pages 63–72, 2009.

[25] B. ten Cate and P. G. Kolaitis and W. C. Tan. Database Constraints and Homomorphism Dualities. *Work in Progress*, 2010.

[26] L.-L. Yan, R. J. Miller, L. M. Haas, and R. Fagin. Data-Driven Understanding and Refinement of Schema Mappings. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 485–496, 2001.