Limits of Schema Mappings

Phokion G. Kolaitis[1], Reinhard Pichler[2], Emanuel Sallinger[3], and Vadim Savenkov[4]

1 University of California – Santa Cruz, Santa Cruz, USA; and
IBM Research-Almaden, San Jose, USA
2 TU Wien, Vienna, Austria
3 University of Oxford, Oxford, UK
4 Vienna University of Economics and Business, Vienna, Austria

Abstract

Schema mappings have been extensively studied in the context of data exchange and data integration, where they have turned out to be the right level of abstraction for formalizing data inter-operability tasks. Up to now and for the most part, schema mappings have been studied as static objects, in the sense that each time the focus has been on a single schema mapping of interest or, in the case of composition, on a pair of schema mappings of interest.

In this paper, we adopt a dynamic viewpoint and embark on a study of sequences of schema mappings and of the limiting behavior of such sequences. To this effect, we first introduce a natural notion of distance on sets of finite target instances that expresses how "close" two sets of target instances are as regards the certain answers of conjunctive queries on these sets. Using this notion of distance, we investigate pointwise limits and uniform limits of sequences of schema mappings, as well as the companion notions of pointwise Cauchy and uniformly Cauchy sequences of schema mappings. We obtain a number of results about the limits of sequences of GAV schema mappings and the limits of sequences of LAV schema mappings that reveal striking differences between these two classes of schema mappings. We also consider the completion of the metric space of sets of target instances and obtain concrete representations of limits of sequences of schema mappings in terms of generalized schema mappings, i.e., schema mappings with infinite target instances as solutions to (finite) source instances.

1998 ACM Subject Classification H.2.0 Database Management – General

Keywords and phrases Limit, Pointwise convergence, Uniform convergence, Schema mapping

Digital Object Identifier 10.4230/LIPIcs.ICDT.2016.19

1 Introduction

Schema mappings have been extensively studied in the context of data exchange and data integration, where they have turned out to be the right level of abstraction for formalizing data inter-operability tasks (see the surveys [11, 12] and the monograph [1]). Up to now and for the most part, schema mappings have been studied as static objects, in the sense that each time the focus has been on a single schema mapping or on a finite and, typically, small number of schema mappings. In the case of data exchange [6], a single schema mapping is used to specify the relationship between a source schema and a target schema. In the case of operators on schema mappings [3], such as the composition operator [14, 8], a fixed number of schema mappings is used as input (e.g., two schema mappings in the case of composition) and return another schema mapping as output. Even the case of schema-mapping evolution [9] entails a finite (but potentially large) number of schema mappings.

19th International Conference on Database Theory (ICDT 2016).
Editors: Wim Martens and Thomas Zeume; Article No. 19; pp. 19:1–19:17
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In this paper, we adopt a dynamic viewpoint and embark on a systematic investigation of sequences of schema mappings and of the limiting behavior of such sequences. The original motivation came from the earlier work [2, 5, 7, 10, 14] on schema-mapping optimization and the study of various notions of equivalence between schema mappings that, intuitively, stipulate that two schema mappings cannot be distinguished using conjunctive queries (CQ-equivalence) or conjunctive queries with at most n variables (CQ_n-equivalence), for some fixed $n \geq 1$. In particular, in [5] and, implicitly, in [14], it was shown that, given an SO-tgd (second-order tuple-generating dependency) σ and a positive integer n, one can construct a GLAV schema mapping that is CQ_n-equivalent to σ. Informally, this means that a given SO tgd can be "approximated" by GLAV schema mappings up to any fixed level of precision, even though an SO tgd is a formula of second-order logic that may not be logically equivalent to any formula of first-order logic and, in particular, to any GLAV schema mapping. A more dynamic interpretation is that, given an SO-tgd σ, one can obtain a sequence of GLAV schema mappings $(\mathcal{M}_n)_{n\geq 1}$, whose "limit" is σ.

Summary of Results. Our contributions are both conceptual and technical. At the conceptual level, we develop a framework for studying sequences of schema mappings by first introducing a natural notion of distance dist on the powerset $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ of the set $\mathrm{Inst}(\mathbf{T})$ of finite instances over a schema $\mathbf{T}$. Intuitively, this notion of distance expresses how "close" two sets of finite $\mathbf{T}$-instances are as regards the certain answers of conjunctive queries on these sets. The pair $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$ is a pseudometric space, which means that the distance function dist is symmetric and obeys the triangle inequality, but different sets of finite target instances may have distance zero; however, two such sets have distance zero if and only if they are CQ-equivalent, i.e., every conjunctive query has the same certain answers on these two sets. Thus, we will also work with the metric space obtained by considering the CQ-equivalence classes of members of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$, and will use the same notation for it.

Sequences of functions from some set to a metric space occupy a central place in the study of metric spaces (see, e.g., [18]). In particular, there are natural notions of a *pointwise limit* and of a *uniform limit* of a sequence $(f_n)_{n\geq 1}$ of functions from some set to a metric space; moreover, there are companion notions of a *pointwise Cauchy* and of a *uniformly Cauchy* sequence of such functions. We now describe briefly how these notions can be applied to sequences of schema mappings. In its most general formulation, a schema mapping $\mathcal{M}$ over a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$ is a set of pairs (I, J), where I is a finite $\mathbf{S}$-instance and J is a finite $\mathbf{T}$-instance. It follows that a schema mapping $\mathcal{M}$ can be also be viewed as a function f from the set $\mathrm{Inst}(\mathbf{S})$ of all finite $\mathbf{S}$-instances to the powerset $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ of the set of all finite $\mathbf{T}$-instances, where $f(I) = \{J : (I, J) \in \mathcal{M}\}$. This way, a sequence $(\mathcal{M}_n)_{n\geq 1}$ of schema mappings over a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$ can be viewed as a sequence of functions from $\mathrm{Inst}(\mathbf{S})$ to the (pseudo)metric space $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$.

After the conceptual framework has been laid out, we study in depth the limiting behavior of sequences of GAV mappings and the convergence of sequences of LAV mappings. We establish a number of technical results that reveal rather dramatic and perhaps unanticipated differences between GAV schema mappings and LAV schema mappings.

For sequences of GAV mappings, we point out that every uniformly Cauchy sequence of GAV mappings is eventually constant, hence it has a GAV mapping as uniform limit. We also show that every pointwise Cauchy sequence of GAV mappings has a pointwise limit, but it need not have a uniform limit; moreover, there are pointwise Cauchy sequences of GAV mappings such that no GAV mapping is their pointwise limit. This raises the question as to when a sequence of GAV mapping has a GAV mapping as a pointwise limit. We prove that a sequence of GAV mappings has a GAV mapping as a pointwise limit if and only if it has a pointwise limit that allows for CQ-rewriting[1].

For sequences of LAV mappings, we show that the notions of uniform limit and pointwise limit coincide; moreover, the same holds true for the notions of uniformly Cauchy and pointwise Cauchy sequences. However, there are uniformly Cauchy sequences of LAV mappings that have no uniform limit. We also establish that a uniformly Cauchy sequence of LAV mappings has a LAV mapping as a uniform limit if and only if it has a uniform limit that admits universal solutions. The aforementioned results lift to sequences of *premise-bounded* sequences of GLAV mappings, i.e., sequences of GLAV mappings for which there is a $k \geq 1$ such that, for every mapping in the sequence, the left-hand side of every GLAV constraint has at most k source atoms (LAV mappings have $k = 1$).

In terms of techniques, we use systematically the structural characterizations of schema-mapping languages established in [19], thus creating a link with a different line of research.

The metric space $(\mathcal{P}(\text{Inst}(\mathbf{T})), \text{dist})$ is incomplete, i.e., there are Cauchy sequences of elements of $\mathcal{P}(\text{Inst}(\mathbf{T}))$ that have no limit in $\mathcal{P}(\text{Inst}(\mathbf{T}))$. It is well known that every incomplete metric space (X, d) has a completion, which means that it can be embedded into a complete metric space (X^*, d^*) so that X is a dense subset of X^*. Moreover, pointwise (respectively, uniformly) Cauchy sequences of functions on X have pointwise (respectively, uniform) limits that take values in X^*. The construction of X^* from X involves equivalence classes of Cauchy sequences of elements of X, thus, in general, the members of X^* do not have a concrete representation. In the last part of the paper, we show that the members of $\mathcal{P}(\text{Inst}(\mathbf{T}))^*$ can be represented by suitably constructed infinite $\mathbf{T}$-instances. As a consequence of this, the pointwise (respectively, uniform) limits of Cauchy sequences of schema mappings can be represented by *generalized* schema mappings, i.e., schema mappings that allow for infinite target instances as solutions to finite source instances.

2 Preliminaries

This section contains a minimum amount of the necessary background material.

Schemas, Instances, and Conjunctive Queries. A *schema* $\mathbf{R}$ is a finite sequence $\langle R_1, \ldots, R_k \rangle$ of relation symbols, where each R_i has a fixed arity. An *instance* I over $\mathbf{R}$, or an $\mathbf{R}$-*instance*, is a sequence $(R_1^I, \ldots, R_k^I)$, where each R_i^I is a finite relation of the same arity as R_i. We will often use R_i to denote both the relation symbol and the relation R_i^I that interprets it. The *active domain* of an instance I is the set of all values occurring in the relations of I. A *fact* of an instance I (over $\mathbf{R}$) is an expression $R_i^I(a_1, \ldots, a_m)$ (or simply $R_i(v_1, \ldots, v_m)$), where R_i is a relation symbol of $\mathbf{R}$ and $(a_1, \ldots, a_m) \in R_i^I$.

A *conjunctive query* is a first-order formula of the form $\exists \mathbf{z}\, \theta(\mathbf{x}, \mathbf{z})$, where $\theta(\mathbf{x}, \mathbf{z})$ is a conjunction of atomic formulas $R_i(v_1, ..., v_m)$ and each v_j is one of the variables in $\mathbf{x}$ and $\mathbf{z}$. A *boolean conjunctive query* is a conjunctive query with no free variables. We write CQ for the class of all conjunctive queries over some schema. For every $n \geq 1$, we let CQ_n denote the class of all conjunctive queries with at most n variables. We also let CQ_0 denote the singleton consisting of a trivially true query.

Schema Mappings, Universal Solutions, Certain Answers. Motivated by the terminology in data exchange [6], we typically work with two schemas, a *source schema* $\mathbf{S}$ and a *target*

[1] Allowing for CQ-rewriting means that the certain answers of every conjunctive query over the target schema is definable by a union of conjunctive queries over the source schema - see [19].

schema $\mathbf{T}$ with no relation symbols in common. We refer to $\mathbf{S}$-instances as *source instances*, and to $\mathbf{T}$-instances as *target instances*. We assume the presence of two kinds of values in instances, namely *constants* and *(labeled) nulls*. We also assume that the active domains of source instances consists of constants; the active domains of target instances may contain both constants and nulls.

A *schema mapping* $\mathcal{M}$ between a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$ is a set of pairs (I, J), where I is source instance and J a target instance. A schema mapping is often (but not always) given as a triple $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$, where Σ is a set of formulas in some suitable logical formalism such that $(I, J) \in \mathcal{M}$ if and only if $I \cup J \models \Sigma$.

Let $\mathcal{M}$ be a fixed schema mapping. In data exchange, the main problem is, given a source instance I, to find a *solution* for I w.r.t. $\mathcal{M}$, that is, a target instance J such that $(I, J) \in \mathcal{M}$ (or determine that no solution exists). We use the notation $\mathsf{Sol}(I, \mathcal{M}) = \{J \mid (I, J) \in \mathcal{M}\}$ to denote the set of all solutions for I w.r.t. $\mathcal{M}$. In data integration, the main problem is to compute the *certain answers* of queries [12]. Specifically, given a query q over the target schema and a source instance I, the *certain answers of* q *on* I *w.r.t.* $\mathcal{M}$ is the set

$$\mathrm{cert}(q, I, \mathcal{M}) = \bigcap \{q(J) \mid J \in \mathsf{Sol}(I, \mathcal{M})\}.$$

On the face of it, the definition of certain answers may entail computing an intersection of infinitely many sets. One of the main findings in [6] is that there is a notion of a "good" solution in data exchange, called *universal solution*, that can also be used to compute the certain answers of conjunctive queries in a much more direct way.

Let J_1 and J_2 be two target instances. A function h is a *homomorphism* from J_1 to J_2 if the following hold: (i) for every constant c, we have that $h(c) = c$; and (ii) for every relation symbol R in $\mathbf{R}$ and every tuple $(a_1, \dots, a_n) \in R^{J_1}$, we have that $(h(a_1), \dots, h(a_n)) \in R^{J_2}$. We write $J_1 \to J_2$ to denote that there is a homomorphism from J_1 to J_2. We say that J_1 is *homomorphically equivalent* to J_2, written $J_1 \leftrightarrow J_2$, if $J_1 \to J_2$ and $J_2 \to J_1$.

Let I be a source instance. A *universal* solution for I w.r.t. $\mathcal{M}$ is a solution J such that for every solution $J' \in \mathsf{Sol}(I, \mathcal{M})$, we have that $J \to J'$. Intuitively, a universal solution for I is a "most general" solution for I. We write $\mathsf{UnivSol}(I, \mathcal{M})$ to denote the set of all universal solutions for I w.r.t. $\mathcal{M}$ (universal solutions need not always exist). As shown in [6], if q is a conjunctive query, I is a source instance, and J is a universal solution for I w.r.t. $\mathcal{M}$, then $\mathrm{cert}(q, I, \mathcal{M}) = q(J)_{\downarrow}$, where $q(J)_{\downarrow}$ is the set of all null-free tuples in $q(J)$.

Structural Properties of Schema Mappings. We now present a number of structural properties that a schema mapping may or may not possess. These properties were investigated in their own right in [19], where they were used to obtain characterizations of schema-mapping languages that will be of great interest to us in this paper. Let $\mathcal{M}$ be a schema mapping.
$\mathcal{M}$ *allows for* CQ*-rewriting* if for every target conjunctive query q, there exists a union q' of source conjunctive queries such that $\mathrm{cert}(I, \mathcal{M}, q) = q'(I)$, for every source instance I.
$\mathcal{M}$ *admits universal solutions* if for every source instance I, there is a universal solution for I w.r.t. $\mathcal{M}$. We write $\mathrm{univ}(I, \mathcal{M})$ to denote some such universal solution.
$\mathcal{M}$ is *closed under target homomorphisms* if $(I, J) \in \mathcal{M}$ and $J \to J'$ imply that $(I, J') \in \mathcal{M}$.
$\mathcal{M}$ is *closed under unions* if $(I_1, J_1) \in \mathcal{M}$ and $(I_2, J_2) \in \mathcal{M}$ imply that $(I_1 \cup I_2, J_1 \cup J_2) \in \mathcal{M}$.
$\mathcal{M}$ is *closed under target intersections* if $J_1 \in \mathsf{Sol}(I, \mathcal{M})$ and $J_2 \in \mathsf{Sol}(I, \mathcal{M})$ imply that $(J_1 \cap J_2) \in \mathsf{Sol}(I, \mathcal{M})$.
$\mathcal{M}$ is *n-modular* if whenever $(I, J) \notin \mathcal{M}$, there is a subinstance $I' \subseteq I$ with at most n elements in its active domain such that $(I', J) \notin \mathcal{M}$ ("small counterexample").

Schema Mapping Languages. A GLAV *(global-and-local-as-view)* constraint is a first-order formula 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 the source schema $\mathbf{S}$, 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 the target schema $\mathbf{T}$ with variables in $\mathbf{x}$ and $\mathbf{y}$. We refer to $\varphi(\mathbf{x})$ as the *left-hand side*, or *premise*, and $\exists \mathbf{y} \psi(\mathbf{x}, \mathbf{y})$ as the *right-hand side*, or *conclusion* of the constraint. Another name for GLAV constraints is *source-to-target tuple-generating dependencies* or, in short, *s-t tgds*.

A LAV (*local-as-view*) constraint is a GLAV constraint whose left-hand side is a single atom over the source, while a GAV (*global-as-view*) constraint is a GLAV constraint whose right-hand side is a single atom over the target (in particular, the right-hand side contains no existential quantifiers). For example, $\forall x, y (E(x, y) \rightarrow \exists z (F(x, z) \wedge F(z, y)))$ is a LAV constraint, and $\forall x, y, z (E(x, z) \wedge E(z, y) \rightarrow F(x, y))$ is a GAV constraint.

A GLAV (*global-and-local-as=view*) mapping is a schema mapping $\mathcal{M} = (\mathbf{S}, \mathbf{T}, \Sigma)$ such that Σ is a finite set of GLAV constraints. The notions of a *LAV mapping* and of a *GAV mapping* are defined analogously.

Every GLAV mapping $\mathcal{M}$ admits universal solutions [6]; furthermore, given a source instance I, a *canonical universal solution* $\text{chase}(I, \mathcal{M})$ can be produced via the *oblivious chase procedure* as follows: whenever the antecedent of an s-t tgd in $\mathcal{M}$ becomes true, fresh null values are introduced and facts involving these nulls are added to $\text{chase}(I, \mathcal{M})$, so that the conclusion of the s-t tgd becomes true. Every GLAV mapping is also known to allow for CQ-rewriting and to be n-modular, for some $n \geq 1$. Moreover, every LAV mapping is closed under unions, while every GAV mapping is closed under target intersections.

Second-Order tgds, or *SO tgds*, were introduced in [8] and were shown to be exactly the constraints needed to express the composition of a finite number of GLAV mappings. Instead of giving the precise definition of an SO tgd, we illustrate this notion with an example from [8]. The formula $\exists f (\forall e (\mathit{Emp}(e) \rightarrow \mathit{Mgr}(e, f(e))) \wedge \forall e (\mathit{Emp}(e) \wedge (e = f(e)) \rightarrow \mathit{SelfMgr}(e)))$ expresses the property that every employee has a manager, and if an employee is the manager of himself/herself, then this employee is a self-manager. The above formula is an SO tgd that is not logically equivalent to any (finite or infinite) set of GLAV constraints [8].

Every SO tgd allows for CQ-rewriting and admits universal solutions; however, an SO tgd may not be closed under target homomorphisms and there may not exist any $n \geq 1$ such that the SO tgd is n-modular (see [8, 19]).

Pseudometric Spaces and Metric Spaces. A *pseudometric space* is a pair (X, d), where X is a set and d is a function from $X \times X$ to the set R^+ of non-negative real numbers with the following properties: (i) $d(x, x) = 0$, for every x in X; (ii) $d(x, y) = d(y, x)$, for every x and y in X; (iii) $d(x, y) \leq d(x, z) + d(y, z)$, for every x, y, z in X (triangle inequality). A *metric space* is a pseudometric space (X, d) such that if $d(x, y) = 0$, then $x = y$. It is easy to see that if (X, d) is a pseudometric space, then the relation $R_d = \{(x, y) \in X \times X \mid d(x, y) = 0\}$ is an equivalence relation on X. From this, it follows that every pseudometric space (X, d) gives rise to a metric space $(\widehat{X}, \widehat{d})$, where $\widehat{X}$ is the set of equivalence classes of elements of X modulo the equivalence relation R_d and $\widehat{d}([x], [y]) = d(x, y)$.

A sequence of elements $x_1, x_2, \ldots$ of X *converges* to an element x of X, denoted by $\lim\limits_{n \to \infty} x_n = x$, if for every $\epsilon > 0$, there is an integer n_0 such that $d(x_n, x) < \epsilon$, for every $n \geq n_0$. We say that x is the *limit* of this sequence (the limit is unique if (X, d) is a metric space). A sequence $x_1, x_2, \ldots$ of elements of X is *Cauchy* if for every $\epsilon > 0$, there is an integer n_0 such that $d(x_n, x_{n'}) < \epsilon$, for every $n, n' \geq n_0$.

Using the triangle inequality, it is easy to see that if a sequence of elements in a (pseudo)metric space has a limit, then the sequence is Cauchy. The converse, however, does not hold for arbitrary (pseudo)metric spaces. A (pseudo)metric space (X, d) is *complete* if every Cauchy sequence of elements of X has a limit in X; otherwise, it is *incomplete*.

It is well known that every incomplete (pseudo)metric space (X, d) can be embedded into a complete (pseudo)metric space (X^*, d^*), called the *completion* of (X, d), in such a way that X is a *dense* subset of X^*, i.e., every member of X^* is the limit of a sequence of members of X. The members of X^* are equivalence classes of Cauchy sequences of X, where two Cauchy sequences $x_1, x_2, ...$ and $y_1, y_2, \ldots$ of elements of X are *equivalent* if $\lim_{n\to\infty} d(x_n, y_n) = 0$, while the distance function d^* is defined as $d^*([x_1, x_2, \ldots], [y_1, y_2, \ldots]) = \lim_{n\to\infty} d(x_n, y_n)$. The proof of correctness of this construction can be found in [18] or any other book on metric spaces.

As a concrete example, the metric space of the real numbers is the completion of the metric space of the rational numbers (both with the standard distance).

3 Metric Space of Target Instances

To study the limits of sequences of schema mappings, we first introduce a pseudometric space of sets of target instances. By considering schema mappings as functions that map each source instance to the set of its solutions, we can view sequences of schema mappings as *sequences of functions*. The (pointwise or uniform) limit of a sequence of schema mappings is then simply defined in the standard way as the limit of a sequence of functions taking values in a pseudometric space. Moreover, by passing to the associated metric space of equivalence classes of sets of target instances, we ensure the uniqueness of the limit. If $\mathbf{T}$ is a schema, we write $\text{Inst}(\mathbf{T})$ for the set of all finite instances of $\mathbf{T}$. We also write $\mathcal{P}(\text{Inst}(\mathbf{T}))$ for the power set of $\text{Inst}(\mathbf{T})$. The notion of distance on $\mathcal{P}(\text{Inst}(\mathbf{T}))$ that we are about to introduce is heavily based on the notion of the certain answers to conjunctive queries and on the idea that two members $\mathcal{J}$ and $\mathcal{J}'$ of $\mathcal{P}(\text{Inst}(\mathbf{T}))$ are "close" to each other if only "big" conjunctive queries can yield different certain answers on $\mathcal{J}$ and $\mathcal{J}'$.

▶ **Definition 1.** Let q be a query over a schema $\mathbf{T}$ and let $\mathcal{J}$ be a member of $\mathcal{P}(\text{Inst}(\mathbf{T}))$. The *certain answers* of q over $\mathcal{J}$ are defined as $\text{cert}(q, \mathcal{J}) = \bigcap\{q(J) \mid J \in \mathcal{J}\}$.

We say that two sets of instances $\mathcal{J}$ and $\mathcal{J}'$ in $\mathcal{P}(\text{Inst}(\mathbf{T}))$ are CQ-*equivalent*, denoted $\mathcal{J} \equiv_{\mathsf{CQ}} \mathcal{J}'$, if $\text{cert}(q, \mathcal{J}) = \text{cert}(q, \mathcal{J}')$ for all conjunctive queries q.

We say that $\mathcal{J}$ and $\mathcal{J}'$ are CQ_n-*equivalent*, denoted $\mathcal{J} \equiv_{\mathsf{CQ}_n} \mathcal{J}'$, if $\text{cert}(q, \mathcal{J}) = \text{cert}(q, \mathcal{J}')$ for all conjunctive queries q with at most n variables (i.e., for all q in CQ_n.) ◁

▶ **Definition 2.** Let $\mathcal{J}$ and $\mathcal{J}'$ be two sets of instances in $\mathcal{P}(\text{Inst}(\mathbf{T}))$. The *similarity* $\text{sim}(\mathcal{J}, \mathcal{J}')$ and the *distance* $\text{dist}(\mathcal{J}, \mathcal{J}')$ between $\mathcal{J}$ and $\mathcal{J}'$ are defined as follows:

- $\text{sim}(\mathcal{J}, \mathcal{J}') = \max\{k \mid \mathcal{J} \equiv_{\mathsf{CQ}_k} \mathcal{J}'\}$;
- $\text{dist}(\mathcal{J}, \mathcal{J}') = 2^{-\text{sim}(\mathcal{J}, \mathcal{J}')}$. ◁

It is easy to verify that the pair $(\mathcal{P}(\text{Inst}(\mathbf{T})), \text{dist})$ is a pseudometric space; in fact, dist is an *ultrametric* distance function, that is, $\text{dist}(\mathcal{J}, \mathcal{J}') \leq \max\{\text{dist}(\mathcal{J}, \mathcal{J}''), \text{dist}(\mathcal{J}'', \mathcal{J}')\}$ holds for all $\mathcal{J}$, $\mathcal{J}'$, $\mathcal{J}''$ in $\mathcal{P}(\text{Inst}(\mathbf{T}))$. Moreover, $\text{dist}(\mathcal{J}, \mathcal{J}') = 0$ if and only if $\mathcal{J}$ and $\mathcal{J}'$ are CQ-equivalent. It is important to note that the pseudo-metric space $(\mathcal{P}(\text{Inst}(\mathbf{T})), \text{dist})$ is incomplete, i.e., there exist Cauchy sequences of elements of $\mathcal{P}(\text{Inst}(\mathbf{T}))$ that do not have a limit in $\mathcal{P}(\text{Inst}(\mathbf{T}))$. We first give an example of a sequence that has a limit in $\mathcal{P}(\text{Inst}(\mathbf{T}))$.

▶ **Example 3.** Let $\mathbf{T}$ be a schema consisting of a single binary relation and let C_m be the undirected cycle of length m, for $m \geq 1$. Consider the sequence $(\{C_{2n+1}\})_{n\geq 1}$ of singletons each containing a cycle of odd size. It is not hard to verify that $\lim_{n\to\infty} (\{C_{2n+1}\})_{n\geq 1} = \{C_2\}$. ◁

In contrast, there are Cauchy sequences of element of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ that have no limit.

▶ **Proposition 4.** *Let* $\mathbf{T}$ *be a schema consisting of a single binary relation and let* K_n *be the clique of size* n*, for* $n \geq 1$*. The sequence* $(\{K_n\})_{n\geq 1}$ *of singletons each containing a clique of different size is Cauchy, but has no limit in* $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$*.*

Proof. The sequence $(\{K_n\})_{n\geq 1}$ is Cauchy because if $m \geq n$, then K_m and K_n satisfy the same first-order sentences with n variables. To show that this sequence has no limit in $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$, assume that there is a set $\mathcal{J}$ of finite instances over $\mathbf{T}$ such that $\lim\limits_{n\to\infty} \{K_n\} = \mathcal{J}$. We distinguish two cases. If $\mathcal{J} = \varnothing$, then $\mathrm{cert}(q, \mathcal{J}) = \mathit{true}$, for every conjunctive query q. In contrast, $\mathrm{cert}(\exists x E(x,x), \{K_n\}) = \mathit{false}$, for every $n \geq 2$. If $\mathcal{J} \neq \varnothing$, consider a member J of $\mathcal{J}$. Let m be the biggest integer such that J contains a clique of size m, and let $\exists K_{m+1}$ be the conjunctive query asserting that there is a clique of size $m+1$. We now have that $\mathrm{cert}(\exists K_{m+1}, \mathcal{J}) = \mathit{false}$, while $\mathrm{cert}(\exists K_{m+1}, \{K_n\}) = \mathit{true}$, for every $n \geq m+1$. ◀

Since $(\{K_n\})_{n\geq 1}$ is a Cauchy sequence, it has a limit in the completion of $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$. A concrete representation of this limit is the singleton $\{K_\infty\}$, where K_∞ is the infinite clique. In Section 6, we will examine the completion of $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$ more closely.

The following definitions are perfectly meaningful for every pseudometric space (X, d) and for every sequence of functions taking values in X. For concreteness, we give the definitions for sequences of functions taking values in $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$.

▶ **Definition 5.** Let A be a set, let $(f_n)_{n\geq 1}$ be a sequence of functions from A to $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$, and let f be a function from A to $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$.

- We say that $(f_n)_{n\geq 1}$ *converges pointwise* to f, denoted as $\overset{p}{\lim\limits_{n\to\infty}} f_n = f$, if for every element $x \in A$, we have that $\lim\limits_{n\to\infty} f_n(x) = f(x)$.
- We say that $(f_n)_{n\geq 1}$ *converges uniformly* to f, denoted as $\overset{u}{\lim\limits_{n\to\infty}} f_n = f$, if for every $\epsilon > 0$, there exists an integer $n_0 \geq 1$ such that for every integer $n \geq n_0$ and for every element $x \in A$, we have $\mathrm{dist}(f_n(x), f(x)) < \epsilon$.
- We say that $(f_n)_{n\geq 1}$ is *pointwise Cauchy*, if for every element $x \in A$, the sequence $(f_n(x))_{n\geq 1}$ is Cauchy.
- We say that $(f_n)_{n\geq 1}$ is *uniformly Cauchy*, if for every $\epsilon > 0$, there exists an integer $n_0 \geq 1$ such that for all integers $n, n' \geq n_0$ and for every element $x \in A$, we have $\mathrm{dist}(f_n(x), f_{n'}(x)) < \epsilon$. ◁

Clearly, if $(f_n)_{n\geq 1}$ converges pointwise (resp., uniformly), then $(f_n)_{n\geq 1}$ is pointwise (resp., uniformly) Cauchy. The converse is not in general true for arbitrary (pseudo)metric spaces; in particular, it is not true for the pseudometric space $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$.

We now bring schema mappings into the picture. Every schema mapping $\mathcal{M}$ over a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$ can be identified with a function $f\colon \mathrm{Inst}(\mathbf{S}) \longrightarrow \mathcal{P}(\mathrm{Inst}(\mathbf{T}))$, where $f(I) = \mathsf{Sol}(I, \mathcal{M})$ (recall that $\mathsf{Sol}(I, \mathcal{M})$ is the set of all solutions of I w.r.t. $\mathcal{M}$, i.e., the set of all finite $\mathbf{T}$ instances J such that $(I, J) \in \mathcal{M}$). Thus, a sequence $(\mathcal{M}_n)_{n\geq 1}$ of schema mappings over a source schema $\mathbf{S}$ and target schema $\mathbf{T}$ can be viewed as a sequence of functions from $\mathrm{Inst}(\mathbf{S})$ to $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$. Therefore, we can talk about a sequence of schema mappings being pointwise Cauchy and uniformly Cauchy if the sequence of the associated functions has these properties. Similarly, we say that a sequence of schema mappings has a pointwise limit (resp., a uniform limit) if the sequence of the associated functions converges pointwise (resp., converges uniformly) to a schema mapping.

The preceding notion of convergence of a sequence of schema mappings allows us to draw a connection to earlier work on schema mapping optimization [5, 7]. Here, we are considering CQ-equivalence and CQ_n-equivalence of *sets of instances*. In previous works, these notions of equivalence have been mainly applied to schema mappings (see, e.g., [5, 7, 14]). Specifically, two schema mappings $\mathcal{M}, \mathcal{M}'$ are CQ-equivalent (resp., CQ_n-equivalent) if for every target conjunctive query q (resp., every target conjunctive query q in CQ_n) and every source instance I, we have that $\mathrm{cert}(q, I, \mathcal{M}) = \mathrm{cert}(q, I, \mathcal{M}')$. In this case, we write $\mathcal{M} \equiv_{\mathsf{CQ}} \mathcal{M}'$ (resp., $\mathcal{M} \equiv_{\mathsf{CQ}_n} \mathcal{M}'$). The notion of CQ_n-equivalence has been studied in the context of schema mapping optimization [5, 7]. Below we discuss its relationship to the convergence of schema mappings.

▶ **Proposition 6.** *Consider a sequence $(\mathcal{M}_n)_{n\geq 1}$ of schema mappings and a schema mapping $\mathcal{M}$. Then $\overset{u}{\lim\limits_{n\to\infty}} \mathcal{M}_n = \mathcal{M}$ if and only if for every integer $k \geq 1$, there is an integer $n_0 \geq 1$ such that for all integers $n \geq n_0$, we have that $\mathcal{M}_n \equiv_{\mathsf{CQ}_k} \mathcal{M}$.* ◁

Intuitively, the preceding proposition states that it takes bigger and bigger conjunctive queries to distinguish the members of a sequence $(\mathcal{M}_n)_{n\geq 1}$ from its uniform limit.

Although never explicitly introduced, the notion of uniform convergence was implicit in [5], where it was shown that for every SO tgd σ and for every $n \geq 1$, there is a GLAV mapping $\mathcal{M}_n$ such that $\sigma \equiv_{\mathsf{CQ}_n} \mathcal{M}_n$. From this, it is easy to see that $\overset{u}{\lim\limits_{n\to\infty}} \mathcal{M}_n = \sigma$. Thus, we have the following result.

▶ **Theorem 7** (implicit in [5]). *Every SO tgd is a uniform limit of a sequence of GLAV mappings.*

There are SO tgds that are not CQ-equivalent to any GLAV mapping [7]. Thus, the point of Theorem 7 is that SO tgds can be "approximated" up to any level of CQ_k-equivalence by GLAV mappings, which are syntactically simpler and generally more well-behaved.

As stated earlier, $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$ is a pseudometric space since it cannot distinguish CQ-equivalent sets of instances. Consequently, the limit of a sequence of sets of instances and the (uniform or pointwise) limit of a sequence of mappings need not be unique. However, the limit is unique up to CQ-equivalence and, as described in Section 2, there is an associated metric space $(\widehat{\mathcal{P}(\mathrm{Inst}(\mathbf{T}))}, \widehat{\mathrm{dist}})$ obtained by considering the equivalence classes of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ modulo the equivalence relation R_{dist}, where $(\mathcal{J}, \mathcal{J}') \in R_{\mathrm{dist}}$ if and only if $\mathrm{dist}(\mathcal{J}, \mathcal{J}') = 0$ (i.e., if and only if $\mathcal{J} \equiv_{\mathsf{CQ}} \mathcal{J}'$).

In subsequent sections, we will work with the metric space $(\widehat{\mathcal{P}(\mathrm{Inst}(\mathbf{T}))}, \widehat{\mathrm{dist}})$. Moreover, we will be interested in schema mappings modulo CQ-equivalence, which means that from now on we will view schema mappings as functions from source instances to equivalence classes of sets of target instances modulo CQ-equivalence. However, for notational simplicity, we will work each time with representatives of the equivalence classes. By a slight abuse of notation, we will write $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$, instead of $(\widehat{\mathcal{P}(\mathrm{Inst}(\mathbf{T}))}, \widehat{\mathrm{dist}})$. Likewise, we will not explicitly distinguish between a schema mapping $\mathcal{M}$ and the equivalence class of schema mappings that are CQ-equivalent to $\mathcal{M}$.

4 Limits of Sequences of GAV Mappings

Our goal in this section is to analyze sequences of GAV mappings. To this effect, we first investigate the existence of limits of such sequences and then examine the definability of limits. As discussed in Section 3, if a sequence $(\mathcal{M}_n)_{n\geq 1}$ of schema mappings has a pointwise (resp., uniform) limit, then the sequence is pointwise (resp., uniformly) Cauchy. The next result asserts that the converse holds for sequences of GAV mappings.

▶ **Theorem 8.** *Let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of GAV mappings.*

- *If $(\mathcal{M}_n)_{n\geq 1}$ is pointwise Cauchy, then it has a pointwise limit.*
- *If $(\mathcal{M}_n)_{n\geq 1}$ is uniformly Cauchy, then it is eventually constant and thus has a GAV schema mapping as a uniform limit.*

Proof Sketch. For showing the first claim, assume that $(\mathcal{M}_n)_{n\geq 1}$ is a pointwise Cauchy sequence of schema mappings and let I be a source instance. For each $n \geq 1$, consider the universal solution chase$(I, \mathcal{M}_n)$ for I w.r.t. $\mathcal{M}_n$ obtained by using the oblivious chase procedure. Since each $\mathcal{M}_n$ is a GAV schema mapping, we have that chase$(I, \mathcal{M}_n)$ contains no nulls. It can be shown that there exists some m_I s.t. for all $n \geq m_I$, we have that chase$(I, \mathcal{M}_n) =$ chase$(I, \mathcal{M}_{m_I})$. In other words, the sequence $(\text{chase}(I, \mathcal{M}_n))_{n\geq 1}$ is eventually constant (does not oscillate). Then the schema mapping $\mathcal{M} = \{(I, \text{chase}(I, \mathcal{M}_{m_I})) \mid I \text{ is a source instance}\}$ is a pointwise limit of the sequence $(\mathcal{M}_n)_{n\geq 1}$.

For showing the second claim, assume that $(\mathcal{M}_n)_{n\geq 1}$ is a uniformly Cauchy sequence of GAV mappings. We claim that $(\mathcal{M}_n)_{n\geq 1}$ is eventually constant, i.e., there is some m such that for all $n \geq m$, $\mathcal{M}_n \equiv_{\mathsf{CQ}} \mathcal{M}_m$ holds. Towards a contradiction, assume that for every m there exists an $i > m$ such that $\mathcal{M}_i \not\equiv_{\mathsf{CQ}} \mathcal{M}_m$. That is, for some source instance I, it is the case that chase$(I, \mathcal{M}_m) \neq$ chase$(I, \mathcal{M}_i)$. Since neither chase$(I, \mathcal{M}_m)$ nor chase$(I, \mathcal{M}_i)$ contain nulls, they can be distinguished using atomic queries from CQ_k, where k is the maximum relation arity of the target schema. Since this is the case for an arbitrarily large m, it follows that $(\mathcal{M}_n)_{n\geq 1}$ is not a uniformly Cauchy sequence, a contradiction. ◀

Next, we point out that even simple sequences of GAV schema mappings may have no uniform limit.

▶ **Proposition 9.** *There exists a sequence of GAV mappings that has a pointwise limit but no uniform limit.*

Proof. For every $n \geq 2$, let $\exists K_n$ be the boolean conjunctive query asserting that there is a clique of size n, i.e., $\exists K_n$ is the expression $\exists x_1, \ldots x_n \bigwedge_{i \neq j}(E(x_i, x_j) \wedge E(x_j, x_i))$.

Let $(\mathcal{M}_n)_{n\geq 1}$ be the sequence of GAV mappings, where $\mathcal{M}_n$ is specified by the constraint $\forall x(P(x) \wedge \exists K_{n+1} \rightarrow P'(x))$. Intuitively, $\mathcal{M}_n$ is a "copy" schema mapping, but the copying action is triggered only if E contains a clique of size $n+1$. One can show that the GAV schema mapping $\mathcal{M} = \{\forall x \forall y(P(x) \wedge E(y, y) \rightarrow P'(x))\}$ is a pointwise limit of $(\mathcal{M}_n)_{n\geq 1}$, but that this pointwise limit is not a uniform limit of $(\mathcal{M}_n)_{n\geq 1}$ and thus no uniform limit of $(\mathcal{M}_n)_{n\geq 1}$ exists.

To see that $\mathcal{M}$ is a pointwise limit of $(\mathcal{M}_n)_{n\geq 1}$, note that for source instances with a self-loop $E(a, a)$ for some a, $\mathcal{M}$ is indistinguishable from every element $\mathcal{M}_i \in (\mathcal{M}_n)_{n\geq 1}$. For source instances without such a self-loop, $\mathcal{M}$ coincides with all members of $(\mathcal{M}_n)_{n\geq 1}$ with an index exceeding the size of the maximal clique in I.

Now towards a contradiction assume that $\mathcal{M}$ is also a uniform limit. Then, there must be an n_0 such that for all $n \geq n_0$, the equivalence $\mathcal{M}_n \equiv_{\mathsf{CQ}_1} \mathcal{M}$ holds. However, taking $n = n_0$ and a source instance $I = K_n \cup \{P(c)\}$, one can observe that a target CQ_1 query $q = \exists x\, P'(x)$ witnesses $\mathcal{M}_n \not\equiv_{\mathsf{CQ}_1} \mathcal{M}$, since I contains no self-loop and thus $\mathsf{UnivSol}(I, \mathcal{M}) = \{\emptyset\}$. ◀

Proposition 9 and Theorem 8 imply that the sequence of GAV mappings in the proof of Proposition 9 is an example of a pointwise Cauchy sequence that is not uniformly Cauchy. More importantly, Theorem 8 gives rise to the following natural question concerning the definability of limits: if a sequence of GAV mappings has a pointwise limit, does it have a GAV mapping as such a limit? We answer this question in the negative: even the much richer language of SO tgds cannot express pointwise limits of sequences of GAV mappings.

▶ **Proposition 10.** *There is a pointwise Cauchy sequence of GAV schema mappings such that no SO tgd is a pointwise limit of that sequence.*

Proof Idea. For every $n \geq 1$, let $P_n(x, y)$ be the conjunctive query expressing the property "there is an E-path of length n from x to y", and let $\mathcal{M}_n$ be the GAV mapping specified by the set $\{\forall x, y(P_i(x, y) \rightarrow F(x, y)) \mid 1 \leq i \leq n\}$. The schema mapping

$$\mathcal{M}^\star = \{(I, J) \mid F^J \text{ contains the transitive closure } TC(I) \text{ of } E^I\}$$

is a pointwise limit of the sequence $(\mathcal{M}_n)_{n\geq1}$. However, note that $\mathcal{M}^\star$ is not CQ-equivalent to any schema mapping $\mathcal{M}'$ that allows for CQ-rewriting: if it were, then there would exist a union q of conjunctive queries over the source such that, for every source instance I, $\text{cert}(F(x, y), I, \mathcal{M}^\star) = TC(I) = \text{cert}(F(x, y), I, \mathcal{M}') = q(I)$. Consequently, the transitive closure of I would be first-order definable over the source, which is not the case. Since every SO tgd allows for CQ-rewriting, no SO tgd is a pointwise limit of the sequence $(\mathcal{M}_n)_{n\geq1}$. ◀

We have just seen that there are sequences of GAV mappings that have a pointwise limit, but no such limit is definable by a GAV mapping. This raises the question of finding necessary and sufficient conditions guaranteeing that a sequence of GAV mappings has a GAV mapping as a pointwise limit. The next result provides an answer to this question.

▶ **Theorem 11.** *Let $(\mathcal{M}_n)_{n\geq1}$ be a pointwise Cauchy sequence of GAV mappings. The following statements are equivalent:*

1. *$(\mathcal{M}_n)_{n\geq1}$ has a GAV mapping as a pointwise limit.*
2. *$(\mathcal{M}_n)_{n\geq1}$ has a pointwise limit that allows for CQ-rewriting.*

Proof Idea. Let $(\mathcal{M}_n)_{n\geq1}$ be a pointwise Cauchy sequence of schema mappings. As seen in the proof sketch of Theorem 8, for every source instance I, there is a positive integer m_I, such that for all $n \geq m_I$ the equality $\text{chase}(I, \mathcal{M}_{m_I}) = \text{chase}(I, \mathcal{M}_n)$ holds for the respective elements $\mathcal{M}_{m_I}$ and $\mathcal{M}_n$ of $(\mathcal{M}_n)_{n\geq1}$. Moreover, the schema mapping $\mathcal{M} = \{(I, \text{chase}(I, \mathcal{M}_{m_I}) \mid I \text{ is a source instance}\}$ is a pointwise limit of $(\mathcal{M}_n)_{n\geq1}$, and so is the CQ-equivalent mapping $\mathcal{M}^\star = \{(I, J) \mid \text{chase}(I, \mathcal{M}_{m_I}) \subseteq J\}$. The result we seek is an immediate consequence of the fact that the following four statements are equivalent:

(a) $(\mathcal{M}_n)_{n\geq1}$ has a GAV mapping as a pointwise limit.
(b) $(\mathcal{M}_n)_{n\geq1}$ has a pointwise limit that allows for CQ-rewriting.
(c) $\mathcal{M}^\star$ allows for CQ-rewriting.
(d) $\mathcal{M}^\star$ is logically equivalent to a GAV mapping.

The proof uses Theorem 3.2 in [19], which asserts that a schema mapping is logically equivalent to a GAV schema mapping if and only if it allows for CQ-rewriting, admits universal solutions, and is closed under both target homomorphisms and target intersections. ◀

▶ **Corollary 12.** *Let $(\mathcal{M}_n)_{n\geq1}$ be a pointwise Cauchy sequence of GAV mappings. The following statements are equivalent:*

1. *$(\mathcal{M}_n)_{n\geq1}$ has a GAV mapping as a pointwise limit.*
2. *$(\mathcal{M}_n)_{n\geq1}$ has an SO tgd as a pointwise limit.*

Proposition 10 and Theorem 11 yield a fairly complete picture of the definability of pointwise limits of GAV mappings. Specifically, there are two mutually exclusive possibilities:

1. No pointwise limit allows for CQ-rewriting and no GAV mapping is a pointwise limit.
2. Every pointwise limit admits CQ-rewriting and there is a GAV mapping that is a pointwise limit. Moreover, this happens precisely when the schema mapping $\mathcal{M}^\star$ in the proof of Theorem 11 allows for CQ-rewriting or, equivalently, when $\mathcal{M}^\star$ is logically equivalent to a GAV mapping.

5 Limits of Sequences of LAV Mappings

In this section, we investigate the existence and definability of limits of sequences of LAV mappings. In fact, we will consider a much broader class of GLAV mappings than LAV, which we call *premise-bounded* GLAV mappings. LAV mappings are the special case of this class when the premise bound is equal to one.

▶ **Definition 13.** Let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of GLAV mappings. We say that $(\mathcal{M}_n)_{n\geq 1}$ is *premise-bounded* if there exists an integer k such that for every element $\mathcal{M}_n$ of $(\mathcal{M}_n)_{n\geq 1}$, the premise of every constraint in $\mathcal{M}_n$ has at most k atoms.

Unlike the case of GAV mappings, the notions of pointwise Cauchy and uniformly Cauchy sequences of premise-bounded GLAV mappings coincide. Moreover, the same holds true for the notions of pointwise limit and uniform limit of sequences of such schema mappings.

▶ **Theorem 14.** *Let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of premise-bounded GLAV mappings.*

1. *The sequence $(\mathcal{M}_n)_{n\geq 1}$ is pointwise Cauchy if and only if it is uniformly Cauchy.*
2. *The sequence $(\mathcal{M}_n)_{n\geq 1}$ has a pointwise limit if and only if it has a uniform limit.*

The following two propositions further demarcate the differences between GAV and premise-bounded mappings. In fact, these differences are already witnessed by sequences of LAV mappings. The first difference concerns the existence of limits of uniformly Cauchy sequences. In contrast to the GAV case, uniformly Cauchy sequences of LAV mappings may have no uniform limit; in fact, they may not even have a pointwise limit.

▶ **Proposition 15.** *There exists a uniformly Cauchy sequence of LAV mappings that has no pointwise limit; in particular, it has no uniform limit either.*

Proof Idea. For every $n \geq 1$, let $\mathcal{M}_n$ be the LAV mapping specified by the constraint $\forall x, y(E(x, y) \rightarrow \exists K_{n+1})$, where, as earlier, $\exists K_{n+1}$ is the boolean conjunctive query asserting that there is a clique of size $n + 1$. Using an argument similar to the one in the proof of Proposition 4, it can be shown that the sequence $(\mathcal{M}_n)_{n\geq 1}$ has no pointwise limit. ◀

The next difference is the definability of uniform limits. In Section 4, we saw that if a sequence of GAV mappings has a uniform limit, then it is eventually constant, hence it has a GAV mapping as a uniform limit. This property need not hold for sequences of LAV mappings (hence, it need not hold for sequences of premise-bounded schema mappings).

▶ **Proposition 16.** *There exists a sequence $(\mathcal{M}_n)_{n\geq 1}$ of LAV mappings that has a uniform limit, but no uniform limit of $(\mathcal{M}_n)_{n\geq 1}$ admits universal solutions. In particular, no SO tgd is a uniform limit of the sequence $(\mathcal{M}_n)_{n\geq 1}$.*

Proof Idea. For every $n \geq 1$, let $\mathcal{M}_n$ be the LAV mapping specified by the constraint $\forall x(V(x) \rightarrow \exists P_n)$, where $\exists P_n$ is a boolean CQ asking for a path of length n in the target instance. We argue that the mapping $\mathcal{M} = \{(\emptyset, \emptyset)\} \cup \{(I, C_k) \mid I \text{ non-empty and } k > 1\}$ is the uniform limit of $(\mathcal{M}_n)_{n\geq 1}$, and that $\mathcal{M}$ does not admit universal solutions. ◀

By Theorem 7, every SO tgd is the uniform limit of a sequence of GLAV mappings. Proposition 16 implies that the converse is false, even for sequences of LAV mappings.

In the previous section, we showed that a sequence of GAV mappings has a GAV mapping as a pointwise limit if and only if it has a pointwise limit that allows for CQ-rewriting. Is there some structural property that characterizes when a sequence of premise-bounded GLAV mappings has a GLAV mapping as a pointwise limit (which, for premise-bounded mappings, is the same as a uniform limit)? We will show that the property of admitting universal solutions is the key to this question. Specifically, we have the following result.

▶ **Theorem 17.** *Let $(\mathcal{M}_n)_{n\geq1}$ be a premise-bounded sequence of GLAV mappings. The following statements are equivalent.*

1. *$(\mathcal{M}_n)_{n\geq1}$ has a GLAV mapping $\mathcal{M}$ as a uniform limit.*
2. *$(\mathcal{M}_n)_{n\geq1}$ has a uniform limit that admits universal solutions.*

Moreover, if $(\mathcal{M}_n)_{n\geq1}$ is a sequence of LAV mappings, then $(\mathcal{M}_n)_{n\geq1}$ has a LAV mapping as a uniform limit if and only $(\mathcal{M}_n)_{n\geq1}$ has a uniform limit that admits universal solutions.

Proof (Hint). The direction (1) ⇒ (2) is obvious. For the direction (2) ⇒ (1), we start with the case when $(\mathcal{M}_n)_{n\geq1}$ is a sequence of LAV mappings. As stepping stones to the proof, the following lemmas can be used, which are of interest in their own right.

▶ **Lemma 18.** *If $\mathcal{M}$ is the uniform limit of a sequence $(\mathcal{M}_n)_{n\geq1}$ of schema mappings each of which allows for CQ-rewriting, then also $\mathcal{M}$ allows for CQ-rewriting.*

▶ **Lemma 19.** *Let $\mathcal{M}$ be a uniform limit of a sequence $(\mathcal{M}_n)_{n\geq1}$ of LAV mappings. If $\mathcal{M}$ admits universal solutions, then it is closed under unions.*

Assume that $\mathcal{M}$ is a uniform limit of a sequence $(\mathcal{M}_n)_{n\geq1}$ of LAV mappings and that $\mathcal{M}$ admits universal solutions. Since the notion of limit is based on CQ-equivalence, we may assume w.l.o.g. that $\mathcal{M}$ is closed under target homomorphisms. Then $\mathcal{M}$ has the following properties: $\mathcal{M}$ admits universal solutions; $\mathcal{M}$ allows for CQ-rewriting (Lemma 18); $\mathcal{M}$ is closed under target homomorphisms; $\mathcal{M}$ is closed under unions (by Lemma 19). From Theorem 3.1 in [19], it follows that $\mathcal{M}$ is logically equivalent to a LAV mapping.

For the case when $(\mathcal{M}_n)_{n\geq1}$ is a sequence of premise-bounded GLAV mappings (but not necessarily LAV mappings), we apply yet another structural characterization theorem from [19], namely Theorem 3.9, which asserts that if a schema mapping allows for CQ-rewriting, admits universal solutions, is closed under target homomorphisms, and is n-modular, for some fixed n, then it is logically equivalent to a GLAV mapping. Using machinery similar to the one used for the closure under unions in Lemma 19, it can be shown that the uniform limit $\mathcal{M}$ of the sequence $(\mathcal{M}_n)_{n\geq1}$ is n-modular for some fixed $n \geq 1$. The other structural properties are handled as in the case of a sequence of LAV mappings. ◀

We conclude this section with a conjecture concerning uniform limits of arbitrary sequences of GLAV mappings.

▶ **Conjecture 20.** *The following statements are equivalent for a sequence $(\mathcal{M}_n)_{n\geq1}$ of GLAV mappings.*

1. *$(\mathcal{M}_n)_{n\geq1}$ has an SO tgd as a uniform limit.*
2. *$(\mathcal{M}_n)_{n\geq1}$ has a uniform limit that admits universal solutions.*

It is not hard to show that the preceding conjecture is implied by a conjecture in [2] to the effect that the language of *plain*[2] SO-tgds can be characterized by the following three properties: allowing for CQ-rewriting, admitting universal solutions, and closure under target homomorphisms. It appears, however, that the technical tools needed to resolve the conjecture in [2] are not available at present.

6 Metric Space Completion and Generalized Schema Mappings

Let $\mathbf{T}$ be a schema containing a binary relation symbol. By Proposition 4, the metric space $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$ is not complete, i.e., there are Cauchy sequences of elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ that have no limit in $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$. Let $(\mathcal{P}(\mathrm{Inst}(\mathbf{T}))^*, \mathrm{dist}^*)$ be the completion of $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), \mathrm{dist})$. As described in Section 2, the elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))^*$ are the equivalence classes of Cauchy sequences of elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$, where two Cauchy sequences $\mathcal{I}_1, \mathcal{I}_2, \ldots$ and $\mathcal{J}_1, \mathcal{J}_2, \ldots$ are equivalent if $\lim_{n\to\infty} \mathrm{dist}(\mathcal{I}_n, \mathcal{J}_n) = 0$. Clearly, this is a rather abstract description of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))^*$. The main result of this section reveals that the elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))^*$ can be represented by suitably constructed infinite $\mathbf{T}$-instances. In turn, this result and basic results about complete metric spaces imply that the (pointwise or uniform) limits of a Cauchy sequence of schema mappings can be represented by a *generalized* schema mapping, that is, a schema mapping in which infinite solutions are allowed.

Let q be a conjunctive query with k free variables and let $\mathbf{a}$ be a k-tuple of constants. We write $q(\mathbf{a})$ to denote the instance K obtained by (i) substituting the free variables of q by the respective elements of $\mathbf{a}$; (ii) replacing the existential variables of q by fresh distinct labeled nulls; and (iii) treating the resulting body atoms of q as facts of the instance K.

We write $J_1 \uplus J_2$ to denote the *disjoint union* of two instances J_1 and J_2, that is, the instance obtained as a union of J_1 and J_2 with all labeled nulls renamed apart. If X is a set of instances, we write $\biguplus X$ to denote the disjoint union of all members of X. Note that we do not necessarily assume X to be finite; thus, $\biguplus X$ may be an infinite instance.

We are now ready to state the main result of this section and sketch its proof.

▶ **Theorem 21.** *Let $(\mathcal{J}_n)_{n\geq 1}$ be a Cauchy sequence of elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$. Then the limit of the sequence $(\mathcal{J}_n)_{n\geq 1}$ is the singleton $\mathbf{T}$-instance set $\mathcal{J}^*$, where*

$$\mathcal{J}^* = \Big\{ \biguplus \{ q(\mathbf{a}) \mid q \in \mathsf{CQ} \textit{ and there is an integer } p \textit{ such that } \mathbf{a} \in \mathrm{cert}(q, \mathcal{J}_i), \textit{ for every } i \geq p \} \Big\}.$$

Proof (Sketch). We have to show that $(\mathcal{J}_n)_{n\geq 1} \longrightarrow \mathcal{J}^*$, which means that for every integer $m \geq 1$ there exists an integer $n_0 \geq 1$ such that $\mathcal{J}_n \equiv_{\mathsf{CQ}_m} \mathcal{J}^*$, for all $n \geq n_0$.

By definition, $\mathcal{J}^*$ is a singleton; we write J to denote the single element of $\mathcal{J}^*$. The first crucial observation is that the set D of constants occurring in J is finite. To show this, we consider *single-atom* conjunctive queries, that is, queries of the form $\exists \mathbf{y} R(\mathbf{x}, \mathbf{y})$, where R is a relation symbol in the target schema $\mathbf{T}$. Clearly, every single-atom query has at most k variables, where k is the maximum arity of the relation symbols in $\mathbf{T}$.

Since the sequence $(\mathcal{J}_n)_{n\geq 1}$ is Cauchy, there exists an integer p_k such that $\mathcal{J}_i \equiv_{\mathsf{CQ}_k} \mathcal{J}_{p_k}$, for all $i \geq p_k$. This implies that the certain answers to single-atom conjunctive queries become fixed in $(\mathcal{J}_n)_{n\geq 1}$ starting from some integer p_k that depends only on the schema $\mathbf{T}$. By definition, the certain answers hold in every instance in $\mathcal{J}_{p_k}$; moreover, every instance in $\mathcal{J}_{p_k}$ is finite. Hence, the set D' of the certain answers to single-atom conjunctive queries

[2] A *plain* SO tgd is an SO tgd that contains no nested terms and no equalities. Every SO tgd is known to be CQ-equivalent to a plain one [2].

that eventually hold in $(\mathcal{J}_n)_{n\geq 1}$ is finite. To complete the proof of the finiteness, we show that the set D of constants occurring in the instance J is contained in D'. To see this, recall that J is composed of the bodies of conjunctive queries $q(\mathbf{a})$ such that $\mathbf{a} \in \mathrm{cert}(q, \mathcal{J}_n)$, for all sufficiently large n. Fix such a conjunctive query q with r atoms and consider its decomposition to single-atom queries $q_i(\mathbf{a}_i), \ldots, q_r(\mathbf{a}_r)$, where q_i has the i-th atom of q as its body and $\mathbf{a}_i$ contains the constants of $\mathbf{a}$ occurring in this atom. Observe that $\mathbf{a} \in \mathrm{cert}(q, \mathcal{J})$ implies $\mathbf{a}_i \in \mathrm{cert}(q_i, \mathcal{J})$, for every set $\mathcal{J}$ of instances. Consequently, the inclusion $D \subseteq D'$ holds, and thus D must be finite.

Now, given m, we need to provide n_0 such that for all $n \geq n_0$, we have that $\mathcal{J}_n \equiv_{\mathsf{CQ}_m} \mathcal{J}^*$ holds. In other words, n_0 has to big enough to ensure the equality $\mathrm{cert}(q, \mathcal{J}_n) = \mathrm{cert}(q, \mathcal{J}^*)$ for every conjunctive query $q \in \mathsf{CQ}_m$. In order to guarantee the inclusion $\mathrm{cert}(q, \mathcal{J}_n) \subseteq \mathrm{cert}(q, \mathcal{J}^*)$, it suffices to choose n_0 greater than the index n_1, starting from which all certain answers to CQ_m queries become fixed in $(\mathcal{J}_n)_{n\geq 1}$. Such an index n_1 exists since the sequence $(\mathcal{J}_n)_{n\geq 1}$ is Cauchy. To ensure $\mathrm{cert}(q, \mathcal{J}^*) \subseteq \mathrm{cert}(q, \mathcal{J}_n)$, we analyze the values of q in the limit instance J (recall that $\mathcal{J}^*$ is a singleton $\{J\}$). By the definition of J^*, atoms witnessing that $J \models q(\mathbf{b})$ stem from the bodies of conjunctive queries $q_1(\mathbf{a}_1), \ldots, q_\ell(\mathbf{a}_\ell)$. All these conjunctive queries hold in $(\mathcal{J}_n)_{n\geq 1}$ starting from some index. Inspecting finitely many conjunctive queries in CQ_m and all possible certain answers to them, one can choose n_0 large enough to ensure that $\mathrm{cert}(q, \mathcal{J}^*) \subseteq \mathrm{cert}(q, \mathcal{J}_n)$ as well. ◀

In their recent monograph [15], Nešetřil and Ossona de Mendez considered a notion of distance between instances, as well as sequences of instances and their limits. However, they considered a different setting and followed a different approach: first, they did not distinguish two classes of domain elements (constants and nulls) and, second, they heavily relied on a quasi-order on instances based on homomorphisms. The limit of a Cauchy sequence of instances is obtained in [15] via the concept of *ideal completion*. If $(\mathcal{J}_n)_{n\geq 1}$ is a Cauchy sequence of elements of $\mathcal{P}(\mathrm{Inst}(\mathbf{T}))$ such that all target instances appearing in this sequence contain only nulls (and no constants), then our description of the limit $\mathcal{J}*$ can be shown to be equivalent in the one in [15]; moreover, in this case, only boolean conjunctive queries contribute to the disjoint unions defining the limit.

We now recall two basic results about complete metric spaces.

▶ **Proposition 22.** *Let (Y, d) be a complete metric space and let $(f_n)_{n\geq 1}$ be a sequence of function from a set X to Y.*

- *If $(f_n)_{n\geq 1}$ is a pointwise Cauchy sequence, then $(f_n)_{n\geq 1}$ has a pointwise limit $f : X \to Y$, where $f(x) = \lim_{n\to\infty} f_n(x)$, for every $x \in X$.*
- *If $(f_n)_{n\geq 1}$ is a uniformly Cauchy sequence, then $(f_n)_{n\geq 1}$ has a uniform limit. Moreover, the pointwise limit $f : X \to Y$ of $(f_n)_{n\geq 1}$ is also the uniform limit of $(f_n)_{n\geq 1}$.*

The proof of the first part of Proposition 22 is immediate from the definitions; the proof of the second part can be found in any standard book on metric spaces (see, e.g., Proposition 3.6.6 in [18]). Note that the second part of Proposition 22 is known as the *Cauchy criterion*.

We are now ready to obtain concrete representations of the (pointwise or uniform) limits of Cauchy sequences of schema mappings.

▶ **Definition 23.** Let $\mathbf{S}, \mathbf{T}$ be two schemas. A *generalized schema mapping* is a set $\mathcal{M}$ of pairs (I, J) such that I is a finite $\mathbf{S}$-instance and J is a possibly infinite $\mathbf{T}$-instance.

▶ **Corollary 24.** *Let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of schema mappings. Consider the generalized schema mapping $\mathcal{M} = \big\{(I, J) \mid J = \biguplus\{q(\mathbf{a}) \mid q \in \mathsf{CQ}$ and $\exists p\ \forall i \geq p\ \mathbf{a} \in \mathrm{cert}(q, I, \mathcal{M}_i)\}\big\}$*

- *If $(\mathcal{M}_n)_{n\geq 1}$ is a pointwise Cauchy sequence, then the schema mapping $\mathcal{M}$ is the pointwise limit of $(\mathcal{M}_n)_{n\geq 1}$.*
- *If $(\mathcal{M}_n)_{n\geq 1}$ is a uniformly Cauchy sequence, then the schema mapping $\mathcal{M}$ is the uniform limit of $(\mathcal{M}_n)_{n\geq 1}$.*

Proof. The first part follows from Theorem 21 and the definition of pointwise convergence. The second part follows from the first part and Proposition 22. ◀

Finally, we consider (pointwise or uniformly) Cauchy sequences of schema mappings admitting universal solutions and obtain a different representation of their limits.

▶ **Corollary 25.** *Let $(\mathcal{M}_n)_{n\geq 1}$ be a pointwise Cauchy sequence of schema mappings over a source schema $\mathbf{S}$ and a target schema $\mathbf{T}$, each admitting universal solutions.*

1. *For every $I \in Inst(\mathbf{S})$, the sequence $(\mathsf{UnivSol}(I, \mathcal{M}_n))_{n\geq 1}$ is Cauchy, and hence it has a limit $\lim\limits_{n\to\infty}(\mathsf{UnivSol}(I, \mathcal{M}_n))$ in the complete metric space $(\mathcal{P}(\mathrm{Inst}(\mathbf{T}))^*, \mathrm{dist}^*)$.*
2. *The generalized schema mapping $\mathcal{M}^* = \{(I, J) \mid I \in Inst(\mathbf{S}), J \in \lim\limits_{n\to\infty}(\mathsf{UnivSol}(I, \mathcal{M}_n))\}$ is a pointwise limit of $(\mathcal{M}_n)_{n\geq 1}$. Moreover, if $(\mathcal{M}_n)_{n\geq 1}$ is a uniformly Cauchy sequence, then $\mathcal{M}^*$ is its uniform limit.*

7 Concluding Remarks

In this paper, we have embarked on a systematic study of the limiting behavior of sequences of schema mappings using concepts and tools from metric spaces. For the important special cases of GAV and LAV mappings, our main results are summarized in Figures 1 and 2.

In words, we have shown that, for GAV mappings, a pointwise Cauchy sequence need not be uniformly Cauchy; moreover, the existence of a pointwise limit does not imply the existence of a uniform limit. This cannot happen for LAV mappings. On the other side, a uniformly Cauchy sequence of LAV mappings need not even have a pointwise limit, which cannot happen for GAV mappings. We have also shown that structural properties of schema mappings can be used to characterize when the limit of a pointwise Cauchy sequence of GAV (or of LAV) mappings is equivalent to a GAV (or to a LAV) mapping. Finally, we have shown that infinite target instances and generalized mappings (i.e., schema mappings where target instances may be infinite) can be used to represent limits of Cauchy sequences of sets of target instances and limits of Cauchy sequences of arbitrary schema mappings.

We believe that the work reported here has laid the foundation for several interesting lines of subsequent investigations. We have seen that our results about sequences of LAV mappings extend in a natural way to sequences of premise-bounded GLAV mappings; an analogous extension of our results about sequences of GAV mappings to sequences of *conclusion-bounded* GLAV mappings is left for future work. We have also seen that there are sequences of LAV mappings for which no SO tgd is a uniform limit. Are there structural properties that characterize when a sequence of GLAV mappings has an SO tgd as a pointwise limit? In this vein, we have offered Conjecture 20. A related interesting open problem is whether schema mappings with target constraints are powerful enough to express pointwise limits or uniform limits of sequences of arbitrary GLAV schema mappings. We have some preliminary evidence that this is plausible, but much more work remains to be done.

We believe that the work reported in this paper provides a new perspective on the study of schema mappings by examining them from a dynamic viewpoint. As stated earlier, our original motivation came from schema-mapping optimization and, in particular, from the idea that “complex” schema mappings can be “approximated” by “simpler” ones. It remains to be

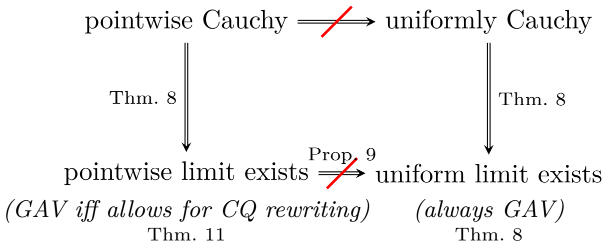

Figure 1 Overall picture for GAV schema mappings.

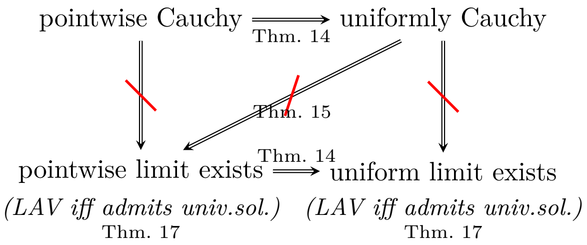

Figure 2 Overall picture for LAV schema mappings.

seen whether the work reported here will lead to applications to schema-mapping optimization. We believe, however, that the study of the limiting behavior of schema mappings via metric spaces is interesting in its own right.

We also note there are several areas in theoretical computer science where the study of limiting behavior of objects has produced results that were significant in their own right and also had fruitful consequences. For example, starting with the work of Fagin [4], there has been an extensive investigation of the asymptotic probabilities of logical properties and of 0-1 laws for various logics of interest in computer science. More recently, there has been a study of *profinite words*, which has found applications to automata theory and to the satisfiability problem for variants of monadic second-order logic (see, e.g., [17, 20]). Note that the profinite words form the completion of a metric space on words in which the distance is based on the size of the largest deterministic finite automaton needed to separate two words. Finally, as mentioned in the previous section, there is a direct connection between graph limits in the monograph [15] by Nešetřil and Ossona de Mendez and the completion of the metric space $(\mathcal{P}(\mathrm{Inst}(\mathbf{T})), d)$, which may merit further exploration. It should also be pointed out that, motivated from the study of large-scale networks, there has been an extensive body of work on a notion of graph limits arising from converging sequences of *homomorphism densities*; a detailed account of this work is given in the monograph [13] by Lovász. In addition, Nešetřil and Ossona de Mendez [16] developed a general framework for limits of graphs and relational structures; in that framework, different fragments of first-order logic are used to define different notions of limits arising from converging sequences of the frequencies that first-order formulas in the fragment at hand are satisfied by an assignment (homomorphism densities correspond to the fragment consisting of all quantifier-free conjunctive queries). Homomorphisms, metric completions, and representations of limits of finite structures play a central role in [13, 16]. The precise connections with the work reported here will have to be worked out in a future investigation.

Acknowledgments. The research of Phokion G. Kolaitis on this paper was partially supported by NSF Grant IIS-1217869. The research of Reinhard Pichler, Emanuel Salinger, and Vadim Savenkov was supported by the Austrian Science Fund, projects (FWF):P25207-N23 and (FWF):Y698, and the Vienna Science and Technology Fund, project ICT12-015. Emanuel Sallinger is currently supported by the EPSRC grant EP/M025268/1.

References

1 Marcelo Arenas, Pablo Barceló, Leonid Libkin, and Filip Murlak. *Foundations of Data Exchange*. Cambridge University Press, 2014.

2 Marcelo Arenas, Jorge Pérez, Juan Reutter, and Cristian Riveros. The language of plain SO-tgds: Composition, inversion and structural properties. *Journal of Computer and System Sciences*, 79(6):763–784, 2013. doi:10.1016/j.jcss.2013.01.002.

3 Philip A. Bernstein. Applying model management to classical meta data problems. In *CIDR*, 2003.

4 Ronald Fagin. Probabilities on finite models. *J. Symb. Log.*, 41(1):50–58, 1976.

5 Ronald Fagin and Phokion G. Kolaitis. Local transformations and conjunctive-query equivalence. In *PODS*, pages 179–190, 2012. doi:10.1145/2213556.2213583.

6 Ronald Fagin, Phokion G. Kolaitis, Renée J. Miller, and Lucian Popa. Data exchange: Semantics and query answering. *Theor. Comput. Sci.*, 336(1):89–124, May 2005.

7 Ronald Fagin, Phokion G. Kolaitis, Alan Nash, and Lucian Popa. Towards a theory of schema-mapping optimization. In *PODS*, pages 33–42, 2008. doi:10.1145/1376916.1376922.

8 Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, and Wang Chiew Tan. Composing schema mappings: Second-order dependencies to the rescue. *ACM Trans. Database Syst.*, 30(4):994–1055, 2005.

9 Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, and Wang Chiew Tan. Schema mapping evolution through composition and inversion. In *Schema Matching and Mapping*, pages 191–222. Springer, 2011.

10 Ingo Feinerer, Reinhard Pichler, Emanuel Sallinger, and Vadim Savenkov. On the undecidability of the equivalence of second-order tuple generating dependencies. *Inf. Syst.*, 48:113–129, 2015. doi:10.1016/j.is.2014.09.003.

11 Phokion G. Kolaitis. Schema mappings, data exchange, and metadata management. In *PODS*, pages 61–75, 2005.

12 Maurizio Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246, 2002.

13 László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, 2012.

14 Jayant Madhavan and Alon Y. Halevy. Composing mappings among data sources. In *VLDB*, pages 572–583, 2003. URL: http://www.vldb.org/conf/2003/papers/S18P01.pdf.

15 Jaroslav Nešetřil and Patrice Ossona de Mendez. *Sparsity – Graphs, Structures, and Algorithms*, volume 28 of *Algorithms and combinatorics*. Springer, 2012. doi:10.1007/978-3-642-27875-4.

16 Jaroslav Nešetřil and Patrice Ossona de Mendez. A unified approach to structural limits, and limits of graphs with bounded tree-depth. arXiv:1303.6471, 2013.

17 Jean-Eric Pin. Profinite methods in automata theory. In *STACS*, pages 31–50, 2009.

18 Satish Shirali and Harkrishan Vasudeva. *Metric spaces*. Springer, 2006.

19 Balder ten Cate and Phokion G. Kolaitis. Structural characterizations of schema-mapping languages. In *ICDT*, pages 63–72, 2009. doi:10.1145/1514894.1514903.

20 Szymon Torunczyk. Languages of profinite words and the limitedness problem. In *ICALP*, pages 377–389, 2012.