Tuple-Independent Representations of Infinite Probabilistic Databases

Nofar Carmeli
snofca@cs.technion.ac.il
Technion—Israel Institute of Technology
Haifa, Israel

Martin Grohe
grohe@informatik.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Peter Lindner
lindner@informatik.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

Christoph Standke
standke@informatik.rwth-aachen.de
RWTH Aachen University
Aachen, Germany

ABSTRACT

Probabilistic databases (PDBs) are probability spaces over database instances. They provide a framework for handling uncertainty in databases, as occurs due to data integration, noisy data, data from unreliable sources or randomized processes. Most of the existing theory literature investigated finite, tuple-independent PDBs (TI-PDBs) where the occurrences of tuples are independent events. Only recently, Grohe and Lindner (PODS '19) introduced independence assumptions for PDBs beyond the finite domain assumption. In the finite, a major argument for discussing the theoretical properties of TI-PDBs is that they can be used to represent any finite PDB via views. This is no longer the case once the number of tuples is countably infinite. In this paper, we systematically study the representability of infinite PDBs in terms of TI-PDBs and the related block-independent disjoint PDBs.

The central question is which infinite PDBs are representable as first-order views over tuple-independent PDBs. We give a necessary condition for the representability of PDBs and provide a sufficient criterion for representability in terms of the probability distribution of a PDB. With various examples, we explore the limits of our criteria. We show that conditioning on first order properties yields no additional power in terms of expressivity. Finally, we discuss the relation between purely logical and arithmetic reasons for (non-)representability.

CCS CONCEPTS

• **Theory of computation → Incomplete, inconsistent, and uncertain databases**; **Logic and databases**.

KEYWORDS

Probabilistic Databases; Infinite Probabilistic Databases; Tuple-Independence; Block-Independent Disjoint Probabilistic Databases; First-Order Logic; Views

PODS '21, June 20–25, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8381-3/21/06…$15.00
https://doi.org/10.1145/3452021.3458315

ACM Reference Format:
Nofar Carmeli, Martin Grohe, Peter Lindner, and Christoph Standke. 2021. Tuple-Independent Representations of Infinite Probabilistic Databases. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS '21), June 20–25, 2021, Virtual Event, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3452021.3458315

1 INTRODUCTION

Probabilistic databases (PDBs) provide a framework for dealing with uncertainty in databases, as could occur due to data integration, the acquisition of noisy data or data from unreliable sources or as outputs of randomized processes. Often the appropriate probability spaces are infinite. For example, fields in a fact may contain measurements from a noisy sensor, which we model as real numbers with some error distribution, or approximate counters, modeled by some probability distribution over the integers, or text data scraped from unreliable web sources.

Formally, PDBs are probability spaces over (relational) database instances. A key issue when working with PDBs in practice is the question of how to represent them. For finite PDBs, this is always possible in principle (ignoring numerical issues) by just listing all instances in the PDB and their probabilities. But of course it is usually infeasible to store the whole sample space of a probability distribution over reasonably sized database instances. Instead, various compact representation systems have been proposed. Arguably the simplest is based on tuple-independent PDBs. In a *tuple-independent PDB (TI-PDB)* the truths of all facts f are regarded as independent events that hold with a probability $p_f \in [0, 1]$. The sample space of a TI-PDB with n facts has size 2^n, but we can represent the TI-PDB by just listing the n marginal probabilities p_f of the facts f. Much of the theoretical work on PDBs is concerned with TI-PDBs. A justification for this focus on TI-PDBs is the following nice Theorem [52]: *every finite PDB can be represented by a first-order view over a (finite) TI-PDB.* In other words: first-order views over TI-PDBs form a complete representation system for finite PDBs.

In this paper we investigate similar questions in a broadened scope. Our main focus lies on probabilistic databases with a countably infinite sample space. The idea of modeling uncertainty in databases by viewing them as infinite collections of possible worlds has been around for quite some time in the context of incomplete databases [31]. While existing PDB systems are already using infinite domains and sample spaces, such as [3, 11, 20, 32, 35, 48], a formal framework of probabilistic databases with infinite sample spaces has only been introduced recently [27, 29]. Let us emphasize that in an infinite PDB, it is the sample space of the probability distribution that is infinite; every single instance of an infinite PDB is still finite. As a special case, it is natural to consider infinite PDBs of *bounded instance size*, which means that all instances have size at most b for some fixed bound b. A simple example that can be modeled as a PDB of bounded size is a table that contains the number of car accidents for different countries where the numbers are regarded as inaccurate and the errors are modeled by some Poisson distribution. This model is often used in the insurance industry.

For infinite PDBs, the issue of representing them becomes even more difficult, because clearly, not every infinite PDB has a finite or computable representation. In this paper, we set out to study representations of infinite PDBs by considering representations over infinite TI-PDBs. Infinite TI-PDBs have been considered in [27], and it has been observed there that it is *not* the case that every countably infinite PDB can be represented by a first-order view over an infinite TI-PDB. The reason for this is relatively simple: it can be shown that the expected instance size in a TI-PDB is always finite, but there exist infinite PDBs with an infinite expected instance size. As first-order views preserve instance size up to a polynomial factor, such infinite PDBs with an infinite expected instance size cannot be represented by first-order views over TI-PDBs.

Let us denote the class of all PDBs that can be represented by a first-order view over a TI-PDB by $\mathsf{FO}(\mathsf{TI})$. We prove that the class $\mathsf{FO}(\mathsf{TI})$ is quite robust: somewhat unexpectedly, conditioning a TI-PDB on a first-order constraint prior to applying a first-order view does not yield additional expressive power, i. e. $\mathsf{FO}(\mathsf{TI} \mid \mathsf{FO}) = \mathsf{FO}(\mathsf{TI})$. This allows us to show that all block-independent disjoint PDBs (BID-PDBs) are in $\mathsf{FO}(\mathsf{TI})$. A BID-PDB is a PDB where the set of facts is partitioned into blocks, such that facts from different blocks are independent, while facts from the same block are disjoint. BID-PDBs form a practically quite relevant class of PDBs that includes PDBs of the form described in the car accident example above, where specific fields of facts in a table store the outcome of a random variable.

In [27], the authors exploited that TI-PDBs have finite expected instance size to construct a PDB that is not contained in $\mathsf{FO}(\mathsf{TI})$. We generalize this idea and prove that for every PDB in $\mathsf{FO}(\mathsf{TI})$, all *size moments*, that is, moments of the random variable that maps each instance to its size, are finite. This imposes a fairly strong restriction on the probability distribution of PDBs in $\mathsf{FO}(\mathsf{TI})$. In addition, we give an example showing that there are even PDBs that have finite size moments but that are still not in $\mathsf{FO}(\mathsf{TI})$. Complementing these non-representability results, we prove that all PDBs of bounded instance size are in $\mathsf{FO}(\mathsf{TI})$. Furthermore, we give a sufficient criterion on the growth rate of the probabilities of PDBs in $\mathsf{FO}(\mathsf{TI})$. This sufficient condition can be used to show that $\mathsf{FO}(\mathsf{TI})$ also contains PDBs of unbounded instance size.

All the non-representability results mentioned so far are caused by unwieldy probability distributions. We say that the reasons for these non-representability results are *arithmetical*. We asked ourselves if it can also happen that a PDB is not in $\mathsf{FO}(\mathsf{TI})$ for *logical* reasons, for example, because there are large gaps in the range of instance sizes. We formalize this question by saying that a PDB is not in $\mathsf{FO}(\mathsf{TI})$ *for logical reasons* if there is no probability distribution that assigns positive probabilities to all instances in the sample space such that the resulting PDB is in $\mathsf{FO}(\mathsf{TI})$. Thus being not representable for logical reasons is a property of the sample space and not of the probability measure. Arguably, this notion is closer to the theory of incomplete databases than to probabilistic databases. Surprisingly, we prove that there are no logical reasons for non-representability.

Related Work. By now, there exists an abundance of theoretical work on finite PDBs (see the surveys [52, 53] as well as [15, 51]). The most prominent theoretical problem regarding finite representation systems is probabilistic query evaluation (PQE) complexity [5, 16, 18, 22], typically subject to independence assumptions. In particular, [4] considers PQE on structurally restricted BID-PDBs. Among query languages, conjunctive queries and unions of conjunctive queries received the most attention [18]. On the side of expressiveness, [6, 9, 24, 47] consider more sophisticated PDB representations built upon independence assumptions. Probabilistic models for sensor networks typically feature a finite number of facts with continuously distributed attributes [20, 21, 49]. Overviews over representation formalisms for probabilistic databases can be found in [24, 46, 52, 53, 55].

Conditioning PDBs has been considered before in order to introduce correlations or dependencies [33, 44], for updates of probabilistic data [37], for making queries tractable [42], in cleaning problems [26, 45] and for introducing ontologies [10, 34]. In [17], the authors consider limit probabilities of conjunctive queries over conditioned PDBs.

There already exist PDB systems supporting infinite probability spaces [3, 11, 20, 32, 35, 48]. The investigations in these works are generally directly tied to their representation mechanisms, which complicates an abstract comparison. However, all of the mentioned approaches directly transition to a uncountable setting with continuous distributions, which is different from our setting of countably infinite PDBs. The formal possible worlds semantics have been extended towards infinite probability spaces, allowing for instances of unbounded size in [27, 29]. Allowing for distributions over worlds of unbounded size is strongly motivated by incorporating the open-world assumption into PDBs [10, 13, 23, 27]. Semi-structured models (probabilistic XML [36]) have been extended towards infinite spaces as well [1, 8].

As discussed by [24], incomplete databases (IDB) [2, 31, 54] are closely related to PDBs. In [19], the authors study expressiveness among classes of IDBs with strong associations with TI- and BID-PDBs. Recent work on IDBs considers (probabilistic) measures of certainty for infinite domain IDBs [14, 40].

It has been noticed that there are strong connections between PDBs and probabilistic models in AI research [53], in particular probabilistic graphical models [38] and weighted model counting (WMC) [25]. There exist well-established modeling formalisms that support infinite spaces, for example [30, 41, 50]. Recently, WMC has also been introduced for infinite universes [7].

Paper Outline. We review the background of PDBs in Section 2. In Section 3, we explore the limits of TI-representations. We extend the aforementioned non-representability result that relies on the expected instance size to apply to all size moments, and we provide a new necessary condition for representability that applies also in the case of finite moments. In Section 4, we show that **FO**(**TI**) is closed under conditioning under first-order constraints. Section 5 contains positive results. In Section 5.1 we prove a sufficient criterion on the growth rate of the probabilities and conclude also that PDBs of bounded instances size are in **FO**(**TI**). In Section 5.2, we show the same for BID-PDBs. Finally, in Section 6, we consider logical reasons. In Section 6.1, we demonstrate how the logic alone can be used to show non-representability of PDBs in the infinite. In Section 6.2, we show that, when the instance size is unbounded, any argument regarding representability using TI-PDBs with FO-views must take the instance probabilities into account. We close the paper with concluding remarks in Section 7.

2 PRELIMINARIES

In this section, we provide definitions and state known results that we use throughout this paper.

We denote the set of non-negative integers by $\mathbb{N}$, whereas the set of positive integers is denoted $\mathbb{N}_+$. We write $(0,1)$, $[0,1)$, $(0,1]$ and $[0,1]$ for the open, half-open and closed intervals of real numbers between 0 an 1.

Probability Spaces. A *discrete probability space* is a pair $\mathcal{S} = (\mathbb{S}, P)$ where $\mathbb{S}$ is a non-empty, countable set, called the *sample space* and P is a probability measure (or *probability distribution*) on $\mathbb{S}$. That is, $P\colon 2^{\mathbb{S}} \to [0,1]$ with the property that $P(A) = \sum_{s \in A} P(\{s\})$ for each $A \subseteq \mathbb{S}$, and $P(\mathbb{S}) = 1$. We denote probability spaces with curly letters and their sample spaces with double-struck letters. We denote probability distributions with variants of P. Throughout this paper, all probability spaces are assumed to be discrete.

We write $S \sim \mathcal{S}$ to indicate that S is a random element, drawn from $\mathbb{S}$ according to distribution P. If P is anonymous, or when we want to emphasize this perspective of drawing a random element, we write

$$\Pr_{S\sim\mathcal{S}}\big(S \text{ has property } \varphi\big) := P\big(\{S \in \mathbb{S} \colon S \text{ has property } \varphi\}\big).$$

Subsets of $\mathbb{S}$ are called *events*. A collection $(A_i)_{i \in I}$ of events in $\mathcal{S}$ is called *(mutually) independent* if $P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$ for all finite subsets J of I. It is called *(mutually) exclusive* if $P(A_i \cap A_j) = 0$ for all $i \neq j$.

A *random variable* X on $\mathcal{S} = (\mathbb{S}, P)$ is a function $X\colon \mathbb{S} \to \mathbb{R}$. It's *expectation* is $\mathrm{E}(X) := \sum_{s \in \mathbb{S}} X(s) \cdot P(\{s\})$ and, in general, its *kth moment* is $\mathrm{E}(X^k)$. We write $\mathrm{E}_{\mathcal{S}}$ for the expectation in $\mathcal{S}$ if the probability space is not clear from the context.

If $\mathcal{S} = (\mathbb{S}, P_{\mathcal{S}})$ is a probability space, $\mathbb{T}$ a (countable) set and $f\colon \mathbb{S} \to \mathbb{T}$ a function, then f introduces a probability distribution on $\mathbb{T}$ via $P_{\mathcal{T}}(\{t\}) := P_{\mathcal{S}}(\{s \in \mathbb{S} \colon f(s) = t\})$.

Relational Databases. We fix some non-empty set $\mathbb{U}$ (called the *universe*). A *database schema* τ is a finite, nonempty set of relation symbols with *arities* $\mathrm{ar}(R) \in \mathbb{N}$ for all $R \in \tau$. A (τ-)fact is an expression of the shape $R(u_1, \dots, u_{\mathrm{ar}(R)})$ where $R \in \tau$ and $u_i \in \mathbb{U}$ for all $1 \leq i \leq \mathrm{ar}(R)$. A *$\tau$-instance* D is a finite set of τ-facts. The *active domain* of D, denoted by $\mathrm{adom}(D)$, is the set of elements of $\mathbb{U}$ appearing among the facts of D. Throughout this paper, we assume that the universe $\mathbb{U}$ is countably infinite.

First-Order Logic. Recall that formulas of *first-order logic* **FO** of schema (or vocabulary) τ are formed from *atomic formulas* of the form $R(u_1, \dots, u_k)$, where $R \in \tau$ is a k-ary relation symbol and the u_i are either variables or elements of $\mathbb{U}$, using the standard Boolean connectives and existential and universal quantification ranging over elements of the universe. An **FO**-formula $\Phi(\bar{x})$ and an instance D (of the same schema) define a relation R where $\bar{t} \in R$ if and only if D satisfies $\Phi(\bar{t})$. *Conjunctive queries*, denoted **CQ**, are a subclass of **FO** formed from the atomic formulas by only taking conjunctions and existential quantification. Adding disjunctions, we obtain the class **UCQ** of *unions of conjunctive queries*. For background on first order logic, see [39].

Probabilistic Databases. A probabilistic database is a probability space over a set of database instances.

Definition 2.1. A *probabilistic database (PDB)* of database schema τ over $\mathbb{U}$ is a discrete probability space $\mathcal{D} = (\mathbb{D}, P)$ where $\mathbb{D}$ is a set of τ-instances. We denote the class of PDBs by **PDB**. If **D** is a class of PDBs and $\mathcal{D} \in \mathbf{D}$, we call $\mathcal{D}$ a **D**-*PDB*.

In a PDB, the instances of positive probability are often called its *possible worlds*. The set $T(\mathcal{D})$ of facts appearing among the possible worlds of a PDB $\mathcal{D}$ is countable. For every fact $t \in T(\mathcal{D})$, the *marginal probability* of t is $\Pr_{D\sim\mathcal{D}}(t \in D)$. Note that while $|\mathbb{D}|$ may be infinite, the instances within $\mathbb{D}$ are always *finite* collections of facts. We call a PDB $(\mathbb{D}, P)$ finite if $|\mathbb{D}|$ is finite.

Remark 2.2. According to Definition 2.1, all PDBs we consider in this paper have a sample space of at most countable size. In general, a PDB may also be an uncountable probability space over some uncountable domain [27, 29]. Usually, *tuple-independent* PDBs are defined by the property that the events "$t \in D$" are stochastically independent (see Definition 2.3 below). In this shape, it only makes sense for countable PDBs, as in every PDB there are at most countably many facts with positive marginal probability (see [27, Proposition 3.4]). There is an extension of tuple-independent PDBs to uncountable domains, so-called Poisson PDBs [28]. It remains future work to extend the results of this paper to the more general setting of Poisson PDBs.

Instance Size. In every PDB $\mathcal{D} = (\mathbb{D}, P)$, the *instance size* $|\cdot|$ is a random variable $|\cdot|\colon \mathbb{D} \to \mathbb{N}$ that maps every database instance D to the number $|D|$ of facts it contains. Its expectation is given by $\mathrm{E}(|\cdot|) = \sum_{D \in \mathbb{D}} |D| \cdot P\big(\{D\}\big)$.

We say that $\mathcal{D} = (\mathbb{D}, P)$ has the *finite moments property* if all moments of its instance size random variable are finite. That is, for all $k \in \mathbb{N}_+$, we have $\mathrm{E}\left(|\cdot|^k\right) = \sum_{D \in \mathbb{D}} |D|^k \cdot P\big(\{D\}\big) < \infty$. A class **D** of PDBs has the *finite moments property* if every $\mathcal{D} \in \mathbf{D}$ does.

Query Semantics. In general, a *view* may be any function that maps database instances of an input schema τ to instances of an output schema τ'. A *query* is a view whose output schema consists of a single relation symbol. Then, a view may also be thought of as a finite collection of queries, one per each relation in the output schema. Applying a view on a PDB yields a new output PDB: If $\mathcal{D} = (\mathbb{D}, P)$ is a PDB, $\mathbb{D}'$ a set of database instances and $V\colon \mathbb{D} \to \mathbb{D}'$ a view, then $V(\mathcal{D}) = (\mathbb{D}', P')$ is a PDB where

$$P'(\{D'\}) = P(\{D \in \mathbb{D} \colon V(D) = D'\})$$

for all $D' \in \mathbb{D}'$.

If **V** is a class of views and **D** is a class of PDBs, then **V**(**D**) denotes the class of images of PDBs of **D** under views of **V**. We call **D** *closed under* **V** if **V**(**D**) = **D**. An **FO**-view is a view that consists of an **FO**-formula for each relation symbol in the target schema. Then, **FO**(**D**) denotes the class of PDBs that are the image of a **D**-PDB under an **FO**-view. These notions are defined analogously for **CQ** and **UCQ**.

Independence Assumptions. Even when leaving out probabilities, the number of possible worlds existing over a fixed set of n facts is exponential in n. In the finite setting, this motivates the introduction of simplifying structural assumptions that allow for succinct representations of PDBs. While for infinite PDBs such a representation might still be infinite, its description still enjoys the simple structure and thus may have advantages with respect to approximate query answering.

Definition 2.3 (Tuple-Independent PDBs)**.** A PDB $\mathcal{D}$ is called *tuple-independent* if for all $k \in \mathbb{N}$ and all pairwise distinct facts $t_1, \dots, t_k \in T(\mathcal{D})$ it holds that

$$\Pr_{D \sim \mathcal{D}}\big(t_1 \in D, \dots, t_k \in D\big) = \prod_{i=1}^{k} \Pr_{D \sim \mathcal{D}}\big(t_i \in D\big).$$

We denote the class of tuple-independent PDBs by **TI**.

For ease of reading, we proceed to denote tuple-independent by $\mathcal{I} = (\mathbb{I}, P)$ instead of $\mathcal{D} = (\mathbb{D}, P)$. Instances of tuple-independent PDBs are denoted by I (or variants) instead of D accordingly. The following provides a necessary and sufficient criterion for the existence of **TI**-PDBs in terms of its marginal probabilities.

Theorem 2.4 ([27, Theorem 4.8])**.** *Let T be a set of facts over a schema τ and let $(p_t)_{t \in T}$ with $p_t \in [0,1]$ for all $t \in T$. The following are equivalent:*

1. *There exists $\mathcal{I} \in$ **TI** with fact set $T(\mathcal{I}) = T$ and marginal probabilities $\Pr_{I \sim \mathcal{I}}(t \in I) = p_t$ for all $t \in T$.*
2. *It holds that $\sum_{t \in T} p_t < \infty$.*

TI-PDBs are a special case of the following, more general model.

Definition 2.5 (Block-Independent Disjoint PDBs)**.** A PDB $\mathcal{D}$ is called *block-independent disjoint* if there exists a partition $\mathcal{B}$ of $T(\mathcal{D})$ into *blocks* such that:

1. for all $k \in \mathbb{N}$ and all $t_1, \dots, t_k$ from pairwise different blocks,
$$\Pr_{D \sim \mathcal{D}}\big(t_1 \in D, \dots, t_k \in D\big) = \prod_{i=1}^{k} \Pr_{D \sim \mathcal{D}}\big(t_i \in D\big).$$
2. for all $B \in \mathcal{B}$ and all $t, t' \in B$ with $t \neq t'$ it holds that $\Pr_{D \sim \mathcal{D}}\big(t \in D \text{ and } t' \in D\big) = 0$.

We denote the class of block-independent disjoint PDBs by **BID**.

A theorem similar to Theorem 2.4 exists for **BID**-PDBs:

Theorem 2.6 ([27, Theorem 4.15])**.** *Let T be a set of facts over a schema τ, let $(p_t)_{t \in T}$ with $p_t \in [0,1]$, and let $\mathcal{B}$ be a partition of T such that $\sum_{t \in B} p_t \leq 1$ for all $B \in \mathcal{B}$. The following are equivalent:*

1. *There exists $\mathcal{D} \in$ **BID** with fact set $T(\mathcal{D}) = T$, blocks $\mathcal{B}$ and marginal probabilities $\Pr_{D \sim \mathcal{D}}(t \in D) = p_t$ for all $t \in T$.*
2. *It holds that $\sum_{t \in T} p_t = \sum_{B \in \mathcal{B}} \sum_{t \in B} p_t < \infty$.*

We note that both **TI**-, and **BID**-PDBs are, in case of existence, uniquely determined by their marginals (and partition into blocks) [28, Section 4].

Finite Representation Systems. Although our focus lies on countably infinite PDBs, we briefly recall the situation in the finite setting. For finite PDBs, the differences in expressiveness are more easily accessible. Let $\mathbf{PDB}_{\text{fin}}$, $\mathbf{TI}_{\text{fin}}$ and $\mathbf{BID}_{\text{fin}}$ denote the classes of finite PDBs, finite **TI**-PDBs and finite **BID**-PDBs, respectively. Then the relationships between these classes are as shown in the Hasse diagram in Figure 1, with respect to strict inclusion. The relationships of Figure 1 are well-known or easy to show. Some proofs and examples can be found in [16, 52]. Proofs of these results (wherever we could not track them down in literature) are contained in the extended version of this paper [12].

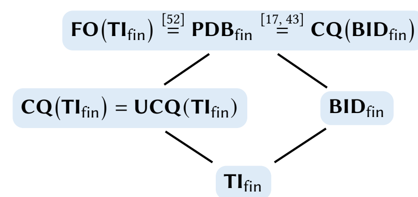

Figure 1: Fin. PDB classes with independence assumptions.

3 LIMITATIONS OF TI REPRESENTATIONS

Unlike the case for finite PDBs [52], we know that there exist infinite PDBs that are not representable by an **FO**-view of a **TI**-PDB.

Proposition 3.1 ([27, Proposition 4.9])**.** **FO**(**TI**) $\subsetneq$ **PDB**.

We devote this section to examine cases of PDBs for which we can prove that such a representation does not exist. In Section 3.1 we generalize the argument of [27] and show that for a PDB to be in **FO**(**TI**) it is a necessary condition that all moments of its instance size are finite. In Section 3.2 we show that this necessary condition is not sufficient: some PDBs that have the finite moments property are not in **FO**(**TI**).

3.1 Infinite Moments

For showing that infinite moments of the instance size random variable imply non-representability in **FO**(**TI**), we first show that every **TI**-PDB has the finite moments property.

Proposition 3.2. **TI** *has the finite moments property.*

Proof. Suppose $T(\mathcal{I}) = \{t_0, t_1, \dots\}$ and let $X_i \colon \mathbb{I} \to \{0,1\}$ be the indicator variable of the event $\{t_i \in I\}$. Since the size of any instance $I \in \mathbb{I}$ is the number of facts in I, we can express the instance size random variable using the indicator variables as $|\cdot| = \sum_{i \in \mathbb{N}} X_i$.

For simplicity, we write E for $\mathrm{E}_{\mathcal{I}}$. Since X_i is an indicator variable, we know that

$$\mathrm{E}(X_i) = \Pr_{I \sim \mathcal{I}}(X_i(I) = 1) = \Pr_{I \sim \mathcal{I}}(t_i \in I) = p_{t_i}$$

which yields

$$\mathrm{E}(|\cdot|) = \mathrm{E}(\sum_{i \in \mathbb{N}} X_i) = \sum_{i \in \mathbb{N}} \mathrm{E}(X_i) = \sum_{i \in \mathbb{N}} p_{t_i}.$$

By Theorem 2.4 the last sum is finite and this proves the claim for $k = 1$. For $k > 1$, we prove the proposition via induction on k by using $\mathrm{E}(|\cdot|^k) \le \mathrm{E}(|\cdot|^{k-1}) \cdot (k - 1 + \mathrm{E}(|\cdot|))$ (the extended version [12] contains a proof of this inequality). The right-hand side is finite by the induction hypothesis and this proves the claim. ◻

Next, we observe that **FO**-views preserve finite moments.

Lemma 3.3. *If $\mathcal{D}'$ is a PDB with the finite moments property and V is an* **FO***-view, then $V(\mathcal{D}')$ also has the finite moments property.*

Proof. Let $\mathcal{D} = V(\mathcal{D}')$. Then $\mathrm{E}_{\mathcal{D}}(|\cdot|^k) = \mathrm{E}_{\mathcal{D}'}(|V(\cdot)|^k)$. Suppose that the schema of $\mathcal{D}$ consists of m relation symbols with arities $r_1, \dots, r_m$. Then V consists of m first order formulas $\varphi_1, \dots, \varphi_m$. Now, letting c_i denote the number of constants in φ_i, it holds that

$$|V(D')| = \sum_{i=1}^{m} |\varphi_i(D')| \le \sum_{i=1}^{m} \big(|\mathrm{adom}(D')| + c_i\big)^{r_i}.$$

Note that $|\mathrm{adom}(D')| \le r' \cdot |D'|$ for the maximum arity r' in the schema of $\mathcal{D}'$. Denoting $r = \max_i r_i$ and $c = \max_i c_i$, we conclude

$$|V(D')| \le m \cdot \big(|\mathrm{adom}(D')| + c\big)^r \le m \cdot \big(r'|D'| + c\big)^r$$

and thus, by the binomial formula,

$$\begin{aligned} \mathrm{E}_{\mathcal{D}}\left(|\cdot|^k\right) &\le \mathrm{E}_{\mathcal{D}'}\left(m^k (r'|\cdot| + c)^{rk}\right) \\ &= m^k \sum_{j=0}^{rk} \tbinom{rk}{j} r'^j c^{rk-j}\, \mathrm{E}_{\mathcal{D}'}\left(|\cdot|^j\right). \end{aligned}$$

Since all the $\mathrm{E}_{\mathcal{D}'}(|\cdot|^j)$ are finite, this is a finite expression. ◻

We note, however, that the property of having a finite kth moment (but not necessarily a finite $(k+1)$st moment) for fixed k, is in general not preserved under **FO**-views.

Using Lemma 3.3, we can extend Proposition 3.2 to **FO**(**TI**).

Proposition 3.4. **FO**(**TI**) *has the finite moments property.*

This yields numerous examples of PDBs that are not in **FO**(**TI**).

Example 3.5. Let $\mathcal{D} = (\mathbb{D}, P)$ be a PDB over a schema consisting of only a single unary relation symbol with $\mathbb{D} = \{D_1, D_2, \dots\}$, $|D_i| = 2^i$ and $P(\{D_i\}) = 3 \cdot 4^{-i}$. Then

$$\mathrm{E}(|\cdot|) = \sum_{D \in \mathbb{D}} |D| \cdot P(\{D\}) = 3 \sum_{i=1}^{\infty} 2^{-i} = 3$$

but

$$\mathrm{E}(|\cdot|^2) = \sum_{D \in \mathbb{D}} |D|^2 \cdot P(\{D\}) = 3 \sum_{i=1}^{\infty} 1 = \infty.$$

According to Proposition 3.4, $\mathcal{D} \notin$ **FO**(**TI**).

3.2 Balancing the Marginal Probabilities

In this section, we prove that there are PDBs having the finite moments property that do not have a representation as an **FO**-view over a **TI**-PDB.

A valid representation needs to balance the marginal probabilities of the independent facts. On the one hand, the sum of these probabilities must converge, so, intuitively, the probabilities have to decrease fast enough. On the other hand, the probabilities must be large enough to represent the possible worlds. In particular, including all elements of the active domain of some possible world has to be at least as probable as this world. In the following, we formalize the implications of this second requirement on the sum of marginal probabilities.

Lemma 3.6. *Consider an* **FO***-view Φ over a* **TI***-PDB $\mathcal{I} = (\mathbb{I}, P')$ and let r be the maximum arity of a relation in $\mathcal{I}$. Let D_n be an instance of $\Phi(\mathcal{I})$, and let V_n be the set of elements in the active domain of D_n that do not appear as constants in Φ. Let E_n be the facts of $\mathbb{I}$ containing at least one element of V_n, and let q_e be the marginal probability of $e \in E_n$ in $\mathcal{I}$. Then,*

$$\Pr_{I \sim \mathcal{I}}(\Phi(I) = D_n) \le |V_n| \left(r^2 |V_n|^{r-1} \sum_{e \in E_n} q_e \right)^{\frac{|V_n|}{r}}.$$

Proof Sketch. We define a (multi-)hypergraph $H = (V, E)$ where $V = \bigcup_{I \in \mathbb{I}} \mathrm{adom}(I)$ is the active domain of $\mathbb{I}$, and there is an (hyper-)edge for every fact of $\mathbb{I}$ containing the elements that appear in the fact. Denote by $\mathrm{EC}_H(S)$ the set of edge covers of the nodes S in the hypergraph H, and denote by $\mathrm{EC}^*_H(S)$ the set of all such minimal edge covers. A necessary condition for $\Phi(I) = D_n$ is that every element of the active domain of D_n appears either as a constant in Φ or in a fact of I. Therefore,

$$\begin{aligned} \Pr_{I \sim \mathcal{I}}(\Phi(I) = D_n) &\le \Pr_{I \sim \mathcal{I}}(\text{every element of } V_n \text{ appears in } I) \\ &= \Pr_{I \sim \mathcal{I}}(I \text{ is an edge cover of } V_n) \\ &= \sum_{S \in \mathrm{EC}_H(V_n)} \Pr_{I \sim \mathcal{I}}(I = S). \end{aligned}$$

To be able to bound the last expression, we translate it to minimal edge covers over a deduplicated hypergraph. We define H'_n to be H restricted to the nodes of V_n and without duplicate edges. Using the inequality of the geometric and arithmetic means, we show that:

$$\begin{aligned} \sum_{S \in \mathrm{EC}_H(V_n)} \Pr_{I \sim \mathcal{I}}(I = S) &\le \sum_{C \in \mathrm{EC}^*_H(V_n)} \prod_{e \in C} q_e \\ &\le \sum_{C' \in \mathrm{EC}^*_{H'_n}(V_n)} \left(\frac{1}{|C'|} \sum_{e \in E_n} q_e \right)^{|C'|}. \end{aligned}$$

As r is the maximum arity of a relation in $\mathcal{I}$, each edge in H'_n contains at most r elements. Therefore, the size of a minimal edge cover of V_n is between $\frac{|V_n|}{r}$ and $|V_n|$. For the same reason, the number of edges in H'_n is at most

$$\sum_{i=1}^{r} \binom{|V_n|}{i} \le \sum_{i=1}^{r} |V_n|^i \le r|V_n|^r.$$

Hence, there are at most $\binom{r|V_n|^r}{k} \le (r|V_n|^r)^k$ minimal edge covers of H'_n of size k. By incorporating these bounds into the previous expression, we obtain the inequality stated in the lemma. The full proof can be found in the extended version of this paper [12]. ◻

The condition of Lemma 3.6 combined with the fact that the sum of all marginal probabilities in a TI-PDB must converge yields a requirement on every possible **FO**-representation over TI-PDBs. We wish to utilize this requirement to formalize a property of the represented PDB itself. This property is concluded from the mere fact that such a representation exists, regardless of the choice of representation. Lemma 3.6 considers individual possible worlds. In order to translate the convergence restriction of the sum of all facts to a restriction involving a single possible world at a time, we consider domain disjoint PDBs: A PDB $\mathcal{D}$ is called *domain disjoint* if $\mathrm{adom}(D) \cap \mathrm{adom}(D') = \emptyset$ for all $D, D' \in \mathbb{D}$ with $D \neq D'$. Investigating domain-disjoint PDBs allows us to identify a connection between the probabilities of the possible worlds and the size of their active domains. Given an instance D_n, we denote $d_n := |\mathrm{adom}(D_n)|$.

Lemma 3.7. *Given a domain disjoint $\mathcal{D} \in \mathsf{FO}(\mathsf{TI})$ with instances $\mathbb{D} = \{D_0, D_1 \dots\}$, there exists a constant $r \in \mathbb{N}_+$ s. t. for every divergent series $\sum_{n\in\mathbb{N}} a_n = \infty$, there are infinitely many $n \in \mathbb{N}$ with*

$$\Pr_{D\sim\mathcal{D}}(D = D_n) < d_n\big(a_n d_n^{r-1}\big)^{\frac{d_n}{r}}.$$

Proof. Consider a representation of $\mathcal{D}$ as an FO-view Φ over the TI-PDB $\mathcal{I} = (\mathbb{I}, P')$, and let r be the maximum arity of a relation of $\mathcal{D}$. As before, let V_n be the active domain of D_n that does not appear in Φ, let E_n be the facts of $\mathbb{I}$ that contain at least one element of V_n, and let q_e be the marginal probability of $e \in E_n$.

In the following, we consider a series that consists of the sum of marginal probabilities of facts that belong to each possible world. Intuitively, since the sum of probabilities of all facts converges, and since there is a bound on the number of possible worlds to which a single fact can contribute, the sum of this series converges as well. More formally, we obtain the following inequalities since for each $v \in V = \bigcup_{I\in\mathbb{I}} \mathrm{adom}(I)$, there is at most one $n \in \mathbb{N}$ such that $v \in V_n$:

$$\sum_{n\in\mathbb{N}}\sum_{e\in E_n} q_e \le \sum_{n\in\mathbb{N}}\sum_{v\in V_n}\sum_{\substack{e\in E\\ v\in e}} q_e \le \sum_{v\in V}\sum_{\substack{e\in E\\ v\in e}} q_e = \sum_{e\in E}\sum_{v\in e} q_e \le \sum_{e\in E} r q_e = r\sum_{e\in E} q_e.$$

Since $\mathcal{I}$ is a well-defined TI-PDB, $\sum_{e\in E} q_e < \infty$. This means that in a valid representation of possible worlds with disjoint domains, the sum over all $\sum_{e\in E_n} q_e$ converges. Thus, every subseries of it converges as well. In particular, since $\sum_{n\in\mathbb{N}} \frac{a_n}{r^2}$ diverges, for infinitely many n values, we get

$$\sum_{e\in E_n} q_e < \tfrac{a_n}{r^2}. \tag{1}$$

Otherwise, $\sum_{n\in\mathbb{N}}\sum_{e\in E_n} q_e$ would contain a divergent subseries.

As there is only a constant number of elements in the view, and the active domains of distinct possible worlds are disjoint, there is a finite number of possible worlds with elements in Φ. Therefore, for n large enough, $V_n = \mathrm{adom}(D_n)$ and $|V_n| = d_n$. We established that there are infinitely many values $n \in \mathbb{N}$ such that inequality (1) holds and $d_n = |V_n|$. Using Lemma 3.6 for these values of n:

$$\Pr_{D\sim\mathcal{D}}(D = D_n) = \Pr_{I\sim\mathcal{I}}(\Phi(I) = D_n) \le |V_n|\Big(r^2|V_n|^{r-1}\sum_{e\in E_n} q_e\Big)^{\frac{|V_n|}{r}} < d_n\big(a_n d_n^{r-1}\big)^{\frac{d_n}{r}}. \qquad \square$$

Remark 3.8. Lemma 3.7 holds also if the domains of the possible worlds are not disjoint, but instead there is a bound on the number of possible worlds in which every domain element appears.

Lemma 3.7 can be used to identify PDBs that do not have a representation as an FO-view over a TI-PDB.

Example 3.9. Let $\mathcal{D} = (\mathbb{D}, P)$ be a PDB over a schema consisting of only a single unary relation symbol with $\mathbb{D} = \{D_1, D_2, \dots\}$, $|\mathrm{adom}(D_n)| = \lceil \log n \rceil$ and $P(\{D_n\}) = \frac{c}{n^2}$. Here, $c = \frac{6}{\pi^2}$ is the constant that scales the probabilities so that they add up to 1.

Assume by contradiction that $\mathcal{D} \in \mathsf{FO}(\mathsf{TI})$. Since $\sum_{n\in\mathbb{N}_+} \frac{1}{n}$ is a divergent series, according to Lemma 3.7, there exists a constant $r \in \mathbb{N}_+$ such that there are infinitely many $n \in \mathbb{N}_+$ with

$$\frac{c}{n^2} = \Pr_{D\sim\mathcal{D}}(D = D_n) < d_n\big(\tfrac{1}{n} d_n^{r-1}\big)^{\frac{d_n}{r}} = \lceil \log(n) \rceil \Big(\frac{\lceil \log(n) \rceil^{r-1}}{n}\Big)^{\frac{\lceil \log(n) \rceil}{r}}$$

For large enough n, this is a contradiction. When $\lceil \log(n) \rceil \le n^{\frac{1}{r}}$, $\lceil \log(n) \rceil \ge 3r^2 + r$, and $n > \frac{1}{c}$, we have:

$$\begin{aligned}\lceil \log(n) \rceil \left(\frac{\lceil \log(n) \rceil^{r-1}}{n}\right)^{\frac{\lceil \log(n) \rceil}{r}} &\le n^{\frac{1}{r}}\left(n^{\frac{r-1}{r}} n^{-1}\right)^{\frac{\lceil \log(n) \rceil}{r}}\\ &= n^{\frac{1}{r}}\left(n^{-\frac{1}{r}}\right)^{\frac{\lceil \log(n) \rceil}{r}} = n^{\frac{r-\lceil \log(n) \rceil}{r^2}} \le n^{-3} < \frac{c}{n^2}.\end{aligned}$$

Note that the PDB $\mathcal{D}$ has the finite moments property. To see that, let $k \in \mathbb{N}_+$. Then there is n_k such that for all $n \ge n_k$ we have $\log(n) \le n^{\frac{1}{2k}}$. Hence,

$$\begin{aligned}\mathrm{E}_{\mathcal{D}}(|\cdot|^k) = \sum_{n=1}^{\infty} \frac{c\log(n)^k}{n^2} &\le \sum_{n=1}^{n_k} \frac{c\log(n)^k}{n^2} + \sum_{n=n_k+1}^{\infty} \frac{c(n^{\frac{1}{2k}})^k}{n^2}\\ &\le \sum_{n=1}^{n_k} \frac{c\log(n)^k}{n^2} + c\sum_{n=n_k+1}^{\infty} n^{-\frac{3}{2}}.\end{aligned}$$

This expression is finite since the first sum is finite and the second sum is well-known to converge.

Example 3.9 proves Theorem 3.10.

Theorem 3.10. *There are PDBs having the finite moments property that are not in* $\mathsf{FO}(\mathsf{TI})$.

4 CONDITIONAL VIEWS

In this section we define conditional representations: representations of PDBs as views over TI-PDBs conditioned on an FO-sentence. The possible worlds are restricted to those satisfying the condition, and the probability mass of the valid instances is scaled to add up to one. As the next sections demonstrate, when constructing a representation to a PDB, conditional representations are often simpler to identify and explain compared to unconditional ones. We show the equivalence between the PDBs that admit an FO-representation and those that admit a conditional FO-representation. As a consequence, we obtain a tool for showing that a PDB has a representation as an FO-view over a TI-PDB: it is enough to identify such a representation that is conditioned on an FO-sentence.

We start by defining conditional views. Given a PDB $\mathcal{D} = (\mathbb{D}, P)$ and an FO-sentence φ such that $\Pr_{D\sim\mathcal{D}}(D \models \varphi) > 0$, we denote by $\mathcal{D} \mid \varphi$ the PDB $(\mathbb{D}_\varphi, P')$ where $\mathbb{D}_\varphi := \{D \in \mathbb{D} \mid D \models \varphi\}$ and, for all $D \in \mathbb{D}_\varphi$,

$$P'(D) := P\big(\{D\} \mid \mathbb{D}\varphi\big) = \frac{P(\{D\})}{P(\mathbb{D}_\varphi)}.$$

Given a class **D** of PDBs, we denote by **D** | **FO** the class of all PDBs obtained by conditioning a PDB of **D** on an **FO**-sentence. That is,

$$\mathbf{D} \mid \mathbf{FO} := \big\{(\mathcal{D} \mid \varphi) \colon \mathcal{D} \in \mathbf{D} \text{ and } \varphi \in \mathbf{FO} \text{ sentence} \\ \text{with } \Pr_{D\sim\mathcal{D}}(D \models \varphi) > 0\big\}.$$

The following is the main result of this section, stating that the class of PDBs that can be represented as **FO**-views over **FO**-*conditioned* **TI**-PDBs coincides with the class of PDBs that can be represented by an **FO**-view of a **TI**-PDB alone.

Theorem 4.1. $\mathbf{FO}(\mathbf{TI} \mid \mathbf{FO}) = \mathbf{FO}(\mathbf{TI})$.

Remark 4.2. Note that $\mathbf{FO}(\mathbf{TI} \mid \mathbf{FO}) \subseteq \mathbf{FO}(\mathbf{TI})$ is by no means trivial. In particular, we cannot simply merge the condition given as an **FO**-sentence into the **FO**-view. Composing them cannot work, as the condition is a sentence, and so the composition can only result in two outcomes and cannot represent PDBs with more than two possible worlds. Intersecting them will not work either. The conditional view removes the the possible worlds that do not meet the condition from the sample space and scales the probability mass of the remaining possible worlds up to one, while intersecting the **FO**-view with the condition keeps the probability mass of the invalid possible worlds but renders these worlds empty.

Proof (Theorem 4.1). The direction $\mathbf{FO}(\mathbf{TI}) \subseteq \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$ is immediate. We show $\mathbf{FO}(\mathbf{TI}|\mathbf{FO}) \subseteq \mathbf{FO}(\mathbf{TI})$. In the following proof, given a PDB $\mathcal{D} = (\mathbb{D}, P)$ and an **FO**-sentence φ, when plugged into P, we treat φ as the event $\{D \in \mathbb{D} \colon D \models \varphi\}$. That is, $P(\varphi) := P(\{D \in \mathbb{D} \colon D \models \varphi\})$ and similarly, for all events $\mathbb{D}_0 \subseteq \mathbb{D}$, we let $P(\mathbb{D}_0 \cap \varphi) := P(\{D \in \mathbb{D}_0 \colon D \models \varphi\})$.

Let $\mathcal{D} = (\mathbb{D}, P_{\mathcal{D}}) \in \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$. Then, there exist a **TI**-PDB $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$, an **FO**-view Φ and an **FO**-sentence φ such that $\mathcal{D} = \Phi(\mathcal{I} \mid \varphi)$. That is, $P_{\mathcal{D}}(\{D\}) = P_{\mathcal{I}}(\Phi^{-1}(D) \mid \varphi)$ for all $D \in \mathbb{D}$. Note that, as $\mathcal{I} \mid \varphi$ is properly defined, this implies that $P_{\mathcal{I}}(\varphi) > 0$.

Fix an instance D_0 of $\mathcal{D}$ with $p_0 := P_{\mathcal{D}}(\{D_0\}) > 0$. If $p_0 = 1$, then $\mathcal{D}$ consists of a single instance of positive probability, and thus $\mathcal{D} \in \mathbf{TI} \subseteq \mathbf{FO}(\mathbf{TI})$ by assigning all facts of D_0 probability 1 and all other facts probability 0. Hence, we assume $p_0 < 1$.

The outline of the rest of the proof is as follows (see Figure 2). The instance D_0 will receive a separate treatment: First, we construct an **FO**-sentence φ_0 that characterizes whether an instance of $\mathcal{I}$ maps to D_0. Next, we create a new **TI**-PDB $\mathcal{I}^{(k)}$ from k independent copies of $\mathcal{I}$. The facts of $\mathcal{I}^{(k)}$ are the original facts from $\mathcal{I}$, extended by an identifier ranging over $1, \dots, k$. We are interested in those instances that, restricted to one of the identifiers, satisfy the condition φ while not already mapping to D_0, that is, satisfying $\varphi \wedge \neg\varphi_0 =: \psi$. These are used as the potential representations of the instances different from D_0. Among the identifiers that could be picked this way, we chose the minimal one to get a unique representation. If we think of drawing from $\mathcal{I}^{(k)}$ as drawing k times independently from $\mathcal{I}$, then the probability of not having encountered a single representation can be arbitrarily low, depending on the parameter k. If k is chosen high enough, the probability mass of always violating the condition is low enough to be concealed within the probability mass of the new representation of D_0.

We start with the introduction of φ_0, describing $\Phi^{-1}(D_0)$.

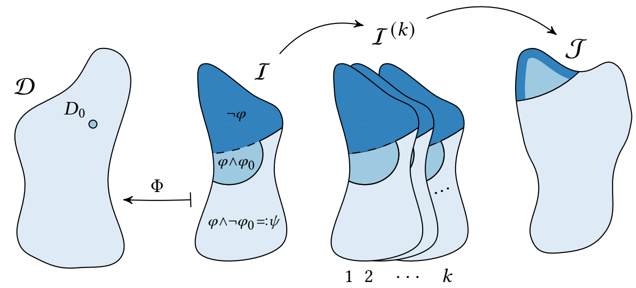

Figure 2: Illustration of the constructions in the proof.

Claim 4.3. *There exists an* **FO***-sentence* φ_0 *with the property that* $I \models \varphi_0$ *if and only if* $\Phi(I) = D_0$.

Proof. Suppose that $\mathcal{D}$ has m relations $R_1, \dots, R_m$. Then Φ consists of m **FO**-formulae $\Phi_1, \dots, \Phi_m$. Further suppose that we have $R_i(D_0) = \{R_i(\bar{a}_{i1}), \dots, R_i(\bar{a}_{in_i})\}$. We define

$$\varphi_0 := \bigwedge\nolimits_{i=1}^{m}\big(\forall \bar{x} = (x_i, \dots, x_{r_i}) \colon \Phi_i(\bar{x}) \leftrightarrow \bigvee\nolimits_{j=i}^{n_i} \bar{x} = \bar{a}_{ij}\big)$$

where r_i is the arity of R_i in $\mathcal{D}$. Then $I \models \varphi_0$ if and only if the satisfying valuations of Φ_i are exactly valuations obtained from $\Phi_i^{-1}(D_0)$. That is, $\{I \in \mathbb{I} \colon I \models \varphi_0\} = \Phi^{-1}(D_0)$. ⌟

From Claim 4.3 we get in particular that $P_{\mathcal{I}}(\varphi_0) = P_{\mathcal{I}}(\Phi^{-1}(D_0))$ and $P_{\mathcal{I}}(\varphi_0 \mid \varphi) = p_0$ since $\mathcal{D} = \Phi(\mathcal{I} \mid \varphi)$.

Now let $\psi := \varphi \wedge \neg\varphi_0$. Then, ψ is an **FO**-sentence with the property that $P_{\mathcal{I}}(\psi \mid \varphi) = P_{\mathcal{I}}(\neg\varphi_0 \mid \varphi) = 1 - P_{\mathcal{I}}(\varphi_0 \mid \varphi) = 1 - p_0$. Moreover, for all $D \in \mathbb{D} \setminus \{D_0\}$ it holds that

$$\Pr_{I\sim\mathcal{I}}\big(\Phi(I) = D \mid I \models \psi\big) = P_{\mathcal{I}}\big(\Phi^{-1}(D) \mid \psi\big) = \frac{P_{\mathcal{I}}\big(\Phi^{-1}(D) \cap \psi\big)}{P_{\mathcal{I}}(\psi)} \\ \overset{(*)}{=} \frac{P_{\mathcal{I}}\big(\Phi^{-1}(D) \cap \varphi\big)}{P_{\mathcal{I}}(\psi \cap \varphi)} = \frac{P_{\mathcal{I}}\big(\Phi^{-1}(D) \mid \varphi\big)}{P_{\mathcal{I}}(\psi \mid \varphi)} = \frac{P_{\mathcal{D}}(\{D\})}{1 - p_0}.$$

For the numerator in the step belonging to $(*)$, note that for any $I \in \mathbb{I}$ with $I \in \Phi^{-1}(D)$, it holds that $I \not\models \varphi_0$ and thus, $I \models \psi$ if and only if $I \models \varphi$. The denominator can be transformed as is done since ψ entails φ. For the step afterwards, we multiplied both the numerator and the denominator with $P_{\mathcal{I}}(\varphi)$. The last equality in the above calculation again holds since $\mathcal{D} = \Phi(\mathcal{I} \mid \varphi)$.

Since $P_{\mathcal{I}}(\psi) = P_{\mathcal{I}}(\psi \mid \varphi) \cdot P_{\mathcal{I}}(\varphi) = (1 - p_0) \cdot P_{\mathcal{I}}(\varphi)$, we have $0 < P_{\mathcal{I}}(\psi) < 1$. Therefore, we can choose $k \in \mathbb{N}_+$ sufficiently large such that $\big(1 - P_{\mathcal{I}}(\psi)\big)^k < p_0$. We construct a **TI**-PDB $\mathcal{I}^{(k)}$ that consists of k independent copies of $\mathcal{I}$. If τ denotes the schema of $\mathcal{I}$, the schema τ_k of $\mathcal{I}^{(k)}$ is defined as follows: For each $R \in \tau$ of arity r, τ_k contains a distinguished relation symbol R' of arity $r + 1$. Additionally, τ_k contains a binary relation $R_\leq$. The facts and marginal probabilities are defined as follows: for all $i, j \in \{1, \dots, k\}$ such that $i \leq j$, we have $(i, j) \in R_\leq(i, j)$ with marginal probability 1; for $i \in \{1, \dots, k\}$ and $R(a_1, \dots, a_r) \in T(\mathcal{I})$ with marginal probability p in $\mathcal{I}$, we have $R'(i, a_1, \dots, a_r) \in T(\mathcal{I}^{(k)})$ with marginal probability p in $\mathcal{I}^{(k)}$; and all other facts have probability 0. Note that $\mathcal{I}^{(k)}$ is then a well-defined **TI**-PDB since the sum of the fact probabilities is bounded by $k^2 + k \cdot \mathrm{E}_{\mathcal{I}}|\cdot| < \infty$.

If I is an instance of $\mathcal{I}^{(k)}$ and $1 \leq i \leq k$, we let $I[i]$ denote the instance that is obtained by selecting the facts with identifier i and projecting the identifier out. That is,

$$I[i] \coloneqq \{R(\bar{a}) : R'(i, \bar{a}) \in I\} \in \mathbb{I}.$$

Let I be an instance of $\mathcal{I}^{(k)}$. An index $i \in \{1, \dots, k\}$ is called *suitable* if $I[i] \models \psi$. We call I a *representation* if a suitable i exists. Note that $\Pr_{I\sim\mathcal{I}^{(k)}}(I[i] \models \psi) = \Pr_{I\sim\mathcal{I}}(I \models \psi) = P_{\mathcal{I}}(\psi)$ by the definition of the marginal probabilities in $\mathcal{I}^{(k)}$. Therefore,

$$\begin{aligned} q \coloneqq \Pr_{I\sim\mathcal{I}^{(k)}}(I \text{ is a representation}) &= 1 - \Pr_{I\sim\mathcal{I}^{(k)}}(\text{no } i \text{ is suitable}) \\ &= 1 - \big(1 - P_{\mathcal{I}}(\psi)\big)^k > 1 - p_0. \end{aligned} \tag{2}$$

We say that an $\mathcal{I}^{(k)}$-instance I *represents* a $\mathcal{D}$-instance D if I is a representation and $\Phi(I[i_0]) = D$ where i_0 is the smallest suitable index in I. For all $D \in \mathbb{D}$,

$$\begin{aligned} &\Pr_{I\sim\mathcal{I}^{(k)}}\big(I \text{ represents } D \mid I \text{ is a representation}\big) \\ &\quad= \Pr_{I\sim\mathcal{I}^{(k)}}(\Phi(I[i_0]) = D \mid i_0 \text{ is suitable in } I) \\ &\quad= \Pr_{I\sim\mathcal{I}}(\Phi(I) = D \mid I \models \psi) = \tfrac{P_{\mathcal{D}}(\{D\})}{1-p_0}. \end{aligned}$$

Therefore, (in the following, rep. stands for representation)

$$\begin{aligned} &\Pr_{I\sim\mathcal{I}^{(k)}}\big(I \text{ represents } D\big) \\ &\quad= \Pr_{I\sim\mathcal{I}^{(k)}}\big(I \text{ is a rep.}\big) \cdot \Pr_{I\sim\mathcal{I}^{(k)}}\big(I \text{ represents } D \mid I \text{ is a rep.}\big) \\ &\quad= q \cdot \tfrac{P_{\mathcal{D}}(\{D\})}{1-p_0} \end{aligned}$$

We now construct a **TI**-PDB $\mathcal{J}$. We let $q_0 \coloneqq (p_0 - 1 + q)/q$ and note that $0 < q_0 < 1$ by Equation (2). We define $\mathcal{J}$ by adding a relation $R_\perp$ to $\mathcal{I}^{(k)}$ that contains one fact, which we denote $\perp$, with marginal probability q_0. All other relations and marginal probabilities are the same as in $\mathcal{I}^{(k)}$. For every instance $J \sim \mathcal{J}$ we let $J_\perp$ denote its restriction to the relations of $\mathcal{I}^{(k)}$. We say that $J \sim \mathcal{J}$ *represents* D_0 if either $J_\perp$ is not a representation or $J_\perp$ is a representation and $\perp \in J$. We say that $J \sim \mathcal{J}$ *represents* $D \neq D_0$ if J does not represent D_0 and its restriction $J_\perp$ represents D. Observe that every $J \sim \mathcal{J}$ represents exactly one instance $D \in \mathbb{D}$. We have

$$\Pr_{J\sim\mathcal{J}}\big(J \text{ represents } D_0\big) = (1-q) + (q \cdot q_0) = p_0,$$

and, for $D \neq D_0$, it holds that

$$\begin{aligned} &\Pr_{J\sim\mathcal{J}}\big(J \text{ represents } D\big) = \Pr_{J\sim\mathcal{J}}\big(\perp \notin J \text{ and } J_\perp \text{ represents } D\big) \\ &= (1-q_0)q\tfrac{P_{\mathcal{D}}(\{D\})}{1-p_0} = \big(q - (p_0 - 1 + q)\big)\tfrac{P_{\mathcal{D}}(\{D\})}{1-p_0} = P_{\mathcal{D}}(\{D\}). \end{aligned}$$

Finally, there exists an **FO**-view Φ' that maps each $J \sim \mathcal{J}$ to the $D \in \mathbb{D}$ it represents. The single, fixed instance D_0 can be dealt with separately using a hard-coded description. For the other instances, the view needs to extract the original relations from the R'-relations we introduced according to the smallest suitable index. Getting a hold on the latter is achievable using the order relation $R_\leq$ we introduced in the construction of $\mathcal{I}^{(k)}$.

Thus, $\mathcal{D} \in$ **FO**(**TI**) as witnessed by $\mathcal{J}$ and Φ'. □

Remark 4.4. An implication of Theorem 4.1 is that **FO**(**TI**) is closed under **FO**-conditioning regardless of whether the conditioning is done before or after applying the view. This holds since **FO**(**TI**) ⊆ **FO**(**TI**)|**FO** ⊆ **FO**(**TI**|**FO**) ⊆ **FO**(**TI**). To prove the middle containment, an **FO**-sentence over the view's output schema can be translated to an **FO**-sentence over the input schema by replacing every relation with the **FO**-view defining it.

5 POWER OF TI REPRESENTATIONS

In this section we establish some positive results regarding representability. In Section 5.1 we prove a sufficient criterion on the growth rate of the probabilities and conclude also that PDBs of bounded instances size are in **FO**(**TI**). Section 5.2 is devoted to **BID**-PDBs.

5.1 A Condition on Sizes and Probabilities

Towards characterizing representability of PDBs using **FO**-views over **TI**-PDBs, so far we have from Proposition 3.4 that the finite moments property is a necessary condition. In this subsection, we present a sufficient condition for membership in **FO**(**TI**), taking the detour over membership in **FO**(**TI** | **FO**). At the end of this subsection, we discuss the implications of this condition.

Lemma 5.1. *Let $\mathcal{D} = (\mathbb{D}, P)$ be a PDB. If there exists $c \in \mathbb{N}_+$ s. t.*

$$\sum_{D\in\mathbb{D}\setminus\{\emptyset\}} |D| \cdot P\big(\{D\}\big)^{\frac{c}{|D|}} < \infty, \tag{3}$$

*then $\mathcal{D} \in$ **FO**(**TI** | **FO**).*

The intuition behind Lemma 5.1 is as follows. Assume for the moment that $c = 1$. If $D \in \mathbb{D}$ and t is a fact appearing in D, we copy t to the **TI**-PDB we construct, tagged by an identifier of the instance D. This creates a copy of t for every instance in which it appears. We say that a subset of the constructed facts is a representation if there exists exactly one instance for which all tagged facts appear. We set the **FO**-condition to imply that the possible world is a representation, while the **FO**-view extracts the original instance from its tagged tuples. We select the probabilities of the facts in our construction to ensure that each possible world is obtained with the correct probability. For the obtained construction to be a valid **TI**-PDB, the sum of all marginal probabilities must be finite, and this is the reason we require the condition described in Equation (3) (with $c = 1$). To relax this condition, we allow $c \geq 1$ by encoding c old facts into a single new fact.

Proof (Lemma 5.1). Let $\mathcal{D} = (\mathbb{D}, P)$ be a PDB with instances $\mathbb{D} = \{D_0, D_1, D_2, \dots\}$ where $D_0 = \emptyset$ is the empty instance. Let $s_i \coloneqq |D_i|$ denote the size of D_i and $p_i \coloneqq P(\{D_i\})$ its probability in $\mathcal{D}$ for all $i \in \mathbb{N}$. Without loss of generality, we assume that $p_i > 0$ for all $i \in \mathbb{N}$. (If $p_0 = 0$, we just need to consider $i \in \mathbb{N}_+$. The proof then works the same way without the special treatment of $i = 0$.)

Let $c \in \mathbb{N}_+$ such that (3) holds. For simplicity, we assume that $\mathcal{D}$ is of schema τ consisting of a single, r-ary relation R. The proof can easily be generalized to arbitrary schemas.

We let $\widehat{\tau} \coloneqq \{\widehat{R}\}$ be a new relational schema, where $\widehat{R}$ is $(3+cr)$-ary. Let $\widehat{\mathbb{U}} \coloneqq \mathbb{U} \cup \{\perp\}$. Denote by $t_{i,j}$ the jth fact of $D_i \in \mathbb{D}$ (according to some arbitrary but fixed order on D_i), stripped of its relation symbol R.

If $j > s_i$, we set $t_{i,j} \coloneqq (\perp, \dots, \perp)$. We define

$$N_{i,j} \coloneqq \begin{cases} j+1 & \text{if } (j+1)\cdot c < s_i \text{ and} \\ \perp & \text{if } (j+1)\cdot c \geq s_i. \end{cases}$$

Using this notation we define a set $\widehat{T}$ of $(\widehat{\tau}, \widehat{\mathbb{U}})$-facts:

$$\widehat{T} := \Big\{ \widehat{R}\big(i, j, N_{i,j}, t_{i,jc+1}, \dots, t_{i,jc+c}\big) : \\ i, j \in \mathbb{N}, j \le \max\left(\left\lceil \tfrac{s_i}{c} \right\rceil - 1, 0\right) \Big\}.$$

Every fact in $\widehat{T}$ corresponds to a segment of (up to) c facts in a database instance of $\mathbb{D}$ (see Figure 3).

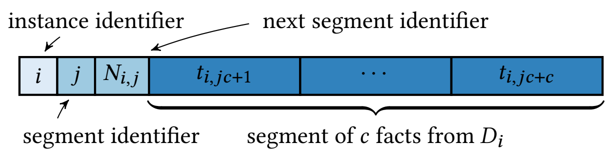

Figure 3: Structure of the new facts.

The first component is the index of the original instance in which the facts appear (*instance identifier*). The second component contains the index of the segment (*segment identifier*) whereas the third entry contains the index of the next segment (*next segment identifier*). The $r \cdot c$ remaining entries contain the c facts of the segment from the original instance. If there are not enough facts left to fill the c fact slots, they are filled with dummy facts $(\bot, \dots, \bot)$.

We proceed to define a **TI**-PDB $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$ of schema $\widehat{\tau}$ over $\widehat{\mathbb{U}}$ such that $T(\mathcal{I}) = \widehat{T}$. Thus, $\mathbb{I}$ is the set of finite subsets of $\widehat{T}$. For every $i \in \mathbb{N}$ we let $\widehat{D}_i \in \mathbb{I}$ be the instance that contains exactly the facts with instance identifier i. We say that $I \in \mathbb{I}$ *represents* $D_i \in \mathbb{D}$ if $\widehat{D}_i \subseteq I$ but $\widehat{D}_j \not\subseteq I$ for all $j \ne i$. We say that $I \in \mathbb{I}$ is a *representation* if it represents some $D_i \in \mathbb{D}$.

Note that the instances $\widehat{D}_i$, $i \in \mathbb{N}$ are pairwise disjoint and cover $\widehat{T}$. Thus, for every $t \in \widehat{T}$ there exists a unique $i := i(t)$ such that $t \in \widehat{D}_i$. Moreover, $\widehat{s}_i := |\widehat{D}_i| = \lceil \frac{s_i}{c} \rceil \ge 1$ for all $i > 0$ and $\widehat{s}_0 = 1$.

For $t \in \widehat{T}$ we define

$$q_t := \left(\frac{p_{i(t)}}{1+p_{i(t)}}\right)^{1/\widehat{s}_{i(t)}},$$

and let $\mathcal{I}$ be the **TI**-PDB spanned by these marginal probabilities. Observe that

$$\sum_{t \in \widehat{T}} q_t = \frac{p_0}{1+p_0} + \sum_{i=1}^{\infty} \left(\frac{p_i}{1+p_i}\right)^{\left\lceil \frac{s_i}{c} \right\rceil^{-1}} \le 1 + \sum_{i=1}^{\infty} \left\lceil \frac{s_i}{c} \right\rceil \cdot p_i^{\left\lceil \frac{s_i}{c} \right\rceil^{-1}}.$$

The convergence of the sum in the last expression is equivalent to (3) (for a proof see the extended version of this paper [12]). Therefore, $\sum_{t \in \widehat{T}} q_t < \infty$, so $\mathcal{I}$ is well-defined.

Let $q_i := \frac{p_i}{1+p_i}$ and observe that for all $i > 0$

$$\Pr_{I \sim \mathcal{I}}\left(\widehat{D}_i \subseteq I\right) = \prod_{t \in \widehat{D}_i} q_t = \left(q_i^{1/\widehat{s}_i}\right)^{\widehat{s}_i} = q_i$$

and, on the other hand, $\Pr_{I \sim \mathcal{I}}\left(\widehat{D}_0 \subseteq I\right) = q_{t_0} = q_0$ where t_0 is the unique fact in $\widehat{D}_0$. Moreover, note that $0 < q_i < 1$ for all $i \in \mathbb{N}$.

Since $\sum_{i=0}^{\infty} q_i \le \sum_{i=0}^{\infty} p_i < \infty$, it holds that

$$Z := \prod_{i=0}^{\infty} (1 - q_i) \in (0, 1].$$

Then for all $i \in \mathbb{N}$ we have

$$\Pr_{I \sim \mathcal{I}}\left(I \text{ represents } D_i\right) = q_i \cdot \textstyle\prod_{j \ne i}(1 - q_j) = Z \cdot \frac{q_i}{1-q_i} = Z \cdot p_i,$$

$$\Pr_{I \sim \mathcal{I}}\left(I \text{ is a representation}\right) = \textstyle\sum_{i=0}^{\infty} Z \cdot p_i = Z,$$

and, moreover,

$$\begin{aligned} &\Pr_{I \sim \mathcal{I}}\left(I \text{ represents } D_i \mid I \text{ is a representation}\right) \\ &= \frac{\Pr_{I \sim \mathcal{I}}\left(I \text{ represents } D_i\right)}{\Pr_{I \sim \mathcal{I}}\left(I \text{ is a representation}\right)} = p_i. \end{aligned} \tag{4}$$

In order to establish $\mathcal{D} \in \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$, it now suffices to prove the following claim.

Claim 5.2. (1) *There exists an* **FO**-*sentence* φ *such that* $I \models \varphi$ *if and only if* I *is a representation for all* $I \in \mathbb{I}$.
(2) *There exists an* **FO**-*view* Φ, *mapping* I *to the database it represents for all* $I \in \mathbb{I}$ *that are representations.*

Proof Sketch. (1) In order to check whether I is a representation, we need to check whether there exists some i such that I represents i. Now I represents i if and only if I contains $\widehat{D}_i$ but does not contain $\widehat{D}_j$ for all $j \ne i$. One can check if $\widehat{D}_i \subseteq I$ (and thus if $\widehat{D}_j \not\subseteq I$ for $j \ne i$) by checking the following: I needs to contain a fact starting with instance identifier i and segment identifier 0; moreover, if I contains a fact with instance identifier i, segment identifier j and next segment identifier $j' \ne \bot$, then I contains a fact with instance identifier i and segment identifier j' (j' will be $j+1$ by the definition of $\widehat{T}$).
(2) Recall that the facts of $\widehat{T}$ contain up to c original facts in succession. These can be recovered by a union of c projections, under omitting the $\bot$ entries. ⌟

Letting Φ and φ be as in Claim 5.2, Equation (4) implies that $\Pr_{I \sim \mathcal{I}}\left(\Phi(I) = D_i \,|\, I \models \varphi\right) = p_i = P\big(\{D_i\}\big)$, so $\mathcal{D} \in \mathbf{FO}(\mathbf{TI} | \mathbf{FO})$. □

From Lemma 5.1 and Theorem 4.1 we get the following.

Theorem 5.3. *Let* $\mathcal{D} = (\mathbb{D}, P)$ *be a PDB. If there exists* $c \in \mathbb{N}_+$ *s. t.*

$$\sum_{D \in \mathbb{D} \setminus \{\emptyset\}} |D| \cdot P\big(\{D\}\big)^{\frac{c}{|D|}} < \infty,$$

then $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$.

In the remainder of this subsection, we discuss applications of Theorem 5.3. We say that a PDB is of *bounded instance size* if there exists some fixed bound c such that all possible worlds have size at most c. Theorem 5.3 implies that every PDB of bounded instance size is representable as an **FO**-view over a **TI**-PDB: The construction in the proof of Lemma 5.1 then represents every instance of the size-bounded PDB by a single segmented fact.

Corollary 5.4. *Every PDB of bounded instance size is in* **FO**(**TI**).

Proof. Let $\mathcal{D} = (\mathbb{D}, P)$ be a PDB and take c to be the bound on its instance sizes. Then, by Theorem 5.3, $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ since

$$\sum_{D \in \mathbb{D} \setminus \{\emptyset\}} |D| \cdot P\big(\{D\}\big)^{\frac{c}{|D|}} \le c \sum_{D \in \mathbb{D}} P\big(\{D\}\big) = c < \infty. \quad \square$$

Remark 5.5. PDBs of bounded instance size are not to be confused with finite PDBs. In particular, they can have infinite domains. As an example, consider the PDB $\mathcal{D} = (\mathbb{D}, P)$ over a schema consisting of a single unary relation symbol R with $\mathbb{D} = \{D_1, D_2, \ldots\}$, where D_n contains a single fact $R(n)$ and $P(\{D_n\}) = \frac{6}{n^2\pi^2}$. In this example, the instance size is bound by 1 though the domain is infinite.

Unfortunately, we have no full characterization of $\mathbf{FO}(\mathbf{TI})$ yet: According to Corollary 5.4, PDBs with bounded instance size are in $\mathbf{FO}(\mathbf{TI})$, and, according to Proposition 3.4, PDBs with infinite expected size, or some infinite size moment, are not in $\mathbf{FO}(\mathbf{TI})$. However, there are PDBs with unbounded size that have the finite moments property. Example 3.9 shows that not all such PDBs are in $\mathbf{FO}(\mathbf{TI})$. The following example shows that some are.

Example 5.6. Let $x \coloneqq \sum_{i=1}^{\infty} 2^{-i^2}$. Then $0 < x < \sum_{i=1}^{\infty} 2^{-i} = 1$. Now define $\mathcal{D} = (\mathbb{D}, P)$ with $\mathbb{D} = \{D_1, D_2, \ldots\}$ with $|D_i| = i$ and let $P(\{D_i\}) = \frac{1}{x} \cdot 2^{-i^2}$. Since $\sum_{D \in \mathbb{D}} P(\{D\}) = 1$, $\mathcal{D}$ is a PDB. Note that $\left(\frac{1}{x}\right)^{\alpha} \leq \frac{1}{x}$ for all $\alpha \in (0, 1]$ because $\frac{1}{x} > 1$. Then

$$\sum_{D \in \mathbb{D}} |D| \cdot P(\{D\})^{\frac{1}{|D|}} = \sum_{i=1}^{\infty} i \cdot \left(\tfrac{1}{x} \cdot 2^{-i^2}\right)^{\frac{1}{i}} \leq \tfrac{1}{x} \sum_{i=1}^{\infty} i \cdot 2^{-i} = \tfrac{2}{x} < \infty.$$

Therefore $\mathcal{D}$ satisfies the condition of Theorem 5.3 for $c = 1$, so $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$. Yet, $\mathcal{D}$ is of unbounded instance size.

The converse of Theorem 5.3 does not hold. In fact, even though every $\mathbf{TI}$-PDB is trivially in $\mathbf{FO}(\mathbf{TI})$, some $\mathbf{TI}$-PDBs violate the condition of the theorem:

Example 5.7. Consider the $\mathbf{TI}$-PDB $\mathcal{I}$ with facts $t_1, t_2, \ldots$ that is spanned by the marginal probabilities $p_i = 1/(i^2 + 1)$ for $i \in \mathbb{N}_+$. Then $\mathcal{I}$ has the finite moments property according to Proposition 3.2 but $\mathcal{I}$ does not satisfy the condition of Theorem 5.3. (This is shown in detail in the extended version [12].)

Thus, we still exhibit a proper gap between our conditions for containment in $\mathbf{FO}(\mathbf{TI})$.

5.2 Block-Independent Disjoint Databases

In this section we prove that every $\mathbf{BID}$-PDB can be represented as an $\mathbf{FO}$ view over a $\mathbf{TI}$-PDB. First note that this does not follow from the previous section, as there are $\mathbf{BID}$-PDBs that are not naturally tuple-independent and such that Theorem 5.3 does not apply to them. Consider the $\mathbf{BID}$-PDB $\mathcal{D}'$ with blocks $(B_i)_{i \in \mathbb{N}_+}$ where each block contains exactly two facts, both of marginal probability $1/\big(2(i^2 + 1)\big)$. Similar to Example 5.7, $\mathcal{D}'$ does not satisfy the condition of Theorem 5.3 (see [12] for a proof).

Again, we take the detour via conditional representations and show that every $\mathbf{BID}$-PDB can be represented as an $\mathbf{FO}$-view of an $\mathbf{FO}$-conditioned $\mathbf{TI}$-PDB.

Lemma 5.8. $\mathbf{BID} \subseteq \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$.

Proof. The basic idea of the proof is as follows. We start with a $\mathbf{BID}$-PDB and forget about its block structure. To compensate, every fact is equipped with an identifier indicating to which block it belongs. The resulting PDB is then treated as a $\mathbf{TI}$-PDB. We define the condition to reject the instances that violate the intended block structure (using the block identifiers to find violations), and then the view can simply project out the block identifiers. The marginal probabilities are carefully chosen to guarantee that this process results in the same probability distribution as in the original PDB.

Let $\mathcal{D} = (\mathbb{D}, P_{\mathcal{D}})$ be a $\mathbf{BID}$-PDB with blocks $\{B_1, B_2, \ldots\}$ such that block B_i contains the facts $\{t_{i,1}, t_{i,2}, \ldots\}$. To simplify notation, assume that B_i is countably infinite for all $i \in \mathbb{N}_+$. This is without loss of generality, since we can add infinitely many artificial facts of marginal probability 0 to every block of $\mathcal{D}$. For all $i, j \in \mathbb{N}_+$, let $p_{i,j} \coloneqq \Pr_{D \sim \mathcal{D}}(t_{i,j} \in D)$. By Theorem 2.6, $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} p_{i,j} < \infty$. The probability to choose no fact from block B_i is $r_i = 1 - \sum_{j=1}^{\infty} p_{i,j}$, called the *residual* (probability mass) in B_i.

We know that the residuals of $\mathbf{BID}$-PDBs form a convergent series [27, see Lemma 4.14]. In particular, there are only finitely many residuals $r_i < \varepsilon$ for every $\varepsilon \in (0, 1)$. Thus, we may assume without loss of generality that the blocks B_i are numbered in increasing order of the r_i. Suppose m is the nonnegative integer with $r_i = 0$ if and only if $1 \leq i \leq m$.

We construct a conditional, tuple-independent representation of $\mathcal{D}$ by altering every relation in the schema of $\mathcal{D}$ to contain an additional *block identifier* attribute. The facts of our new $\mathbf{TI}$-PDB $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$ are the facts from $\mathcal{D}$, augmented by the number of their block in the additional attribute. The marginal probability $q_{i,j} = \Pr_{I \sim \mathcal{I}}(t_{i,j} \in I)$ of $t_{i,j}$ in $\mathcal{I}$ is defined in the following way:

$$q_{i,j} \coloneqq \begin{cases} \dfrac{p_{i,j}}{1 + p_{i,j}} & r_i = 0 \\[2ex] \dfrac{p_{i,j}}{r_i + p_{i,j}} & r_i > 0 \end{cases}$$

We show that these marginal probabilities indeed span a well-defined $\mathbf{TI}$-PDB. Consider $t_{i,j}$ with i and j arbitrary.

- If $r_i > 0$ then $q_{i,j} = \frac{p_{i,j}}{r_i + p_{i,j}} \leq \frac{p_{i,j}}{r_i} \leq \frac{p_{i,j}}{r_{m+1}}$.
- If $r_i = 0$, then $q_{i,j} = \frac{p_{i,j}}{1 + p_{i,j}} \leq \frac{p_{i,j}}{1} \leq \frac{p_{i,j}}{r_{m+1}}$.

Thus, $\mathcal{I}$ is a well-defined $\mathbf{TI}$-PDB since

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} q_{i,j} \leq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{p_{i,j}}{r_{m+1}} = \frac{1}{r_{m+1}} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} p_{i,j} < \infty.$$

The instances of the new schema that are of interest are those belonging to $\mathcal{D}$-instances that obey the block structure of $\mathcal{D}$. The following claim asserts that this property is $\mathbf{FO}$-definable.

Claim 5.9. *There exists an* $\mathbf{FO}$*-sentence* φ *such that for all* $I \in \mathbb{I}$ *it holds that* $I \models \varphi$ *if and only if* I *contains*

- *at most one fact with block identifier* i *for all* $i \in \mathbb{N}_+$*; and*
- *exactly one fact with block identifier* i *for all* $i \leq m$*.*

Proof. For simplicity, assume that the schema of $\mathcal{D}$ consists of a single, unary relation symbol R. Let

$$\varphi \coloneqq \big(\forall j \, \exists^{\leq 1} x \colon R'(x, j)\big) \wedge \textstyle\bigwedge_{i=1}^{m} \exists^{=1} x \colon R'(x, i)$$

where R' is the augmented version of R. Note that the $\exists^{\leq 1}$ and $\exists^{=1}$ quantifiers are expressible in plain $\mathbf{FO}$. The formula above can easily be generalized to arbitrary schemas. ⌟

Let Φ be the view that projects the block identifier out of the facts of $\mathbb{I}$. The PDB $\mathcal{I}$ together with the condition φ and the $\mathbf{FO}$-view Φ is our representation of $\mathcal{D} = (\mathbb{D}, P_{\mathcal{D}})$. It is left to show that the probability distribution we obtain this way is the same as in the original PDB. That is,

$$\Pr_{I\sim\mathcal{I}}\big(\Phi(I) = D \mid I \models \varphi\big) = P_{\mathcal{D}}\big(\{D\}\big) \quad \text{for all} \quad D \in \mathbb{D}.$$

We first prove that it suffices to check every block independently. Given a database D, we denote by $D[i]$ the restriction of D to B_i. Since the blocks are independent in $\mathcal{D}$, for all $D \in \mathbb{D}$,

$$P_{\mathcal{D}}\big(\{D\}\big) = \Pr_{D'\sim\mathcal{D}}\big(\forall i : D'[i] = D[i]\big) = \prod_{i\in\mathbb{N}_+} \Pr_{D'\sim\mathcal{D}}\big(D'[i] = D[i]\big).$$

Let us now inspect the conditional representation. We denote by φ_i the condition that if $r_i > 0$, there is at most one element of B_i, and if $r_i = 0$, there is exactly one element of B_i. (Note that $I \sim \mathcal{I}$ satisfies φ if and only if $I \models \varphi_i$ for all $i \in \mathbb{N}_+$.) Let $D \in \mathbb{D}$ be an instance with positive probability in $\mathcal{D}$. By our definitions, if $\Phi(I) = D$, then in particular $I \models \varphi$. Since the facts and therefore also the blocks are independent in $\mathcal{I}$,

$$\Pr_{I\sim\mathcal{I}}(\Phi(I) = D \mid I \models \varphi) = \frac{\Pr_{I\sim\mathcal{I}}(\Phi(I) = D \text{ and } I \models \varphi)}{\Pr_{I\sim\mathcal{I}}(I \models \varphi)}$$
$$= \frac{\Pr_{I\sim\mathcal{I}}(\forall i : \Phi(I)[i] = D[i])}{\Pr_{I\sim\mathcal{I}}(\forall i : I[i] \models \varphi_i)} = \frac{\prod_{i\in\mathbb{N}_+} \Pr_{I\sim\mathcal{I}}(\Phi(I)[i] = D[i])}{\prod_{i\in\mathbb{N}_+} \Pr_{I\sim\mathcal{I}}(I[i] \models \varphi_i)}.$$

Therefore, in order to show that the probability distribution over the conditional representation is the same as the original PDB, it is enough to show that for all $D \in \mathbb{D}$ and every block B_i:

$$\frac{\Pr_{I\sim\mathcal{I}}(\Phi(I)[i] = D[i])}{\Pr_{I\sim\mathcal{I}}(I[i] \models \varphi_i)} = \Pr_{D'\sim\mathcal{D}}(D'[i] = D[i]). \tag{5}$$

Thus, fix an arbitrary instance $D \in \mathbb{D}$ with positive probability in $\mathcal{D}$ and $i \geq 1$. We denote $Z_i := \prod_{t_{i,j}\in B_i} (1 - q_{i,j})$. Note that the probability of selecting only $t_{i,j}$ from block B_i is

$$q_{i,j} \cdot \prod_{t_{i,k}\in B_i,\, k\neq j} \big(1 - q_{i,k}\big) = \frac{q_{i,j}}{1 - q_{i,j}} Z_i.$$

We distinguish cases according to whether B_i has a positive residual and, if so, whether D contains a fact from B_i.

(1) If $r_i = 0$, the probability of having only fact $t_{i,j}$ from B_i is

$$\tfrac{q_{i,j}}{1-q_{i,j}} Z_i = \tfrac{p_{i,j}}{1+p_{i,j}} \cdot \Big(1 - \tfrac{p_{i,j}}{1+p_{i,j}}\Big)^{-1} \cdot Z_i = p_{i,j} \cdot Z_i.$$

Denote by $t_{i,k}$ the unique fact from B_i in D. Then

$$\frac{\Pr_{I\sim\mathcal{I}}(\Phi(I)[i] = D[i])}{\Pr_{I\sim\mathcal{I}}(I[i] \models \varphi_i)} = \frac{p_{i,k} Z_i}{\sum_{j\in\mathbb{N}_+} p_{i,j} Z_i} = \frac{p_{i,k}}{\sum_{j\in\mathbb{N}_+} p_{i,j}}$$
$$= p_{i,k} = \Pr_{D'\sim\mathcal{D}}(D'[i] = D[i]).$$

(2) If $r_i > 0$, the probability of having no fact from B_i in $\mathcal{I}$ is Z_i, and the probability of having only fact $t_{i,j}$ from B_i is

$$\tfrac{q_{i,j}}{1-q_{i,j}} Z_i = \tfrac{p_{i,j}}{r_i+p_{i,j}} \cdot \Big(1 - \tfrac{p_{i,j}}{r_i+p_{i,j}}\Big)^{-1} Z_i = \tfrac{p_{i,j}}{r_i} Z_i.$$

Therefore, the probability of satisfying the condition φ_i is:

$$\Pr_{I\sim\mathcal{I}}\big(I[i] \models \varphi_i\big) = Z_i + \sum_{j\in\mathbb{N}_+} \tfrac{p_{i,j}}{r_i} Z_i$$
$$= \tfrac{Z_i}{r_i}\Big(r_i + \sum_{j\in\mathbb{N}_+} p_{i,j}\Big) = \tfrac{Z_i}{r_i}.$$

(a) In case there exists a fact $t_{i,k}$ from B_i in D, it holds that

$$\frac{\Pr_{I\sim\mathcal{I}}(\Phi(I)[i] = D[i])}{\Pr_{I\sim\mathcal{I}}(I[i] \models \varphi_i)} = \frac{\frac{p_{i,k}}{r_i} Z_i}{\frac{Z_i}{r_i}}$$
$$= p_{i,k} = \Pr_{D'\sim\mathcal{D}}(D'[i] = D[i]).$$

(b) In case no fact from B_i appears in D, it holds that

$$\frac{\Pr_{I\sim\mathcal{I}}(\Phi(I)[i] = D[i])}{\Pr_{I\sim\mathcal{I}}(I[i] \models \varphi_i)} = \frac{Z_i}{\frac{Z_i}{r_i}}$$
$$= r_i = \Pr_{D'\sim\mathcal{D}}(D'[i] = D[i]).$$

Since i is arbitrary, this holds for all i. Together, we have verified Equation (5). Thus, $\mathcal{D}$ is an **FO**-view of an **FO**-conditioned **TI**-PDB, witnessed by $\mathcal{I}$, Φ and φ. □

From Lemma 5.8 and Theorem 4.1 we get the desired representability result for **BID**-PDBs.

Theorem 5.10. **BID** ⊆ **FO**(**TI**).

Remark 5.11. Since

$$\mathbf{FO}(\mathbf{BID}) \subseteq \mathbf{FO}(\mathbf{FO}(\mathbf{TI})) = \mathbf{FO}(\mathbf{TI}) \subseteq \mathbf{FO}(\mathbf{BID}),$$

this yields **FO**(**BID**) = **FO**(**TI**). This also entails that **FO**(**BID**) (and **BID** in particular) inherits the finite moments property from **FO**(**TI**).

Similarly to Remark 4.4, **FO**(**BID**) is also closed under **FO**-conditioning, regardless of whether the conditioning is done before or after applying the view. So in particular it holds that

$$\mathbf{FO}(\mathbf{BID} \mid \mathbf{FO}) = \mathbf{FO}(\mathbf{BID}) \mid \mathbf{FO} = \mathbf{FO}(\mathbf{BID}).$$

6 SEEKING LOGICAL REASONS

In the previous sections, we investigated the question *whether* a given PDB is representable in a particular way. In this section, we address the question *why* a PDB is not representable. In Section 3, we proved that some PDBs are not in **FO**(**TI**) by using the probabilities of the possible worlds to find a contradiction to some convergence property of **FO**(**TI**). The goal of this section is to separate such arithmetical reasons for non-representability from purely logical reasons, namely, reasons that hold independently of the probabilities of the possible worlds.

To this aim, we use the notion of incomplete databases [31]. We call a set of database instances over some schema τ an *incomplete database (IDB)* over τ. Given a PDB $\mathcal{D} = (\mathbb{D}, P)$, its *induced incomplete database* $\mathrm{IDB}(\mathcal{D})$ is the set of instances with positive probabilities. That is,

$$\mathrm{IDB}(\mathcal{D}) := \{D \in \mathbb{D} \mid P(D) > 0\}.$$

If $\mathbf{D}$ is a class of PDBs, then $\mathrm{IDB}(\mathbf{D}) := \{\mathrm{IDB}(\mathcal{D}) \mid \mathcal{D} \in \mathbf{D}\}$.

In this section we aim to investigate what we can conclude about the possible representations of a PDB based solely on its induced IDB. In Section 6.1 we show how this can be used to show non-representability of infinite PDBs. In Section 6.2 we show what we can conclude about **FO**(**TI**) for logical reasons.

6.1 Proving Non-Representability

In this section, we explore how to use the induced IDB of a PDB to show that it cannot be represented as some view over a **TI**-PDB.

A view $V\colon \mathbb{D}_1 \to \mathbb{D}_2$ is called *monotone* if for any instances D, D' in $\mathbb{D}_1$, we have that $D \subseteq D'$ implies $V(D) \subseteq V(D')$. For example, **UCQ**- and Datalog views are monotone [2]. In the finite case, the monotonicity of **UCQ**-views is used to prove that some PDBs cannot be represented by a **UCQ**-view over a **TI**-PDB by considering only their induced IDBs [52, Proposition 2.17].

The proof in the reference uses the existence of a maximal possible world, and does therefore not directly extend to the infinite case. However, a simple argument that works in both the finite and the infinite setting demonstrates how the induced IDB can be used to conclude non-representability in the infinite too. To this aim, we first examine the structure of IDBs induced by **TI**-PDBs.

Observation 6.1. *Let $\mathcal{I}$ be a* **TI***-PDB with fact set $T(\mathcal{I})$ and let p_t denote the marginal probability of $t \in T(\mathcal{I})$. Then $T(\mathcal{I})$ is partitioned into $T_{\text{always}} := \{t\colon p_t = 1\}$, $T_{\text{sometimes}} := \{t\colon 0 < p_t < 1\}$ and $T_{\text{never}} := \{t\colon p_t = 0\}$ and*

$$\mathrm{IDB}(\mathcal{I}) = \big\{T_{\text{always}} \cup T : T \subseteq T_{\text{sometimes}}, |T| < \infty\big\}.$$

We next inspect the effect of applying views over PDBs and observe that stripping PDBs of their probabilities commutes with applying views.

Observation 6.2. *Let $\mathcal{D} = (\mathbb{D}_1, P)$ be a PDB, $\mathbb{D}_2$ a set of database instances and $V\colon \mathbb{D}_1 \to \mathbb{D}_2$ a view. Then, $V(\mathrm{IDB}(\mathcal{D})) = \mathrm{IDB}(V(\mathcal{D}))$.*

Proof. If $D_2 \in V(\mathrm{IDB}(\mathcal{D}))$, then there exists $D_1 \in \mathrm{IDB}(\mathcal{D})$ with $V(D_1) = D_2$. Since $D_1 \in \mathrm{IDB}(\mathcal{D})$, we have $P_{\mathcal{D}}(D_1) > 0$. Hence, $P_{V(\mathcal{D})}(\{D_2\}) = P_{\mathcal{D}}(V^{-1}(\{D_2\})) \geq P_{\mathcal{D}}(\{D_1\}) > 0$. This means that $D_2 \in \mathrm{IDB}(V(\mathcal{D}))$.

We now show the opposite direction. If $D_2 \in \mathrm{IDB}(V(\mathcal{D}))$, then $P_{\mathcal{D}}(V^{-1}(\{D_2\})) = P_{V(\mathcal{D})}(D_2) > 0$. That is, there exists D_1 in $V^{-1}(\{D_2\})$ with $P_{\mathcal{D}}(D_1) > 0$. We have found $D_1 \in \mathrm{IDB}(\mathcal{D})$ with $D_2 = V(D_1)$, and so $D_2 \in V(\mathrm{IDB}(\mathcal{D}))$. □

Using Observation 6.2, we establish a connection between PDB representations and representations of incomplete databases.

Proposition 6.3. *Let* **D** *be a class of PDBs and* **V** *a class of views. For every PDB $\mathcal{D} \in \mathbf{V}(\mathbf{D})$, we have $\mathrm{IDB}(\mathcal{D}) \in \mathbf{V}(\mathrm{IDB}(\mathbf{D}))$.*

Proof. If $\mathcal{D} \in \mathbf{V}(\mathbf{D})$, there exists $V \in \mathbf{V}$ and $\mathcal{D}' \in \mathbf{D}$ with $\mathcal{D} = V(\mathcal{D}')$. Then, $\mathrm{IDB}(\mathcal{D}) = \mathrm{IDB}(V(\mathcal{D}')) = V(\mathrm{IDB}(\mathcal{D}'))$. □

Proposition 6.3 can be interpreted as a purely logical necessary condition for representability: if $\mathrm{IDB}(\mathcal{D}) \notin \mathbf{V}(\mathrm{IDB}(\mathbf{D}))$, then $\mathcal{D} \notin \mathbf{V}(\mathbf{D})$. Next, we build on Proposition 6.3 and Observation 6.1, and we demonstrate how monotonicity can be used to conclude non-representability in the infinite.

Proposition 6.4. *Let* **V** *be a class of monotone views, and let $\mathcal{D}$ be a PDB with two mutually exclusive facts (that is, facts t_1 and t_2 of positive marginal probability with $\Pr_{D\sim\mathcal{D}}(\{t_1, t_2\} \subseteq D) = 0$). Then, $\mathcal{D} \notin \mathbf{V}(\mathbf{TI})$.*

Proof. Assume by contradiction that $\mathcal{D} \in \mathbf{V}(\mathbf{TI})$, and D_1 and D_2 instances in $\mathrm{IDB}(\mathcal{D})$ with $t_i \in D_i$. By Proposition 6.3, we know that $\mathrm{IDB}(\mathcal{D}) \in \mathbf{V}(\mathrm{IDB}(\mathbf{TI}))$. Hence, there is an incomplete database $\mathbb{I} \in \mathrm{IDB}(\mathbf{TI})$ and a view $V \in \mathbf{V}$ with $V(\mathbb{I}) = \mathrm{IDB}(\mathcal{D})$. Since $D_1, D_2 \in \mathrm{IDB}(\mathcal{D})$, there are $I_1, I_2 \in \mathbb{I}$ with $V(I_i) = D_i$. By Observation 6.1, we also have $I_1 \cup I_2 \in \mathbb{I}$ and so $V(I_1 \cup I_2) \in \mathrm{IDB}(\mathcal{D})$. Since V is monotone, $V(I_1 \cup I_2) \supseteq V(I_1) \cup V(I_2) = D_1 \cup D_2 \supseteq \{t_1, t_2\}$, in contradiction to t_1 and t_2 being mutually exclusive. □

Proposition 6.4 immediately yields that $\mathbf{UCQ}(\mathbf{TI})$ does not contain any **BID**-PDBs that are not already tuple-independent. Since $\mathbf{BID} \subseteq \mathbf{FO}(\mathbf{BID}) = \mathbf{FO}(\mathbf{TI})$, this implies $\mathbf{UCQ}(\mathbf{TI}) \subsetneq \mathbf{FO}(\mathbf{TI})$.

6.2 First Order Views

Proposition 6.4 demonstrates that it is possible to determine non-representability of an infinite PDB based on purely logical reasons. However, in the following we show that there are no logical reasons to show that a PDB is not in $\mathbf{FO}(\mathbf{TI})$.

Lemma 6.5. *Let $\mathbb{D}$ be an IDB. Then there exists $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ with $\mathrm{IDB}(\mathcal{D}) = \mathbb{D}$.*

Proof Sketch. Let $\mathbb{D} = \{D_1, D_2, \dots\}$. Let $x_i := (2^i \cdot |D_i|)^{-|D_i|}$ if $|D_i| \neq 0$ and $x_i := 1$ otherwise. We define $P(\{D_i\}) = x_i / \sum_{i=1}^{\infty} x_i$, and show that

$$\sum_{i \in \mathbb{N}_+,\, |D_i| > 0} |D_i| \cdot P(\{D_i\})^{1/|D_i|} < \infty.$$

By Theorem 5.3, this PDB is in $\mathbf{FO}(\mathbf{TI})$. For the details left out above, we refer to the extended version [12]. □

Lemma 6.5 shows that to prove that a PDB is not in $\mathbf{FO}(\mathbf{TI})$, it is not enough to consider which are the possible worlds with positive probability. Such a proof must take the values of the probabilities into account.

So far we discussed logical reasons for showing that a PDB is *not* representable. We can also examine logical reasons for showing that a PDB *is* representable. According to Corollary 5.4, we know that if the underlying incomplete database has bounded size, then the PDB is in $\mathbf{FO}(\mathbf{TI})$ regardless of the probabilities. We next show that if an incomplete database is of unbounded size, we can assign probabilities that result in infinite expected size. According to Proposition 3.4, this implies that the resulting PDB is not in $\mathbf{FO}(\mathbf{TI})$.

Lemma 6.6. *Let $\mathbb{D}$ be an IDB of unbounded instance size. Then there exists a PDB $\mathcal{D}$ with $\mathrm{E}(|\cdot|) = \infty$ such that $\mathrm{IDB}(\mathcal{D}) = \mathbb{D}$.*

Proof. Since the size of instances in $\mathbb{D}$ is unbounded, there exists an infinite sequence of non-empty instances $(D_{i_k})_{k \in \mathbb{N}_+}$ such that $|D_{i_k}|$ is strictly increasing and therefore $|D_{i_k}| \geq k$. We define $P(\{D_{i_k}\}) = \frac{c}{k^2}$ where $c = \frac{3}{\pi^2}$ is the factor that scales the sum of the probabilities to $\frac{1}{2}$. We assign positive probabilities to the instances in $\mathbb{D} \setminus \{D_{i_k} \mid k \in \mathbb{N}_+\}$ such that they also add up to $\frac{1}{2}$. Then the probabilities of all instances add up to 1, but the expected instance size is infinite, since

$$\begin{aligned}\mathrm{E}_{\mathcal{D}}(|\cdot|) &\geq \sum_{k=1}^{\infty} |D_{i_k}| \cdot P(\{D_{i_k}\}) \\ &\geq \sum_{k=1}^{\infty} k \cdot \tfrac{c}{k^2} = \sum_{k=1}^{\infty} \tfrac{c}{k} = \infty.\end{aligned}$$

□

Theorem 6.7 summarizes our results regarding $\mathbf{FO}(\mathbf{TI})$ based purely on logical reasons.

Theorem 6.7. *Let $\mathbb{D}$ be an incomplete database.*

- *If $\mathbb{D}$ has bounded instance size, then for every PDB $\mathcal{D}$ with $\mathrm{IDB}(\mathcal{D}) = \mathbb{D}$, we have $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$.*
- *Otherwise, there exist PDBs $\mathcal{D}_1 \in \mathbf{FO}(\mathbf{TI})$ and $\mathcal{D}_2 \notin \mathbf{FO}(\mathbf{TI})$ such that $\mathrm{IDB}(\mathcal{D}_1) = \mathrm{IDB}(\mathcal{D}_2) = \mathbb{D}$.*

Proof. If $\mathbb{D}$ has bounded instance size, using Corollary 5.4, every PDB with $\mathbb{D}$ as the induced incomplete database is in $\mathbf{FO}(\mathbf{TI})$. Otherwise, $\mathbb{D}$ has unbounded instance size. Lemma 6.5 shows the existence of $\mathcal{D}_1$, and Lemma 6.6 shows the existence of $\mathcal{D}_2$. The latter is not in $\mathbf{FO}(\mathbf{TI})$ due to Proposition 3.4. □

In short, assume we are given a PDB and we want to determine whether it is in $\mathbf{FO}(\mathbf{TI})$. If its induced incomplete database has bounded instance size, we know that the PDB is in $\mathbf{FO}(\mathbf{TI})$. Otherwise, we must consider the probabilities to settle this question.

7 CONCLUDING REMARKS

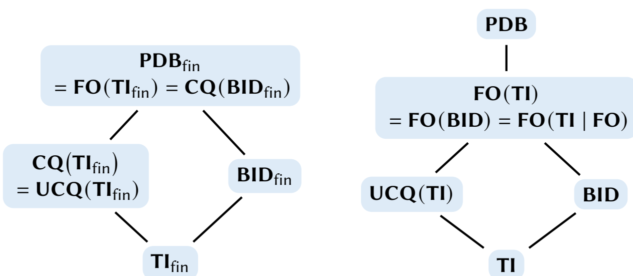

Figure 4: PDB classes with independence assumptions in the finite and the countable setting.

We initiate research on representations of infinite probabilistic databases by systematically studying representability by first-order views over tuple-independent PDBs. Figure 4 compares our findings with the finite setting. In this paper, we showed that $\mathbf{FO}(\mathbf{TI}) = \mathbf{FO}(\mathbf{BID}) = \mathbf{FO}(\mathbf{TI}|\mathbf{FO})$, rendering $\mathbf{FO}(\mathbf{TI})$ quite robust. Although it is known that $\mathbf{FO}(\mathbf{TI}) \subsetneq \mathbf{PDB}$, $\mathbf{FO}(\mathbf{TI})$ can represent any bounded size PDB. While the finite moments property is necessary for containment in $\mathbf{FO}(\mathbf{TI})$, it is not sufficient. Reasons for (non-)representability are twofold in general: arithmetical or purely logical. The class $\mathbf{FO}(\mathbf{TI})$, however, eludes purely logical arguments for non-representability. The mentioned relationships to IDBs extend existing work to infinite PDBs [9, 24].

As of now, our characterization of $\mathbf{FO}(\mathbf{TI})$ is only partial. A full characterization and relationships to other interesting classes of PDBs are left for future work. This may include $\mathbf{CQ}(\mathbf{BID})$, $\mathbf{UCQ}(\mathbf{BID})$ and $\mathbf{TI} \mid \mathbf{FO}$ or using first-order views with inbuilt relations. It could also yield interesting insights to relate our work to recent notions of measuring uncertainty in IDBs [14, 40].

Preliminary results on query evaluation in infinite TI-PDBs were given in [27]. Beyond this, an abstract investigation of query evaluation in infinite PDBs is open as of now.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) under grants GR 1492/16-1 and KI 2348/1-1 "Quantitative Reasoning About Database Queries" and GRK 2236 (UnRAVeL). Nofar Carmeli was supported by the Google PhD Fellowship.

REFERENCES

[1] Serge Abiteboul, T.-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Capturing Continuous Data and Answering Aggregate Queries in Probabilistic XML. *ACM Transactions on Database Systems*, 36(4), 2011. doi:10.1145/2043652.2043658.

[2] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, Boston, MA, USA, 1st edition, 1995.

[3] Parag Agrawal and Jennifer Widom. Continuous Uncertainty in Trio. In *Proceedings of the Third VLDB workshop on Management of Uncertain Data (MUD2009) in conjunction with VLDB 2009, Lyon, France, August 28th, 2009*, volume WP09-14 of *CTIT Workshop Proceedings Series*, pages 17–32. Centre for Telematics and Information Technology (CTIT), University of Twente, The Netherlands, 2009.

[4] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Provenance Circuits for Trees and Treelike Instances. In *Automata, Languages, and Programming (ICALP 2015)*, volume 9135 of *Lecture Notes in Computer Science*, pages 56–68, Berlin, Heidelberg, 2015. Springer. doi:10.1007/978-3-662-47666-6_5.

[5] Antoine Amarilli and İsmail İlkan Ceylan. A Dichotomy for Homomorphism-Closed Queries on Probabilistic Graphs. In *23rd International Conference on Database Theory (ICDT 2020)*, volume 155 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 5:1–5:20, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ICDT.2020.5.

[6] Lyublena Antova, Christoph Koch, and Dan Olteanu. $10^{(10^6)}$ Worlds and Beyond: Efficient Representation and Processing of Incomplete Information. *The VLDB Journal*, 18(5):1021, 2009. doi:10.1007/s00778-009-0149-y.

[7] Vaishak Belle. Open-Universe Weighted Model Counting. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, pages 3701–3708. AAAI Press, 2017.

[8] Michael Benedikt, Evgeny Kharlamov, Dan Olteanu, and Pierre Senellart. Probabilistic XML via Markov Chains. *Proceedings of the VLDB Endowment*, 3(1–2):770–781, 2010. doi:10.14778/1920841.1920939.

[9] Omar Benjelloun, Anish Das Sarma, Alon Halevy, and Jennifer Widom. ULDBs: Databases with Uncertainty and Lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, pages 953–964. VLDB Endowment, 2006.

[10] Stefan Borgwardt, İsmail İlkan Ceylan, and Thomas Lukasiewicz. Ontology-Mediated Queries for Probabilistic Databases. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, (AAAI 2017)*, pages 1063–1069. AAAI Press, 2017.

[11] Zhuhua Cai, Zografoula Vagena, Luis Perez, Subramanian Arumugam, Peter J. Haas, and Christopher Jermaine. Simulation of Database-Valued Markov Chains Using SimSQL. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD 2013)*, pages 637–648, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2463676.2465283.

[12] Nofar Carmeli, Martin Grohe, Peter Lindner, and Christoph Standke. Tuple-Independent Representations of Infinite Probabilistic Databases. arXiv: 2008.09511 [cs.DB], 2020. URL: https://arxiv.org/abs/2008.09511.

[13] İsmail İlkan Ceylan, Adnan Darwiche, and Guy Van den Broeck. Open-World Probabilistic Databases. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, pages 339–348, Palo Alto, CA, USA, 2016. AAAI Press.

[14] Marco Console, Matthias Hofer, and Leonid Libkin. Queries with Arithmetic on Incomplete Databases. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2020)*, pages 179–189, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3375395.3387666.

[15] Nilesh Dalvi, Christopher Ré, and Dan Suciu. Probabilistic Databases: Diamonds in the Dirt. *Communications of the ACM*, 52(7):86–94, 2009. An unpublished, extended version is available at https://homes.cs.washington.edu/~suciu/cacm-paper.pdf. doi:10.1145/1538788.1538810.

[16] Nilesh Dalvi, Christopher Re, and Dan Suciu. Queries and Materialized Views on Probabilistic Databases. *Journal of Computer and System Sciences*, 77(3):473–490, 2011. doi:10.1016/j.jcss.2010.04.006.

[17] Nilesh Dalvi and Dan Suciu. Management of Probabilistic Data: Foundations and Challenges. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2007)*, pages 1–12, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1265530.1265531.

[18] Nilesh Dalvi and Dan Suciu. The Dichotomy of Probabilistic Inference for Unions of Conjunctive Queries. *Journal of the ACM*, 59(6), 2012. doi:10.1145/2395116.

2395119.
[19] Anish Das Sarma, Omar Benjelloun, Alon Halevy, and Jennifer Widom. Working Models for Uncertain Data. In *22nd International Conference on Data Engineering (ICDE 2006)*, volume 1, pages 7:1–7:12, Los Alamitos, CA, USA, 2006. IEEE Computer Society. doi:10.1109/ICDE.2006.174.
[20] Amol Deshpande, Carlos Guestrin, Samuel R. Madden, Joseph M. Hellerstein, and Wei Hong. Model-Driven Data Acquisition in Sensor Networks. In *Proceedings of the 30th VLDB Conference (VLDB 2004)*, pages 588–599. Morgan Kaufmann, 2004. doi:10.1016/B978-012088469-8.50053-X.
[21] Anton Faradjian, Johannes Gehrke, and Philippe Bonnet. GADT: A Probability Space ADT for Representing and Querying the Physical World. In *Proceedings of the 18th International Conference on Data Engineering (ICDE 2002)*, pages 201–211. IEEE Computer Society, 2002. doi:10.1109/ICDE.2002.994710.
[22] Robert Fink and Dan Olteanu. Dichotomies for Queries with Negation in Probabilistic Databases. *ACM Transactions on Database Systems*, 41(1), 2016. doi:10.1145/2877203.
[23] Tal Friedman and Guy Van den Broeck. On Constrained Open-World Probabilistic Databases. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5722–5729. International Joint Conferences on Artificial Intelligence Organization, 2019. doi:10.24963/ijcai.2019/793.
[24] Todd J. Green and Val Tannen. Models for Incomplete and Probabilistic Information. In *Current Trends in Database Technology – EDBT 2006*, Lecture Notes in Computer Science, pages 278–296. Springer, 2006. doi:10.1007/11896548_24.
[25] Eric Gribkoff, Dan Suciu, and Guy Van den Broeck. Lifted Probabilistic Inference: A Guide for the Database Researcher. *Bulletin of the Technical Committee on Data Engineering*, 37(3):6–17, 2014.
[26] Eric Gribkoff, Guy Van den Broeck, and Dan Suciu. The Most Probable Database Problem. In *Proceedings of the First international workshop on Big Uncertain Data (BUDA 2014)*, pages 1–7, 2014.
[27] Martin Grohe and Peter Lindner. Probabilistic Databases with an Infinite Open-World Assumption. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2019)*, pages 17–31, New York, NY, USA, 2019. ACM. Extended version available on arXiv e-prints: arXiv:1807.00607 [cs.DB]. doi:10.1145/3294052.3319681.
[28] Martin Grohe and Peter Lindner. Generalized Independence Assumptions in Probabilistic Databases, 2020. To appear.
[29] Martin Grohe and Peter Lindner. Infinite Probabilistic Databases. In *23rd International Conference on Database Theory (ICDT 2020)*, volume 155 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 16:1–16:20, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ICDT.2020.16.
[30] Bernd Gutmann, Manfred Jaeger, and Luc De Raedt. Extending ProbLog with Continuous Distributions. In *Inductive Logic Programming (ILP 2010)*, Lecture Notes in Computer Science, pages 76–91, Berlin, Heidelberg, Germany, 2011. Springer. doi:10.1007/978-3-642-21295-6_12.
[31] Tomasz Imieliński and Witold Lipski. Incomplete Information in Relational Databases. *Journal of the ACM*, 31(4):761–791, 1984. doi:10.1145/1634.1886.
[32] Ravi Jampani, Fei Xu, Mingxi Wu, Luis Perez, Chris Jermaine, and Peter J. Haas. The Monte Carlo Database System: Stochastic Analysis Close to the Data. *ACM Transactions on Database Systems*, 36(3), 2011. doi:10.1145/2000824.2000828.
[33] Abhay Jha and Dan Suciu. Probabilistic Databases with MarkoViews. *Proceedings of the VLDB Endowment*, 5(11):1160–1171, 2012. doi:10.14778/2350229.2350236.
[34] Jean Christoph Jung and Carsten Lutz. Ontology-Based Access to Probabilistic Data with OWL QL. In *The Semantic Web – ISWC 2012*, Lecture Notes in Computer Science, pages 182–197, Berlin, Heidelberg, Germany, 2012. Springer. doi:10.1007/978-3-642-35176-1_12.
[35] Oliver Kennedy and Christoph Koch. PIP: A Database System for Great and Small Expectations. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 157–168. IEEE, 2010. doi:10.1109/ICDE.2010.5447879.
[36] Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and Complexity. In *Advances in Probabilistic Databases for Uncertain Information Management*, volume 304 of *Studies in Fuzziness and Soft Computing*, pages 39–66. Springer, 2013. doi:10.1007/978-3-642-37509-5_3.
[37] Christoph Koch and Dan Olteanu. Conditioning Probabilistic Databases. *Proceedings of the VLDB Endowment*, 1(1):313–325, 2008. doi:10.14778/1453856.1453894.
[38] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models—Principles and Techniques*. MIT Press, 2009.
[39] Leonid Libkin. *Elements of Finite Model Theory*. Texts in Theoretical Computer Science (TTCS). Springer-Verlag Berlin Heidelberg, Berlin and Heidelberg, Germany, 2004. doi:https://doi.org/10.1007/978-3-662-07003-1.
[40] Leonid Libkin. Certain Answers Meet Zero-One Laws. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2018)*, pages 195–207, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3196959.3196983.
[41] Brian Milch, Bhaskara Marthi, Stuart Russell, David Sontag, Daniel L. Ong, and Andrey Kolobov. BLOG: Probabilistic Models with Unknown Objects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1352–1359, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.
[42] Dan Olteanu, Jiewen Huang, and Christoph Koch. SPROUT: Lazy vs. Eager Query Plans for Tuple-Independent Probabilistic Databases. In *Proceedings of the 25th International Conference on Data Engineering, (ICDE 2009)*, pages 640–651. IEEE Computer Society, 2009. doi:10.1109/ICDE.2009.123.
[43] David Poole. Probabilistic Horn Abduction and Bayesian Networks. *Artificial Intelligence*, 64(1):81–129, 1993. doi:https://doi.org/10.1016/0004-3702(93)90061-F.
[44] Christopher Ré. Applications of Probabilistic Constraints. Technical Report TR2007-03-03, University of Washington, Seattle, WA, USA, 2007.
[45] Christopher De Sa, Ihab F. Ilyas, Benny Kimelfeld, Christopher Ré, and Theodoros Rekatsinas. A Formal Framework for Probabilistic Unclean Databases. In *22nd International Conference on Database Theory (ICDT 2019)*, volume 127 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 6:1–6:18, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ICDT.2019.6.
[46] Sumit Sarkar and Debabrata Dey. Relational Models and Algebra for Uncertain Data. In *Managing and Mining Uncertain Data*, chapter 3. Springer, Boston, MA, USA, 2009. doi:10.1007/978-0-387-09690-2_3.
[47] Prithviraj Sen, Amol Deshpande, and Lise Getoor. PrDB: Managing and Exploiting Rich Correlations in Probabilistic Databases. *The VLDB Journal*, 18(5):1065–1090, 2009. doi:10.1007/s00778-009-0153-2.
[48] Sarvjeet Singh, Chris Mayfield, Sagar Mittal, Sunil Prabhakar, Susanne Hambrusch, and Rahul Shah. Orion 2.0: Native Support for Uncertain Data. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, pages 1239–1242, New York, NY, USA, 2008. Association for Computing Machinery. doi:10.1145/1376616.1376744.
[49] Sarvjeet Singh, Chris Mayfield, Rahul Shah, Sunil Prabhakar, Susanne E. Hambrusch, Jennifer Neville, and Reynold Cheng. Database Support for Probabilistic Attributes and Tuples. In *Proceedings of the 24th International Conference on Data Engineering (ICDE 2008)*, pages 1053–1061, 2008. doi:10.1109/ICDE.2008.4497514.
[50] Parag Singla and Pedro Domingos. Markov Logic in Infinite Domains. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI 2007)*, pages 368–375, Arlington, Virginia, USA, 2007. AUAI Press.
[51] Dan Suciu. Probabilistic Databases for All. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2020)*, pages 19–31, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3375395.3389129.
[52] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool, San Rafael, CA, USA, 1st edition, 2011. doi:10.2200/S00362ED1V01Y201105DTM016.
[53] Guy Van den Broeck and Dan Suciu. Query Processing on Probabilistic Data: A Survey. *Foundations and Trends® in Databases*, 7(3–4):197–341, 2017. doi:10.1561/1900000052.
[54] Ron van der Meyden. Logical Approaches to Incomplete Information: A Survey. In *Logics for Databases and Information Systems*, The Springer International Series in Engineering and Computer Science, pages 307–356. Springer US, Boston, MA, USA, 1998. doi:10.1007/978-1-4615-5643-5_10.
[55] Wenjie Zhang, Xuemin Lin, Jian Pei, and Ying Zhang. Managing Uncertain Data: Probabilistic Approaches. In *The Ninth International Conference on Web-Age Information Management, (WAIM 2008)*, pages 405–412. IEEE Computer Society, 2008. doi:10.1109/WAIM.2008.42.