

# Tuple-Independent Representations of Infinite Probabilistic Databases

Nofar Carmeli<sup>1</sup>, Martin Grohe<sup>2</sup>, Peter Lindner<sup>2</sup>, and Christoph Standke<sup>2</sup>

<sup>1</sup>snofca@cs.technion.ac.il

<sup>2</sup>{grohe,lindner,standke}@informatik.cs.rwth-aachen.de

<sup>1</sup>Technion—Israel Institute of Technology, Haifa, Israel

<sup>2</sup>RWTH Aachen University, Aachen, Germany

April 20, 2022

## 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. In addition, we show that conditioning on first-order properties yields no additional power in terms of expressivity, and we discuss the relation between purely logical and arithmetic reasons for (non-)representability. Finally, we inspect the relative expressivity of additional representations, where the views are restricted to be fragments of first-order logic (such as conjunctive queries with or without self-joins and unions of conjunctive queries).

## 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, modeled by a probability distribution over the strings over a suitable finite alphabet.

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 [56]: *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 [35]. While existing PDB systems are already using infinite domains and sample spaces, such as [23, 52, 4, 39, 36, 14], a formal framework of probabilistic databases with infinite sample spaces has only been introduced recently [31, 33]. 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 describing a fixed number of samples of the fermentation process when brewing Guinness beer, containing the date of each sample and the number of yeast cells in the sample. This number is known to be Poisson distributed [11].

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 [31], 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 **FO(TI)**. We prove that the class **FO(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. **FO(TI | FO) = FO(TI)**. This allows us to show that

all block-independent disjoint PDBs (**BID**-PDBs) are in **FO(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 yeast cell example above, where specific fields of facts in a table store the outcome of a random variable.

In [31], the authors exploited that **TI**-PDBs have finite expected instance size to construct a PDB that is not contained in **FO(TI)**. We generalize this idea and prove that for every PDB in **FO(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 **FO(TI)**. In addition, we give an example showing that there are even PDBs that have finite size moments but that are still not in **FO(TI)**. Complementing these non-representability results, we prove that all PDBs of bounded instance size are in **FO(TI)**. Furthermore, we give a sufficient criterion on the growth rate of the probabilities of PDBs for membership in **FO(TI)**.

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 **FO(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 **FO(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 **FO(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.

As we have seen that various classes of first-order views over **TI**- and **BID**-PDBs coincide with **FO(TI)**, one might wonder about the consequences of allowing only fragments of first-order logic in representation. We investigate such classes and show that, while most of these classes collapse in the finite setting, they remain separated in the infinite.

**Related Work** By now, there exists an abundance of theoretical work on finite PDBs (see the surveys [56, 57] as well as [18, 55]). The most prominent theoretical problem regarding finite representation systems is probabilistic query evaluation (PQE) complexity [19, 21, 25, 6], typically subject to independence assumptions. In particular, [5] considers PQE on structurally restricted **BID**-PDBs. Among query languages, conjunctive queries and unions of conjunctive queries received the most attention [21]. On the side of expressiveness, [10, 28, 51, 7] 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 [24, 23, 53]. Overviews over representation formalisms for probabilistic databases can be found in [28, 59, 50, 56, 57].

Conditioning PDBs has been considered before in order to introduce correlations or dependencies [48, 37], for updates of probabilistic data [42], for making queries tractable [47], in cleaning problems [30, 49] and for introducing ontologies [38, 12]. In [20], the authors consider limit probabilities of conjunctive queries over conditioned PDBs.

There already exist PDB systems supporting infinite probability spaces [23, 52, 4, 39, 36, 14].

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 [31, 33]. Allowing for distributions over worlds of unbounded size is strongly motivated by incorporating the open-world assumption into PDBs [16, 31, 12, 26]. Semi-structured models (probabilistic XML [40]) have been extended towards infinite spaces as well [1, 9].

As discussed by [28], incomplete databases [35, 2, 58] are closely related to PDBs. In [22], the authors study expressiveness among classes of incomplete databases with strong associations with **TI**- and **BID**-PDBs. Recent work on incomplete databases considers (probabilistic) measures of certainty for infinite domains [45, 17].

It has been noticed that there are strong connections between PDBs and probabilistic models in AI research [57], in particular probabilistic graphical models [43] and weighted model counting (WMC) [29]. There exist well-established modeling formalisms that support infinite spaces, for example [46, 54, 34]. Recently, WMC has also been introduced for infinite universes [8].

**Paper Outline** We review the background of probabilistic databases in Section 2. In Section 3, we explore the limits of tuple-independent representations by first-order views. 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 establish 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. 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. In Section 7, we consider various important fragments of first-order logic and investigate their relative expressive power for representations over PDBs with independence assumptions. We close the paper with concluding remarks in Section 8.

This paper is the journal version of the paper [15] that was presented at PODS 2021. It contains detailed proofs and examples that were excluded from the conference version (Theorems 3.2, 3.10, 5.1, 5.9, 5.10 and 6.6). Some of the arguments have been streamlined and simplified. In addition to the main body of results from the conference version, this paper contains a new section (Section 7) that focuses on representations using fragments of **FO** and their relative expressiveness. This answers open questions from the original paper.

## 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 and 1.

### 2.1 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,  $P: 2^{\mathbb{S}} \rightarrow [0, 1]$  with the property that  $P(A) = \sum_{s \in A} P(\{s\})$  for all  $A \subseteq \mathbb{S}$ , 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

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

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  $X: \mathbb{S} \rightarrow \mathbb{R}$ . Its *expectation* is  $E(X) := \sum_{s \in \mathbb{S}} X(s) \cdot P(\{s\})$  and, in general, its *kth moment* is  $E(X^k)$ . We write  $E_{\mathcal{S}}$  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  $f: \mathbb{S} \rightarrow \mathbb{T}$  a function, then  $f$  introduces a probability distribution on  $\mathbb{T}$  via  $P_{\mathcal{T}}(\{t\}) := P_{\mathcal{S}}(\{s \in \mathbb{S} : f(s) = t\})$ .

### 2.2 Relational Databases

We fix some non-empty set  $\mathbb{U}$  (called the *universe*). A *database schema*  $\tau$  is a finite, nonempty set of relation symbols with *arities*  $\text{ar}(R) \in \mathbb{N}$  for all  $R \in \tau$ . A  $(\tau)$ -*fact* is an expression of the shape  $R(u_1, \dots, u_{\text{ar}(R)})$  where  $R \in \tau$  and  $u_i \in \mathbb{U}$  for all  $1 \leq i \leq \text{ar}(R)$ . A  $\tau$ -*instance*  $D$  is a finite set of  $\tau$ -facts. The *active domain* of  $D$ , denoted by  $\text{adom}(D)$ , 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.

In general, a *view* may be any function that maps database instances of an input schema  $\tau$  to instances of an output schema  $\tau'$ . 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. In this paper, we focus on views and queries that are expressed in first-order logic.

### 2.3 First Order Logic, Relational Calculus, and Fragments

Let  $\tau$  be a database schema. A *relational atom* is an expression  $R(\mathbf{u})$  where  $\mathbf{u}$  is a tuple over variables or constants from  $\mathbb{U}$ . An *equality atom* is an expression  $x = u$  where  $x$  is a variable and  $u$  is a variable or a constant. A formula of *first-order logic* **FO** over  $\tau$  is built from atoms using the standard Boolean connectives and existential and universal quantification. When using first-order logic as a database query language, we assume the *active domain semantics* in this paper, see [2].<sup>1</sup> That is, when a formula  $\Phi$  is evaluated on a  $\tau$ -instance  $D$ , the quantifiers, and all valuations are taken to range over  $\text{adom}(D, \Phi) := \text{adom}(D) \cup \text{adom}(\Phi)$ , where  $\text{adom}(\Phi)$  is the set of constants appearing in  $\Phi$ .

We denote the free variables of an **FO**-formula  $\Phi$  by  $\text{free}(\Phi)$  and write  $\Phi = \Phi(\mathbf{x})$  where  $\mathbf{x}$  consists of pairwise distinct variables to indicate that  $\Phi$  has exactly the free variables  $\mathbf{x}$ . A formula  $\Phi$  is called a *sentence* or *Boolean* if  $\text{free}(\Phi) = \emptyset$ . Essentially, we employ first-order logic in the guise of *relational calculus*. That is, if  $D$  is a  $\tau$ -instance, and  $\Phi(\mathbf{x})$  is an **FO**-formula over  $\tau$ , the result of  $\Phi = \Phi(x_1, \dots, x_n)$  on  $D$  is given as

$$\Phi(D) := \{R_\Phi(\mathbf{a}) : \mathbf{a} = (a_1, \dots, a_n) \in \text{adom}(D, \Phi)^n \text{ such that } D \models \Phi[\mathbf{x}/\mathbf{a}]\}, \quad (1)$$

where  $\models$  refers to the active domain semantics, and  $\Phi[\mathbf{x}/\mathbf{a}]$  is the sentence that emerges from  $\Phi$  by replacing  $x_i$  with the constant  $a_i$  for all  $i = 1, \dots, n$ . Moreover,  $R_\Phi$  is a designated relation symbol of the output database schema. Thus, we treat  $\Phi(D)$  as a database instance. Of course, (1) depends on the order imposed on the free variables. We make this order explicit by writing  $\Phi = \Phi(\mathbf{x})$ .

In general, a tuple  $\mathbf{x}$  may contain repeated variables or constants, like in  $\Phi(x, x, y, a, y)$ . This is equivalent to the first-order formula  $\Phi'(x, x', y, x'', y')$  that emerges from  $\Phi$  by conjoining the formula with  $x' = x$ ,  $x'' = a$  and  $y'$  equals  $y$ , where  $x', x'', y'$  are new variables.

A first-order formula is a *conjunctive query* (**CQ**) if it is built from atoms by only using  $\exists$  and  $\wedge$ . Every **CQ** is equivalent to a formula in the shape  $\exists \mathbf{y} : \bigwedge_{i=1}^m \Phi_i$  where the  $\Phi_i$  are atomic formulae (hence, the name **CQ**). A **CQ** is called *self-join free* (**sjfCQ**), if every relation symbol occurs at most once in the formula. Note that we can rewrite every satisfiable conjunctive query  $\Phi(\mathbf{x})$  into a conjunctive query  $\Phi'(\mathbf{x}')$  without equality atoms using substitutions. This might then introduce constants or repeated variables in  $\mathbf{x}'$  (see [2, Chapter 4]).

A first-order formula is a *union of conjunctive queries* (**UCQ**) if it is built from atoms by only using  $\exists$  and  $\wedge$  and  $\vee$ . Every **UCQ** is equivalent to a formula in the shape  $\bigvee_{i=1}^n \Phi_i$  where the  $\Phi_i$  are **CQ**s.

For more background on first-order logic, relational calculus and the fragments introduced above, see [2, 44].

### 2.4 Probabilistic Databases

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

---

<sup>1</sup>We could equivalently restrict ourselves to *domain-independent* formulae and use the standard semantics of first-order logic [2, pp. 77ff]. As we will construct a plethora of formulae in this paper, using the active domain semantics is more convenient though.

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

Note that while  $|\mathbb{D}|$  may be (countably) infinite,  $\mathbb{D}$  still only contains database instances, that is, *finite* sets of facts. We call a PDB *finite* if  $|\mathbb{D}|$  is finite. The class of finite PDBs is denoted by **PDB<sub>fin</sub>**.

**Remark 2.2.** As seen in Theorem 2.1, all probabilistic databases that are considered 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 [31, 33].

In a (discrete) probabilistic database  $\mathcal{D}$ , the instances of positive probability are often called its *possible worlds*. We denote the set of possible worlds of the PDB  $\mathcal{D}$  by  $\text{worlds}(\mathcal{D})$ .<sup>2</sup> We let  $\text{facts}(\mathcal{D}) := \bigcup_{D \in \text{worlds}(\mathcal{D})} D$  denote the set of all facts appearing among the instances of  $\mathcal{D}$ . For all  $f \in \text{facts}(\mathcal{D})$ , the *marginal probability* of  $f$  is  $\Pr_{D \sim \mathcal{D}}(f \in D)$ , and is usually denoted by  $p_f$ .

#### 2.4.1 Instance Size

In every probabilistic database  $\mathcal{D} = (\mathbb{D}, P)$ , the *instance size*  $|\cdot|$  is a random variable  $|\cdot|: \mathbb{D} \rightarrow \mathbb{N}$  that maps every database instance  $D$  to the number  $|D|$  of facts it contains. Its expectation is given by

$$\mathbb{E}(|\cdot|) = \sum_{D \in \mathbb{D}} |D| \cdot P(\{D\}).$$

**Definition 2.3** (Finite Moments Property). A PDB  $\mathcal{D} = (\mathbb{D}, P)$  has the *finite moments property* if for all  $k \in \mathbb{N}_+$ , the  $k$ th moment of the instance size is finite, that is,

$$\mathbb{E}(|\cdot|^k) = \sum_{D \in \mathbb{D}} |D|^k \cdot P(\{D\}) < \infty.$$

A class  $\mathbf{D}$  of PDBs has the *finite moments property*, if every  $\mathcal{D} \in \mathbf{D}$  does.

#### 2.4.2 Probabilistic Query Semantics

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  $V: \mathbb{D} \rightarrow \mathbb{D}'$  a view, then  $V(\mathcal{D}) = (\mathbb{D}', P')$  is a PDB where

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

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

If  $\mathbf{V}$  is a class of views and  $\mathbf{D}$  is a class of PDBs, then  $\mathbf{V}(\mathbf{D})$  denotes the class of images of PDBs of  $\mathbf{D}$  under views of  $\mathbf{V}$ . We call  $\mathbf{D}$  *closed under*  $\mathbf{V}$  if  $\mathbf{V}(\mathbf{D}) = \mathbf{D}$ .

An **FO**-view is a view that consists of an **FO**-formula for each relation symbol in the target schema. Then, **FO**( $\mathbf{D}$ ) denotes the class of PDBs that are the image of a  $\mathbf{D}$ -PDB under an **FO**-view. These notions are defined analogously for **sjfCQ**, **CQ** and **UCQ**.

<sup>2</sup>In most cases, we can just assume that  $\mathbb{D}$  contains no instances of probability zero. However, the notation we introduce here proves convenient later in the paper.



### 2.4.3 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.

For ease of reading, we denote PDBs that obey independence assumptions by  $\mathcal{I} = (\mathbb{I}, P)$  rather than  $\mathcal{D} = (\mathbb{D}, P)$ . Instances are then denoted by  $I$  accordingly.

**Definition 2.4** (Tuple-Independent PDBs). A PDB  $\mathcal{I}$  is called *tuple-independent* if for all  $k \in \mathbb{N}$  and all pairwise distinct facts  $f_1, \dots, f_k \in \text{facts}(\mathcal{I})$  it holds that

$$\Pr_{I \sim \mathcal{I}} (f_1 \in I, \dots, f_k \in I) = \prod_{i=1}^k \Pr_{I \sim \mathcal{I}} (f_i \in I).$$

We let **TI** and **TI<sub>fin</sub>** denote the class of tuple-independent PDBs and finite tuple-independent PDBs, respectively.<sup>3</sup>

The following provides a necessary and sufficient criterion for the existence of **TI**-PDBs in terms of its marginal probabilities.

**Theorem 2.5** ([31, Theorem 4.8]). *Let  $F$  be a set of facts over a schema  $\tau$  and let  $(p_f)_{f \in F}$  with  $p_f \in [0, 1]$  for all  $f \in F$ . The following are equivalent:*

1. *There exists  $\mathcal{I} \in \mathbf{TI}$  with  $\text{facts}(\mathcal{I}) = F$  and marginal probabilities  $\Pr_{I \sim \mathcal{I}} (f \in I) = p_f$  for all  $f \in F$ .*
2. *It holds that  $\sum_{f \in F} p_f < \infty$ .*

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

**Definition 2.6** (Block-Independent Disjoint PDBs). A PDB  $\mathcal{I}$  is called *block-independent disjoint* if there exists a partition  $\mathcal{B}$  of  $\text{facts}(\mathcal{I})$  into *blocks* such that:

1. for all  $k \in \mathbb{N}$  and all  $f_1, \dots, f_k$  from pairwise different blocks,

$$\Pr_{I \sim \mathcal{I}} (f_1 \in I, \dots, f_k \in I) = \prod_{i=1}^k \Pr_{I \sim \mathcal{I}} (f_i \in I).$$

2. for all  $B \in \mathcal{B}$  and all  $f, f' \in B$  with  $f \neq f'$  it holds that  $\Pr_{I \sim \mathcal{I}} (f \in I \text{ and } f' \in I) = 0$ .

We let **BID** and **BID<sub>fin</sub>** denote the class of block-independent disjoint and finite block-independent disjoint PDBs, respectively.

---

<sup>3</sup>Usually, *tuple-independent* PDBs are defined by the property that the events “ $f \in I$ ” are stochastically independent. This definition only makes sense for countable PDBs, as in every PDB there are at most countably many facts with positive marginal probability (see [31, Proposition 3.4]). There is an extension of tuple-independent PDBs to uncountable domains [32].



A theorem similar to Theorem 2.5 exists for **BID**-PDBs:

**Theorem 2.7** ([31, Theorem 4.15]). *Let  $F$  be a set of facts over a schema  $\tau$ , let  $(p_f)_{f \in F}$  with  $p_f \in [0, 1]$ , and let  $\mathcal{B}$  be a partition of  $F$  such that  $\sum_{f \in B} p_f \leq 1$  for all  $B \in \mathcal{B}$ . The following are equivalent:*

1. *There exists  $\mathcal{I} \in \mathbf{BID}$  with  $\text{facts}(\mathcal{I}) = F$ , blocks  $\mathcal{B}$  and marginal probabilities  $\Pr_{\mathcal{I} \sim \mathcal{I}}(f \in \mathcal{I}) = p_f$  for all  $f \in F$ .*
2. *It holds that  $\sum_{f \in F} p_f = \sum_{B \in \mathcal{B}} \sum_{f \in B} p_f < \infty$ .*

We note that both **TI**-, and **BID**-PDBs are, in case of existence, uniquely determined by their marginals (and partition into blocks) [32, Section 4].

### 3 Limitations of Tuple-Independent Representations

Recall that every finite probabilistic database can be represented as an **FO**-view over a finite **TI**-PDB [56, Proposition 2.16]. A similar statement does *not* hold for infinite PDBs.

**Proposition 3.1** ([31, Proposition 4.9]).  $\mathbf{FO}(\mathbf{TI}) \subsetneq \mathbf{PDB}$ .

In this section, we investigate classes of PDBs for which we can prove that they do not belong to **FO(TI)**. The argument of [31] that establishes the proposition above, shows that in any **FO(TI)**-PDB the expected instance size is necessarily finite (while there exist well-defined PDBs of infinite expected instance size). In Section 3.1, we will first generalize their argument to show that for membership in **FO(TI)** it is even necessary that *all* moments of the instance size are finite, that is, that the PDB satisfies the finite moments property. In Section 3.2, however, we show that this requirement is not *sufficient* for membership in **FO(TI)** by giving constructions of non-**FO(TI)**-PDBs that nevertheless exhibit the finite moments property.

#### 3.1 Infinite Moments

Before considering the class **FO(TI)** directly, we show that all **TI**-PDBs have the finite moments property. This relies upon the following observation about independent  $\{0, 1\}$ -valued random variables.

**Lemma 3.2.** *Suppose  $(X_i)_{i \in \mathbb{N}}$  is a family of  $\{0, 1\}$ -valued, independent random variables and let  $X := \sum_{i=1}^{\infty} X_i$  with  $E(X) < \infty$ . Then for all  $k \in \mathbb{N}_+$ ,  $k \geq 2$  it holds that*

$$E(X^k) \leq E(X^{k-1}) \cdot (k - 1 + E(X)). \quad (2)$$

*Proof.* Let  $k \in \mathbb{N}_+$  be arbitrary with  $k \geq 2$ . It holds that<sup>4</sup>

$$E(X^k) = E\left(\left(\sum_{i=1}^{\infty} X_i\right)^k\right) = E\left(\sum_{i_1, \dots, i_k} X_{i_1} \cdot \dots \cdot X_{i_k}\right) = \sum_{i_1, \dots, i_k} E(X_{i_1} \cdot \dots \cdot X_{i_k})$$

<sup>4</sup>The calculation is not as trivial as it may seem, since we sum up infinitely many random variables. However, the equalities can easily be verified using [41, Theorem 5.3].

(with the indices of sums on the right ranging over non-negative integers). Now consider the product  $X_{i_1} \cdot \dots \cdot X_{i_k}$  for arbitrary  $i_1, \dots, i_k \in \mathbb{N}$ . Using that the  $X_i$  are  $\{0, 1\}$ -valued and independent, it holds that

$$\mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_k}) = \begin{cases} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) & \text{if } i_k \in \{i_1, \dots, i_{k-1}\} \text{ and} \\ \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) \cdot \mathbb{E}(X_{i_k}) & \text{otherwise.} \end{cases}$$

Using this to continue our calculation from before, we get

$$\begin{aligned} \mathbb{E}(X^k) &= \sum_{i_1, \dots, i_k} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_k}) \\ &= \sum_{i_1, \dots, i_{k-1}} \left( \sum_{i_k} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_k}) \right) \\ &= \sum_{i_1, \dots, i_{k-1}} \left( \sum_{i_k \in \{i_1, \dots, i_{k-1}\}} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) + \sum_{i_k \notin \{i_1, \dots, i_{k-1}\}} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) \cdot \mathbb{E}(X_{i_k}) \right) \\ &= \sum_{i_1, \dots, i_{k-1}} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) \cdot \left( \sum_{i_k \in \{i_1, \dots, i_{k-1}\}} 1 + \sum_{i_k \notin \{i_1, \dots, i_{k-1}\}} \mathbb{E}(X_{i_k}) \right) \\ &\leq \left( \sum_{i_1, \dots, i_{k-1}} \mathbb{E}(X_{i_1} \cdot \dots \cdot X_{i_{k-1}}) \right) \cdot \left( k - 1 + \sum_{i_k} \mathbb{E}(X_{i_k}) \right) \\ &= \mathbb{E}(X^{k-1}) \cdot (k - 1 + \mathbb{E}(X)). \end{aligned} \quad \square$$

In our application, the random variables  $X_i$  from the previous lemma will correspond to the indicator random variables of the facts of a **TI**-PDB.

**Proposition 3.3.** *TI has the finite moments property.*

*Proof.* Suppose facts( $\mathcal{I}$ ) =  $\{f_0, f_1, \dots\}$  and let  $X_i: \mathbb{I} \rightarrow \{0, 1\}$  be the indicator random variable of the event  $\{f_i \in I\}$  for  $I \sim \mathcal{I}$ . Since the size of any instance  $I \in \mathbb{I}$  is exactly the number of facts in  $I$ , we can express the instance size random variable  $|\cdot|$  using the indicator random variables  $X_i$  as follows:

$$|\cdot| = \sum_{i=1}^{\infty} X_i.$$

For simplicity, we write  $\mathbb{E}$  for  $\mathbb{E}_{\mathcal{I}}$ . Since  $X_i$  is  $\{0, 1\}$ -valued, we know that

$$\mathbb{E}(X_i) = \Pr_{I \sim \mathcal{I}}(X_i(I) = 1) = \Pr_{I \sim \mathcal{I}}(f_i \in I) = p_i$$

for all  $i \in \mathbb{N}$ , where  $p_i$  is the marginal probability of  $f_i$  in  $\mathcal{I}$ . Thus,

$$\mathbb{E}(|\cdot|) = \mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i) = \sum_{i=1}^{\infty} p_i.$$

By Theorem 2.5 the last sum is finite, so  $\mathbb{E}(|\cdot|) < \infty$ . In order to obtain this for higher moments  $|\cdot|^k$  with  $k > 1$ , we proceed inductively (going from  $k - 1$  to  $k$ ) using Theorem 3.2. This yields

$$\mathbb{E}(|\cdot|^k) \leq \mathbb{E}(|\cdot|^{k-1}) \cdot (k - 1 + \mathbb{E}(|\cdot|)).$$

The right-hand side of the above is finite by the induction hypothesis, i. e.  $E(|\cdot|^k) < \infty$ .  $\square$

**Remark 3.4.** The same argument can be employed to show that **BID** has the finite moments property. For this,  $X_i$  needs to be taken as the indicator random variable of the  $i$ th block of the **BID**-PDB. Then one can proceed analogously to the proof of Theorem 3.3.

We next show that **FO**-views over classes with the finite moments property preserve the finite moments property. This is because **FO**-views can only cause a polynomial blowup in the instance size per individual possible world.

**Lemma 3.5.** *If  $\mathcal{D}$  is a PDB with the finite moments property and  $\Phi$  is an **FO**-view, then  $\Phi(\mathcal{D})$  also has the finite moments property.*

*Proof.* Let  $\mathcal{D}$  be a PDB with the finite moments property and let  $\Phi$  be an **FO**-view over the schema of  $\mathcal{D}$ . Suppose that  $\Phi$  consists of  $m$  first-order formulae  $\Phi_1, \dots, \Phi_m$  with arities  $r_1, \dots, r_m$ , respectively. Observe that due to the active domain semantics, for all  $D \sim \mathcal{D}$  we have

$$|\Phi(D)| = \sum_{i=1}^m |\Phi_i(D)| \leq \sum_{i=1}^m |\text{adom}(D, \Phi_i)|^{r_i} \leq m \cdot |\text{adom}(D, \Phi)|^r \leq m \cdot (r_{\max} \cdot |D| + |\text{adom}(\Phi)|)^r$$

where  $r = \max_{i=1, \dots, m} r_i$  and  $r_{\max}$  is the maximum arity of a relation symbol in the schema of  $\mathcal{D}$ . Thus, for all  $k \in \mathbb{N}_+$  it follows that

$$E_{\Phi(\mathcal{D})}(|\cdot|^k) \leq E_{\mathcal{D}}\left(m^k \cdot (|\cdot| r_{\max} + |\text{adom}(\Phi)|)^{rk}\right) = m^k \cdot \sum_{j=0}^{rk} \binom{rk}{j} \cdot r_{\max}^j \cdot |\text{adom}(\Phi)|^{rk-j} \cdot E_{\mathcal{D}}(|\cdot|^j),$$

using the binomial formula and linearity of expectation. Since  $\mathcal{D}$  has the finite moments property, all the  $E_{\mathcal{D}}(|\cdot|^j)$  with  $j = 0, 1, \dots, rk$  are finite. Thus,  $E_{\Phi(\mathcal{D})}(|\cdot|^k)$  is finite as well.  $\square$

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

**Example 3.6.** Consider the schema that only contains a single unary relation symbol  $R$  and let  $D_i = \{R(1), \dots, R(i)\}$  for all  $i \in \mathbb{N}_+$ . Let  $P(\{D_i\}) = \frac{Z}{i^3}$  where  $Z = \sum_{i=1}^{\infty} \frac{1}{i^3}$  is the normalizing constant. Then  $\mathcal{D} = (\mathbb{D}, P)$  is a PDB with expected instance size  $\sum_{i=1}^{\infty} i \cdot \frac{Z}{i^3} = Z \cdot \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$ . Note, however, that the second moment of the instance size of  $\mathcal{D}$  is infinite.

Now consider the view  $\Phi(x, y) = R(x) \vee R(y)$ . Then  $|\Phi(D_i)| = |D_i|^2$ . Thus,  $\Phi(\mathcal{D})$  has expected instance size  $\sum_{i=1}^{\infty} i^2 \cdot \frac{Z}{i^3} = Z \cdot \sum_{i=1}^{\infty} \frac{1}{i} = \infty$ .

With Theorem 3.5 the finite moments property of **TI** (Theorem 3.3) directly extends to **FO(TI)**.

**Proposition 3.7.** **FO(TI)** has the finite moments property.

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

**Example 3.8.** Let  $\mathcal{D} = (\mathbb{D}, P)$  be a PDB with  $\mathbb{D} = \{D_1, D_2, \dots\}$ , where  $|D_i| = 2^i$  and  $P(\{D_i\}) = \frac{3}{4^i}$ . Then,

$$E_{\mathcal{D}}(|\cdot|) = \sum_{D \in \mathbb{D}} |D| \cdot P(\{D\}) = 3 \cdot \sum_{i=1}^{\infty} 2^i \cdot \frac{1}{4^i} = 3 \cdot \sum_{i=1}^{\infty} \frac{1}{2^i} = 3,$$

but

$$E_{\mathcal{D}}(|\cdot|^2) = \sum_{D \in \mathbb{D}} |D|^2 \cdot P(\{D\}) = 3 \cdot \sum_{i=1}^{\infty} 2^{2i} \cdot \frac{1}{4^i} = 3 \cdot \sum_{i=1}^{\infty} 1 = \infty.$$

That is,  $\mathcal{D}$  is a PDB of finite expected instance size where the second moment of the instance size is infinite. According to Theorem 3.7,  $\mathcal{D} \notin \mathbf{FO}(\mathbf{TI})$ .

### 3.2 Balancing the Marginal Probabilities

In this subsection, we show that the finite moments property is not a sufficient condition for membership in  $\mathbf{FO}(\mathbf{TI})$ . That is, the main result of this subsection is the following.

**Theorem 3.9.** *There are PDBs having the finite moments property that are not in  $\mathbf{FO}(\mathbf{TI})$ .*

The proof of this theorem is quite involved. However, it may be skipped by the reader, as the techniques are not essential for the understanding of the rest of the paper.

To obtain Theorem 3.9, we essentially aim to break a balance in fact probabilities that is required for representations: On the one hand, in  $\mathbf{TI}$ -PDBs, the sum of fact probabilities must converge, meaning that these probabilities have to approach zero fast enough. On the other hand, fact probabilities in the  $\mathbf{TI}$ -PDB need to be sufficiently high in order to represent instances of a certain probability with respect to an  $\mathbf{FO}$ -view on  $\Phi$ . This is due to the fact that images of  $\mathbf{FO}$ -views only use domain elements that either appeared in the input instance, or as constants in the view itself. While the first part of this (the convergence of fact probabilities in the  $\mathbf{TI}$ -PDB) is clear, this section focuses first on formalizing the second part of our statement. In particular, we start by turning this idea into an upper bound for the probabilities of instances in the image of an  $\mathbf{FO}$ -view on a  $\mathbf{TI}$ -PDB.

**Lemma 3.10.** *Let  $\mathcal{I}$  be a  $\mathbf{TI}$ -PDB and let  $\Phi$  be an  $\mathbf{FO}$ -view over the schema of  $\mathcal{I}$ . Then for all instances  $D$  of  $\Phi(\mathcal{I})$  it holds that*

$$\Pr_{\mathcal{I} \sim \mathcal{I}}(\Phi(\mathcal{I}) = D) \leq |A_D^*| \cdot \left( r^2 \cdot |A_D^*|^{r-1} \cdot \sum_{f \in F_D^*} p_f \right)^{|A_D^*|/r},$$

where  $A_D^*$  is the set of domain elements appearing in  $D$  that are not constants in  $\Phi$ ,  $F_D^*$  is the set of facts in  $D$  that contain at least one element from  $A_D^*$  and  $r$  is the maximum arity among the relations of  $\mathcal{I}$ .

The idea behind this auxiliary result connects to the previous discussion as follows: If  $D$  is an instance in the image of  $\Phi$  on  $\mathcal{I}$ , then the active domain elements of  $D$  must appear either in the active domain of an input instance from  $\mathcal{I}$  or as a constant in  $\Phi$ . Summing over all the possible ways to “cover” the active domain of  $D$  (minus the constants of  $\Phi$ ) with facts from the  $\mathcal{I}$  can thus be turned into an upper bound for the probability of  $D$  being the output of the view.

*Proof.* Let  $\mathcal{I} = (\mathbb{I}, P)$  be a **TI**-PDB and let  $\Phi$  be an **FO**-view over the schema of  $\mathcal{I}$ . We let  $\mathbb{A} := \bigcup_{I \in \mathbb{I}} \text{adom}(I)$  be the set of domain elements appearing among the instances of  $\mathcal{I}$  and  $\mathbb{F} := \text{facts}(\mathcal{I})$ . We say that a set  $F \subseteq \mathbb{F}$  covers a finite set  $A \subseteq \mathbb{A}$  if for all  $a \in A$  there exists a fact  $f \in F$  such that  $a$  appears in  $f$ . In that case, we call  $F$  a *fact cover* of  $A$ . A fact cover  $F \subseteq \mathbb{F}$  of  $A$  is called *minimal* if no proper subset of  $F$  is a fact cover of  $A$ . Note that every fact cover contains a minimal fact cover. The set of all fact covers of  $A$  is denoted by  $\text{FC}(A)$  and the set of all minimal fact covers of  $A$  is denoted by  $\text{FC}_{\min}(A)$ .

We start by bounding the probability of representing an instance in terms of minimal fact covers.

**Claim 3.11.** *For all instances  $D$  of  $\Phi(\mathcal{I})$  it holds that*

$$\Pr_{I \sim \mathcal{I}} (\Phi(I) = D) \leq \sum_{F \in \text{FC}_{\min}(A_D^*)} \prod_{f \in F} p_f, \quad (3)$$

where  $A_D^* = \text{adom}(D) \setminus \text{adom}(\Phi)$ .

*Proof.* Let  $D$  be an arbitrary instance of  $\Phi(\mathcal{I})$ . For all instances  $I \in \mathbb{I}$  with  $\Phi(I) = D$  we have  $\text{adom}(D) \subseteq \text{adom}(I) \cup \text{adom}(\Phi)$ , that is,  $I$  is a fact cover of  $A_D^*$ . Therefore,

$$\Pr_{I \sim \mathcal{I}} (\Phi(I) = D) \leq \Pr_{I \sim \mathcal{I}} (I \in \text{FC}(A_D^*)) = \sum_{F \in \text{FC}(A_D^*)} \Pr_{I \sim \mathcal{I}} (I = F) \leq \sum_{F' \in \text{FC}_{\min}(A_D^*)} \sum_{F \subseteq \mathbb{F} \setminus F'} \Pr_{I \sim \mathcal{I}} (I = F' \cup F'').$$

The last step above is not an equality, because any fact cover can contain multiple minimal fact covers, so we might be double counting on the right-hand side.

Now, for any set  $F$  of facts, we let  $\mathcal{I}[F]$  denote the restriction of the **TI**-PDB  $\mathcal{I}$  to the fact set  $F$ . This is again a **TI**-PDB that retains the marginal probabilities of the facts of  $F$  (and has marginal probability 0 for all other facts). By splitting instances into a minimal fact cover and the remaining facts, we then get

$$\begin{aligned} \sum_{F' \in \text{FC}_{\min}(A_D^*)} \sum_{F'' \subseteq \mathbb{F} \setminus F'} \Pr_{I \sim \mathcal{I}} (I = F' \cup F'') &= \sum_{F' \in \text{FC}_{\min}(A_D^*)} \sum_{F'' \subseteq \mathbb{F} \setminus F'} \Pr_{I \sim \mathcal{I}[F']} (I = F') \cdot \Pr_{I \sim \mathcal{I}[\mathbb{F} \setminus F']} (I = F'') \\ &= \sum_{F' \in \text{FC}_{\min}(A_D^*)} \Pr_{I \sim \mathcal{I}[F']} (I = F') \cdot \underbrace{\sum_{F'' \subseteq \mathbb{F} \setminus F'} \Pr_{I \sim \mathcal{I}[\mathbb{F} \setminus F']} (I = F'')}_{=1} \\ &= \sum_{F' \in \text{FC}_{\min}(A_D^*)} \Pr_{I \sim \mathcal{I}[F']} (I = F') \\ &= \sum_{F' \in \text{FC}_{\min}(A_D^*)} \prod_{f \in F'} p_f. \quad \lrcorner \end{aligned}$$

From now on, let  $D$  be an arbitrary but fixed instance of  $\Phi(\mathcal{I})$ . We let  $F_D^*$  denote the set of facts of  $\mathcal{I}$  that contain at least one element from  $A_D^* = \text{adom}(D) \setminus \text{adom}(\Phi)$  and define a function  $s_D^*: F_D^* \rightarrow 2^{A_D^*}$  as follows:

$$s_D^*(f) := \{a \in A_D^* : a \text{ occurs in } f\}.$$

Essentially,  $s_D^*$  turns  $f$  into the set of the occurring domain elements (by forgetting the order and multiplicities) and restricts this set to  $A_D^*$ . The function  $s_D^*$  is lifted to sets of facts  $F$  by letting  $s_D^*(F) = \{s_D^*(f) : f \in F\} \subseteq 2^{A_D^*}$ .

Let  $H$  be the hypergraph with vertex set  $A_D^*$  and (hyper)edge set

$$s_D^*(F_D^*) = \{e \subseteq A_D^* : e = s_D^*(f) \text{ for some } f \in F_D^*\}.$$

A set of hyperedges  $C$  is an *edge cover* of  $H$  if every vertex appears in at least one of the hyperedges from  $C$ . Such an edge cover  $C$  is called *minimal* if no proper subset of  $C$  covers all vertices. We let  $\text{EC}_{\min}(H)$  denote the set of minimal edge covers of  $H$ . The connection between minimal fact covers of  $A_D^*$  and minimal edge covers of  $H$  is as follows.

**Claim 3.12.** *A set of facts  $F \subseteq F_D^*$  is a minimal fact cover of  $A_D^*$ , if and only if the set  $s_D^*(F)$  is a minimal edge cover of  $H$  with  $|F| = |s_D^*(F)|$ .*

*Proof.* First, let  $F$  be a minimal fact cover of  $A_D^*$ . Since  $F$  is a fact cover of  $A_D^*$ , the hyperedges  $s_D^*(F)$  are an edge cover of  $H$ . Note that  $s_D^*$  is injective on  $F$ : if there were  $f \neq f'$  with  $s_D^*(f) = s_D^*(f')$ , then  $F \setminus \{f'\}$  would be a fact cover of  $A_D^*$ , contradicting the minimality of  $F$ . Thus,  $|s_D^*(F)| = |F|$ . We still have to show that  $s_D^*(F)$  is a *minimal* edge cover. If  $s_D^*(F)$  is not minimal, then there exists  $e \in s_D^*(F)$  such that  $s_D^*(F) \setminus \{e\}$  is still an edge cover of  $H$ . Then there exists  $f \in F$  with  $s_D^*(f) = e$  such that the domain elements occurring in  $f$  are already covered by other facts from  $F$ . But this means that  $F \setminus \{f\}$  is a fact cover of  $A_D^*$  in contradiction to  $F$  being minimal. Thus,  $s_D^*(F)$  is a minimal edge cover of  $H$ .

For the other direction, suppose that  $F \subseteq F_D^*$  is a set of facts such that  $s_D^*(F)$  is a minimal edge cover of  $H$  with  $|F| = |s_D^*(F)|$ . First note that  $F$  is a fact cover of  $A_D^*$ . Suppose  $F$  is not minimal. Then there exists  $F' \subsetneq F$  such that  $F'$  is a minimal fact cover of  $A_D^*$ . But then  $s_D^*(F') \subseteq s_D^*(F)$  is a minimal edge cover of  $H$  by the first part of this proof. This contradicts  $|s_D^*(F')| = |F'| < |F| = |s_D^*(F)|$ .  $\square$

From the above, it follows that

$$\text{FC}_{\min}(A_D^*) = \bigcup_{C \in \text{EC}_{\min}(H)} \{F \subseteq F_D^* : s_D^*(F) = C \text{ and } |C| = |F|\}.$$

We emphasize that the union on the right-hand side is disjoint, since  $s_D^*$  is a function. Next, we transform the bound from (3) into a bound in terms of minimal edge covers.

**Claim 3.13.** *The probability of representing  $D$  is bounded as follows:*

$$\Pr_{I \sim I} (\Phi(I) = D) \leq \sum_{C \in \text{EC}_{\min}(H)} \left( \frac{1}{|C|} \sum_{f \in F_D^*} p_f \right)^{|C|}.$$

*Proof.* We continue our calculation of an upper bound for  $\Pr_{I \sim I} (\Phi(I) = D)$  from (3) as follows:

$$\Pr_{I \sim I} (\Phi(I) = D) \stackrel{(3)}{\leq} \sum_{F \in \text{FC}_{\min}(A_D^*)} \prod_{f \in F} p_f = \sum_{C \in \text{EC}_{\min}(H)} \sum_{\substack{F \in \text{FC}_{\min}(A_D^*) \\ s_D^*(F) = C}} \prod_{f \in F} p_f = \sum_{C \in \text{EC}_{\min}(H)} \sum_{\substack{F \subseteq F_D^* \\ s_D^*(F) = C \\ |F| = |C|}} \prod_{f \in F} p_f.$$

Suppose now that  $C = \{e_1, \dots, e_k\}$  is a minimal edge cover of size  $k$ . Then

$$\sum_{\substack{F \subseteq F_D^* \\ s_D^*(F)=C \\ |F|=|C|}} \prod_{f \in F} p_f = \sum_{\substack{f_1 \in F_D^* \\ s_D^*(f_1)=e_1}} \cdots \sum_{\substack{f_k \in F_D^* \\ s_D^*(f_k)=e_k}} p_{f_1} \cdots p_{f_k} = \left( \sum_{\substack{f_1 \in F_D^* \\ s_D^*(f_1)=e_1}} p_{f_1} \right) \cdots \left( \sum_{\substack{f_k \in F_D^* \\ s_D^*(f_k)=e_k}} p_{f_k} \right) = \prod_{e \in C} \left( \sum_{\substack{f \in F_D^* \\ s_D^*(f)=e}} p_f \right).$$

Applying the inequality of arithmetic and geometric means, we get that

$$\prod_{e \in C} \left( \sum_{\substack{f \in F_D^* \\ s_D^*(f)=e}} p_f \right) \leq \left( \frac{1}{|C|} \sum_{e \in C} \sum_{\substack{f \in F_D^* \\ s_D^*(f)=e}} p_f \right)^{|C|} \leq \sum_{C \in \text{EC}_{\min}(H)} \left( \frac{1}{|C|} \sum_{f \in F_D^*} p_f \right)^{|C|}. \quad \lrcorner$$

Finally, we are ready to complete the proof of Theorem 3.10. Let  $r$  denote the maximum arity of any relation in the schema of  $\mathcal{I}$ . Then each hyperedge in  $H$  contains at most  $r$  elements. Thus, the size of any minimal edge cover of  $A_D^*$  is at least  $\lceil |A_D^*|/r \rceil \geq |A_D^*|/r$  and (trivially) at most  $|A_D^*|$ .

The number of hyperedges in  $H$  is at most  $\sum_{i=1}^r \binom{|A_D^*|}{i} \leq \sum_{i=1}^r |A_D^*|^i \leq r \cdot |A_D^*|^r$ . In particular, there are at most  $\binom{r \cdot |A_D^*|^r}{k} \leq (r \cdot |A_D^*|^r)^k$  minimal edge covers of size  $k$  in  $H$ . Therefore,

$$\begin{aligned} \Pr_{I \sim \mathcal{I}} (\Phi(I) = D) &\leq \sum_{C \in \text{EC}_{\min}(A_D^*)} \left( \frac{1}{|C|} \sum_{f \in F_D^*} p_f \right)^{|C|} = \sum_{k=\lceil |A_D^*|/r \rceil}^{|A_D^*|} \sum_{\substack{F \in \text{FC}_{\min}(A_D^*) \\ |F|=k}} \left( \frac{1}{k} \sum_{f \in F_D^*} p_f \right)^k \\ &\leq \sum_{k=\lceil |A_D^*|/r \rceil}^{|A_D^*|} (r \cdot |A_D^*|^r)^k \cdot \left( \frac{1}{k} \sum_{f \in F_D^*} p_f \right)^k \\ &\leq \sum_{k=\lceil |A_D^*|/r \rceil}^{|A_D^*|} \left( r^2 \cdot |A_D^*|^{r-1} \sum_{f \in F_D^*} p_f \right)^k \end{aligned}$$

The inequality remains intact if we decrease the value of the exponent on the right-hand side to a smaller non-negative number: if the value inside the parentheses is larger than 1, the complete right-hand side will still be larger than 1 as well. If the expression in the parentheses is at most one, then making the exponent smaller can only increase the value. Thus, we can infer that

$$\Pr (\Phi(I) = D) \leq \sum_{k=\lceil |A_D^*|/r \rceil}^{|A_D^*|} \left( r^2 \cdot |A_D^*|^{r-1} \sum_{f \in F_D^*} p_f \right)^{|A_D^*|/r} \leq |A_D^*| \cdot \left( r^2 \cdot |A_D^*|^{r-1} \sum_{f \in F_D^*} p_f \right)^{|A_D^*|/r}.$$

for all instances  $D$  of  $\Phi(\mathcal{I})$ , concluding the proof of Theorem 3.10.  $\square$

We can think of Theorem 3.10 as another necessary condition for representability in **FO(TI)**. However, it is tied to a concrete representation  $\Phi(\mathcal{I})$ . In the next step, we turn this into a general necessary condition for the represented probabilistic database that holds regardless of



the choice of representation. For this, we also include the convergence requirement of fact probabilities in any **TI**-PDB that serves as a base for the representation. To cast this into requirement over the probabilities of the probabilities of possible world, we consider the class of domain disjoint PDBs: A probabilistic database  $\mathcal{D} = (\mathbb{D}, P)$  is called *domain disjoint* if  $\text{adom}(D) \cap \text{adom}(D') = \emptyset$  for all  $D, D' \in \mathbb{D}$  with  $D \neq D'$ . Investigating this class allows us to identify a connection between the probabilities of possible worlds and the size of their active domains.

**Lemma 3.14.** *Let  $\mathcal{D} = (\mathbb{D}, P) \in \mathbf{FO}(\mathbf{TI})$  be domain-disjoint with instances  $\mathbb{D} = \{D_1, D_2, \dots\}$ . Let  $d_n := |\text{adom}(D_n)|$ . Then there exists a constant  $r \in \mathbb{N}_+$  such that for every divergent series  $\sum_{n=1}^{\infty} a_n = \infty$  of non-negative integers  $a_n$  there are infinitely many  $n \in \mathbb{N}_+$  with*

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

*Proof.* Suppose  $\mathcal{D} = \Phi(\mathcal{I})$  where  $\mathcal{I}$  is a **TI**-PDB and  $\Phi$  is an **FO**-view over the schema of  $\mathcal{I}$ . Let  $r$  be the maximum arity among the relations of the schema  $\mathcal{I}$ . For all  $n \in \mathbb{N}_+$ , we define  $A_n := \text{adom}(D_n) \setminus \text{adom}(\Phi)$  and let  $F_n \subseteq \mathbb{F} = \text{facts}(\mathcal{I})$  denote the set of facts in  $\mathcal{I}$  that contain at least one domain element from  $A_n$ . Note that the  $A_n$  are pairwise disjoint.

In the following, we abuse notation and write  $a \in f$  if  $a$  is a domain element appearing in the fact  $f$ . We know that for every  $a \in \bigcup_{I \in \text{worlds}(\mathcal{I})} \text{adom}(I) = \mathbb{A}$  there exists at most one  $n \in \mathbb{N}_+$  such that  $a \in A_n$ . Thus, it holds that

$$\sum_{n=1}^{\infty} \sum_{f \in F_n} p_f \leq \sum_{n=1}^{\infty} \sum_{a \in A_n} \sum_{\substack{f \in \mathbb{F} \\ a \in f}} p_f \leq \sum_{a \in \mathbb{A}} \sum_{\substack{f \in \mathbb{F} \\ a \in f}} p_f = \sum_{f \in \mathbb{F}} \sum_{\substack{a \in \mathbb{A} \\ a \in f}} p_f \leq \sum_{f \in \mathbb{F}} r \cdot p_f = r \cdot \sum_{f \in \mathbb{F}} p_f < \infty, \quad (4)$$

since  $\mathcal{I}$  is a well-defined **TI**-PDB. We claim that for infinitely many  $n \in \mathbb{N}_+$  we have  $\sum_{f \in F_n} p_f < \frac{a_n}{r^2}$ . Assume otherwise. Then for all but finitely many  $n$  it holds that  $\sum_{f \in F_n} p_f \geq \frac{a_n}{r^2}$ . Let  $N_0$  denote the set of these  $n$ . Of course,  $\sum_{n \in N_0} a_n$  still diverges. But then  $\sum_{n \in N_0} \sum_{f \in F_n} p_f \geq \frac{1}{r^2} \cdot \sum_{n \in N_0} a_n = \infty$ , which means that  $\sum_{n=1}^{\infty} \sum_{f \in F_n} p_f$  contains a divergent subseries in contradiction to (4).

Recall that the active domains of distinct possible worlds of  $\mathcal{D}$  are disjoint. This means that for all large enough  $n$  it holds that  $\text{adom}(D_n) \cap \text{adom}(\Phi) = \emptyset$  and, in particular,  $A_n = \text{adom}(D_n)$  and  $d_n = |A_n|$ . Let  $N$  be the (infinite) set of indices  $n$  for which  $\sum_{f \in F_n} p_f < \frac{a_n}{r^2}$ . Then, with Theorem 3.10, for all  $n \in N$  it holds that

$$\Pr_{D \sim \mathcal{D}} (D = D_n) = \Pr_{I \sim \mathcal{I}} (\Phi(I) = D_n) \leq |A_n| \cdot \left( r^2 \cdot |A_n|^{r-1} \cdot \sum_{f \in F_n} p_f \right)^{|A_n|/r} < d_n \cdot (a_n \cdot d_n^{r-1})^{d_n/r}. \quad \square$$

**Remark 3.15.** The argument from the proof of Theorem 3.14 remains valid also in the case where the active domains of the instances  $D_n$  are not disjoint, but there exists a constant bound on the number of worlds  $D_n$  that contain a domain element  $a \in \mathbb{A}$ .

The condition from Theorem 3.14 is a necessary condition for the representability of domain-disjoint PDBs in **FO(TI)**. We can now construct a domain-disjoint PDB that violates this condition, despite having the finite moments property.

**Example 3.16.** Let  $\mathcal{D} = (\mathbb{D}, P)$  be a domain-disjoint PDB such that  $\mathbb{D} = \{D_1, D_2, \dots\}$  with  $d_n := |\text{adom}(D_n)| = \lceil \log(n) \rceil$  and  $P(\{D_n\}) = \frac{Z}{n^2}$  for all  $n \in \mathbb{N}_+$ , where  $Z = \frac{6}{\pi^2}$  is the normalizing constant that makes  $P$  a probability distribution. Note that  $d_n = \lceil \log(n) \rceil \in o(n^\varepsilon)$  for all  $\varepsilon > 0$ . Thus, for all  $\varepsilon > 0$  there exists  $n(\varepsilon) \in \mathbb{N}$  such that for all  $n > n(\varepsilon)$  it holds that  $d_n \leq n^\varepsilon$ .

First, note that this PDB has the finite moments property. To see this, let  $k \in \mathbb{N}_+$  be arbitrary but fixed and let  $\varepsilon := \frac{1}{2k}$ . Then

$$\mathbb{E}_{\mathcal{D}}(| \cdot |^k) = \sum_{n=1}^{\infty} \frac{Z \cdot d_n^k}{n^2} \leq \sum_{n=1}^{n(\varepsilon)} \frac{Z \cdot d_n^k}{n^2} + \sum_{n > n(\varepsilon)} \frac{Z \cdot n^{\varepsilon k}}{n^2} = \sum_{n=1}^{n(\varepsilon)} \frac{Z \cdot d_n^k}{n^2} + Z \cdot \sum_{n > n(\varepsilon)} n^{-3/2},$$

which is finite, since the first sum has only finitely many terms and the second sum is well-known to converge.

Now towards a contradiction, assume that  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ . As  $\sum_{n=1}^{\infty} \frac{1}{n}$  diverges, by Theorem 3.14 there exists a constant  $r \in \mathbb{N}_+$  such that infinitely many  $n \in \mathbb{N}_+$  satisfy

$$\frac{Z}{n^2} = \Pr_{D \sim \mathcal{D}}(D = D_n) < d_n \cdot \left(\frac{d_n^{r-1}}{n}\right)^{d_n/r}. \quad (5)$$

For all  $n > n(\frac{1}{r})$ , it holds that  $d_n \leq n^{1/r}$ . We fix such an  $n$  that additionally satisfies  $n > \frac{1}{Z}$  and  $d_n \geq 3r^2 + r$ . Then it holds that

$$d_n \cdot \left(\frac{d_n^{r-1}}{n}\right)^{d_n/r} \leq n^{1/r} \left((n^{1/r})^{r-1} \cdot n^{-1}\right)^{d_n/r} = n^{1/r} (n^{-1/r})^{d_n/r} = (n^{1/r})^{1 - \frac{d_n}{r}} = n^{(r-d_n)/r^2} \leq n^{-3} < \frac{Z}{n^2},$$

contradicting (5).

The PDB from Theorem 3.16 is a PDB with the finite moments property that has no representation in  $\mathbf{FO}(\mathbf{TI})$ . In particular, it proves the main theorem (Theorem 3.9) of this section.

## 4 Conditional Views

In this section, we define conditional representations: representations of PDBs as views over  $\mathbf{TI}$ -PDBs conditioned on an  $\mathbf{FO}$ -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 for a given PDB, conditional representations are often simpler to identify and explain compared to unconditional ones. We show the equivalence between the PDBs that admit an  $\mathbf{FO}$ -representation and those that admit a conditional  $\mathbf{FO}$ -representation. As a consequence, we obtain a tool for showing that a PDB has a representation as an  $\mathbf{FO}$ -view over a  $\mathbf{TI}$ -PDB: it is enough to identify such a representation that is conditioned on an  $\mathbf{FO}$ -sentence.

We start by defining conditional views. Given a PDB  $\mathcal{D} = (\mathbb{D}, P)$  and an  $\mathbf{FO}$ -sentence  $\varphi$  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  $D \in \mathbb{D}_\varphi$ ,

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

Given a class  $\mathbf{D}$  of PDBs, we denote by  $\mathbf{D} \mid \mathbf{FO}$  the class of all PDBs obtained by conditioning a PDB of  $\mathbf{D}$  on an  $\mathbf{FO}$ -sentence. That is,

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

The following is the main result of this section, stating that the class of PDBs that can be represented as  $\mathbf{FO}$ -views over  $\mathbf{FO}$ -conditioned  $\mathbf{TI}$ -PDBs coincides with the class of PDBs that can be represented by an  $\mathbf{FO}$ -view of a  $\mathbf{TI}$ -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  $\mathbf{FO}$ -sentence into the  $\mathbf{FO}$ -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 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  $\mathbf{FO}$ -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. Instead of showing  $\mathbf{FO}(\mathbf{TI} \mid \mathbf{FO}) \subseteq \mathbf{FO}(\mathbf{TI})$ , we prove the equivalent statement  $\mathbf{TI} \mid \mathbf{FO} \subseteq \mathbf{FO}(\mathbf{TI})$  (this is equivalent due to  $\mathbf{TI} \mid \mathbf{FO} \subseteq \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$  and  $\mathbf{FO}(\mathbf{TI}) = \mathbf{FO}(\mathbf{FO}(\mathbf{TI}))$ ). Thus, let  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$  be a  $\mathbf{TI}$ -PDB and let  $\varphi$  be an  $\mathbf{FO}$ -sentence for which  $\mathcal{I} \mid \varphi$  is well-defined (that is,  $P_{\mathcal{I}}(\varphi) > 0$ ). Our goal is to show that  $\mathcal{I}_\varphi := \mathcal{I} \mid \varphi \in \mathbf{FO}(\mathbf{TI})$ . If  $P_{\mathcal{I}}(\varphi) = 1$ , then we are done, since then  $\mathcal{I}_\varphi = \mathcal{I} \in \mathbf{TI} \subseteq \mathbf{FO}(\mathbf{TI})$ . Therefore, for the remainder of the proof, we assume that

$$p_\varphi := P_{\mathcal{I}}(\varphi) \in (0, 1).$$

We fix some instance  $I_0 \in \mathbb{I}$  of positive probability in  $\mathcal{I}_\varphi$ . If  $\Pr_{I \sim \mathcal{I}_\varphi}(I = I_0) = 1$ , then  $I_0$  is the only instance of positive probability in  $\mathcal{I}_\varphi$ . This means that  $\mathcal{I}_\varphi$  is a  $\mathbf{TI}$ -PDB where any fact  $R(\mathbf{t})$  has marginal probability 1 if  $R(\mathbf{t}) \in I_0$ , and marginal probability 0 otherwise. Consequentially,  $\mathcal{I}_\varphi \in \mathbf{TI} \subseteq \mathbf{FO}(\mathbf{TI})$  and we are done. Thus, from now on, assume that

$$p_0 := \Pr_{I \sim \mathcal{I}_\varphi}(I = I_0) \in (0, 1).$$

Before we continue we want to convey the main idea of the proof. We are going to construct a new  $\mathbf{TI}$ -PDB consisting of multiple independent copies of the  $\mathbf{TI}$ -PDB  $\mathcal{I}$ . We define how instances of the new PDB represent instances of the PDB  $\mathcal{I}_\varphi$  other than  $I_0$ . This representation will change instance probabilities by a constant factor greater than 1. All remaining instances of the new PDB are taken as the representation of  $I_0$ . Since the other probabilities are scaled up, the probability mass of these instances is smaller than the probability of  $I_0$  in  $\mathcal{I}_\varphi$ . Therefore,

we install a gadget that acts like a switch between the two cases and carefully configure its probability to reconstruct the probability distribution of  $\mathcal{I}_\varphi$ . Finally, we show that the original instances can be reobtained from our representation by the means of an **FO**-view.

Assume that  $I_0 = \bigcup_{n=1}^N \{R_n(\mathbf{a}_{ni}) : i = 1, \dots, r_n\}$  where  $N$  is the number of relations  $(R_1, \dots, R_N)$  in the schema of  $\mathcal{I}$  and for each  $n = 1, \dots, N$ , the number  $r_n$  is the arity of the relation symbol  $R_n$ . Then  $I_0$  can be exactly characterized by an **FO**-sentence  $\varphi_0$  as follows:

$$\varphi_0 := \bigwedge_{n=1}^N \left( \forall \mathbf{x} : R_n(\mathbf{x}) \leftrightarrow \bigvee_{i=1}^{r_n} \mathbf{x} = \mathbf{a}_{ni} \right).$$

That is, for all  $I \in \mathbb{I}$  it holds that  $I \models \varphi_0$  if and only if  $I = I_0$ . We let

$$\psi := \varphi \wedge \neg \varphi_0$$

and

$$p_\psi := P_{\mathcal{I}}(\psi).$$

**Claim 4.3** (Separating  $I_0$ ). *It holds that*

$$p_\psi = (1 - p_0) \cdot p_\varphi \in (0, 1) \quad (6)$$

and for all  $I \in \mathbb{I} \setminus \{I_0\}$  it holds that

$$P_{\mathcal{I}}(\{I\} \mid \psi) = \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0}. \quad (7)$$

*Proof.* We start by calculating the probability of  $\psi$  being true in  $\mathcal{I}$ , given that  $\varphi$  holds:

$$P_{\mathcal{I}}(\psi \mid \varphi) = P_{\mathcal{I}}(\neg \varphi_0 \mid \varphi) = 1 - P_{\mathcal{I}}(\varphi_0 \mid \varphi) = 1 - p_0.$$

Moreover, for all  $I \in \mathbb{I} \setminus \{I_0\}$  it holds that

$$P_{\mathcal{I}}(\{I\} \mid \psi) = \frac{P_{\mathcal{I}}(\{I\}, \psi)}{P_{\mathcal{I}}(\psi)} = \frac{P_{\mathcal{I}}(\{I\}, \varphi)}{P_{\mathcal{I}}(\psi, \varphi)} = \frac{P_{\mathcal{I}}(\{I\} \mid \varphi)}{P_{\mathcal{I}}(\psi \mid \varphi)} = \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0}. \quad (8)$$

Therein, for the second equality we used that  $\psi \equiv \psi \wedge \varphi$  in the denominator. For the numerator, we used the fact that if  $I \not\models \varphi_0$  (which is the case since  $I \neq I_0$ ), then  $I \models \psi$  if and only if  $I \models \varphi$ . The probability of the condition  $\psi$  being satisfied in  $\mathcal{I}$  is thus calculated as follows:

$$p_\psi := P_{\mathcal{I}}(\psi) = P_{\mathcal{I}}(\psi \mid \varphi) \cdot P_{\mathcal{I}}(\varphi) = (1 - p_0) \cdot p_\varphi. \quad \dashv$$

Note that since  $p_0, p_\varphi \in (0, 1)$ , it follows that  $p_\psi = (1 - p_0) \cdot p_\varphi \in (0, 1)$ . Thus, we can choose  $k \in \mathbb{N}$  large enough such that

$$(1 - p_\psi)^k = (1 - P_{\mathcal{I}}(\psi))^k < p_0.$$

We fix such a number  $k$  and construct a new **TI**-PDB  $\mathcal{J} = (\mathbb{J}, P_{\mathcal{J}})$  with schema and marginal probabilities as follows:

### 1. Schema of $\mathcal{J}$ :

- Per relation symbol  $R$  of arity  $r$  of the schema of  $\mathcal{I}$ , the schema of  $\mathcal{J}$  contains a distinguished relation symbol  $R'$  of arity  $r + 1$ .
- We add the set  $\{1, \dots, k\}$  to the domain of  $\mathcal{J}$  (assuming, without loss of generality, that these are *new* domain elements that do not exist in  $\mathcal{I}$ ).

### 2. Marginal probabilities in $\mathcal{J}$ :

- For every fact  $R(\mathbf{a})$  appearing in  $\mathcal{I}$  with marginal probability  $p$ ,  $\mathcal{J}$  contains the facts  $R'(1, \mathbf{a}), \dots, R'(k, \mathbf{a})$ , each with marginal probability  $p$ .
- All other facts have marginal probability 0.

Since  $\mathcal{I}$  is well-defined, so is  $\mathcal{J}$ , as the sum of all fact probabilities in  $\mathcal{J}$  is bounded from above by  $k \cdot E_{\mathcal{I}}(|\cdot|) < \infty$ . The **TI**-PDB  $\mathcal{J}$  can be thought of as consisting of  $k$  independent copies of the original **TI**-PDB  $\mathcal{I}$ . The copies in  $\mathcal{J}$  can be distinguished by the identifier  $i \in \{1, \dots, k\}$  in its facts.

If  $J$  is an instance of  $\mathcal{J}$ , and  $i \in \{1, \dots, k\}$ , we let

$$J[i] := \{R(\mathbf{a}) : R'(i, \mathbf{a}) \in J\}.$$

This is then an instance in  $\mathbb{I}$ . Intuitively,  $J[i]$  is obtained from  $J$  by selecting the facts with copy identifier  $i$ , and projecting the identifier out. We introduce the following terminology:

- We call  $J$  a *representation* if there exists some  $i \in \{1, \dots, k\}$  such that  $J[i] \models \psi$ .
- We say  $J$  *represents*  $I \in \mathbb{I} \setminus \{I_0\}$  if  $J$  is a representation such that for  $i^* = \min\{i : J[i] \models \psi\}$  we have  $J[i^*] = I$ .

Note that we excluded the special instance  $I_0$  from the definition above, and that the set of instances  $I$  we can represent in  $\mathcal{J}$  is exactly the set of instances in  $\mathcal{I}$  that satisfy  $\varphi$  (minus the instance  $I_0$ ). The reason for this is that we tie representations to the sentence  $\psi = \varphi \wedge \neg\varphi_0$ , and that  $J[i] \models \psi$  entails that  $J[i] \neq I_0$ . Observe that for all  $i \in \{1, \dots, k\}$  it holds that

$$\Pr_{J \sim \mathcal{J}}(J[i] \models \psi) = \Pr_{I \sim \mathcal{I}}(I \models \psi) = p_\psi \quad (9)$$

and

$$\Pr_{J \sim \mathcal{J}}(J[i] = I \mid J[i] \models \psi) = P_{\mathcal{I}}(\{I\} \mid \psi) \stackrel{(7)}{=} \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \quad (10)$$

for all  $I \in \mathbb{I} \setminus \{I_0\}$ . The probability that an instance  $J \sim \mathcal{J}$  is a representation is given by

$$\begin{aligned} p_{\text{rep}} &:= \Pr_{J \sim \mathcal{J}}(J \text{ is a representation}) = 1 - \Pr_{J \sim \mathcal{J}}(J[i] \not\models \psi \text{ for all } i = 1, \dots, k) \\ &= 1 - (1 - p_\psi)^k > 1 - p_0, \end{aligned} \quad (11)$$

by the choice of  $k$ . Note that  $p_{\text{rep}} < 1$ , as  $p_\psi < 1$ .

**Claim 4.4** (Representations in  $\mathcal{J}$ ). *For all  $I \in \mathbb{I} \setminus \{I_0\}$  it holds that*

$$\Pr_{J \sim \mathcal{J}} (J \text{ represents } I) = \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \cdot p_{\text{rep}} \in (P_{\mathcal{I}_\varphi}(\{I\}), 1)$$

*Proof.* Let  $I \in \mathbb{I} \setminus \{I_0\}$ , and first consider the probability of the event “ $J$  represents  $I$ ”, conditioned on  $J$  being a representation:

$$\begin{aligned} & \Pr_{J \sim \mathcal{J}} (J \text{ represents } I \mid J \text{ is a representation}) \\ &= \sum_{i^*=1}^k \Pr_{J \sim \mathcal{J}} (J[i^*] = I \mid J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^*) \\ & \quad \cdot \Pr_{J \sim \mathcal{J}} (J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^* \mid J \text{ is a representation}) \\ &= \sum_{i^*=1}^k \Pr_{J \sim \mathcal{J}} (J[i^*] = I \mid J[i^*] \models \psi) \\ & \quad \cdot \Pr_{J \sim \mathcal{J}} (J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^* \mid J \text{ is a representation}) \\ &\stackrel{(10)}{=} \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \cdot \sum_{J \sim \mathcal{J}} (J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^* \mid J \text{ is a representation}). \end{aligned}$$

In the above, in the first step, we used that the events  $(J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^*)_{i^*}$  form a partition of  $I$ . In the second step, in addition to (10), we exploited that statements about  $J[i]$  are independent of statements about the  $J[i']$  with  $i' \neq i$ . Furthermore, it follows from  $\sum_{J \sim \mathcal{J}} (J[i^*] \models \psi \text{ and } J[i] \not\models \psi \text{ for all } i < i^* \mid J \text{ is a representation}) = 1$  that

$$\Pr_{J \sim \mathcal{J}} (J \text{ represents } I \mid J \text{ is a representation}) = \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0}.$$

Combining this with (11), we conclude that for  $I \in \mathbb{I} \setminus \{I_0\}$  it holds that

$$\begin{aligned} \Pr_{J \sim \mathcal{J}} (J \text{ represents } I) &= \Pr_{J \sim \mathcal{J}} (J \text{ represents } I \mid J \text{ is a representation}) \cdot \Pr_{J \sim \mathcal{J}} (J \text{ is a representation}) \\ &= \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \cdot p_{\text{rep}} \in (P_{\mathcal{I}_\varphi}(\{I\}), 1). \end{aligned} \quad \dashv$$

Starting from  $\mathcal{J}$ , we construct yet another **TI**-PDB  $\mathcal{J}_\perp$  that emerges from  $\mathcal{J}$ . We add a new, independent dummy fact  $f_\perp = R_\perp(\perp)$  to  $\mathcal{J}$  (where  $R_\perp$  is a new unary relation symbol and  $\perp$  is a new domain element), with marginal probability

$$p_\perp := \frac{p_{\text{rep}} - (1 - p_0)}{p_{\text{rep}}},$$

leaving all other facts and their marginal probabilities unchanged. Note that it follows from (11) that  $0 < p_\perp < 1$ . Note that for all  $J \sim \mathcal{J}_\perp$  it holds that  $J \setminus \{f_\perp\}$  is an instance of  $\mathcal{J}$ . With the new **TI**-PDB  $\mathcal{J}_\perp$ , we now incorporate the representation of the special instance  $I_0$  as follows: we say that  $J \sim \mathcal{J}_\perp$

- represents  $I_0$  if either  $J \setminus \{f_\perp\}$  is *not* a representation in  $\mathcal{J}$  **or**  $f_\perp \in J$  and  $J \setminus \{f_\perp\}$  is a representation in  $\mathcal{J}$ ,
- represents  $I \neq I_0$  if  $J$  does not represent  $I_0$  **and**  $J \setminus \{f_\perp\}$  represents  $I$  in  $\mathcal{J}$ .

Observe that every instance  $J$  of  $\mathcal{J}_\perp$  represents exactly one instance of  $\mathcal{I}$ . A sketch of the overall situation that we have so far is shown in Figure 1. We now show that in  $\mathcal{J}_\perp$ , the probability of the set of instances representing  $I \in \mathbb{I}$  corresponds to the probability of  $I$  in  $\mathcal{I}_\varphi$ :

**Claim 4.5** (Representations in  $\mathcal{J}_\perp$ ). *For all  $I \in \mathbb{I}$  it holds that*

$$\Pr_{J \sim \mathcal{J}_\perp} (J \text{ represents } I) = P_{\mathcal{I}_\varphi}(\{I\}) = P_{\mathcal{I}}(\{I\} \mid \varphi).$$

*Proof.* This is easily verified by direct calculation. It holds that

$$\begin{aligned} & \Pr_{J \sim \mathcal{J}_\perp} (J \text{ represents } I_0 \text{ (in } \mathcal{J}_\perp)) \\ &= \Pr_{J \sim \mathcal{J}_\perp} (J \setminus \{f_\perp\} \text{ is no representation in } \mathcal{J}) + \Pr_{J \sim \mathcal{J}_\perp} (f_\perp \in J \text{ and } J \setminus \{f_\perp\} \text{ is a representation in } \mathcal{J}) \\ &= (1 - p_{\text{rep}}) + (p_\perp \cdot p_{\text{rep}}) = (1 - p_{\text{rep}}) + (p_{\text{rep}} - (1 - p_0)) = p_0 = P_{\mathcal{I}_\varphi}(\{I_0\}). \end{aligned}$$

Moreover, for all  $I \in \mathbb{I} \setminus \{I_0\}$  it holds that

$$\begin{aligned} & \Pr_{J \sim \mathcal{J}_\perp} (J \text{ represents } I \text{ (in } \mathcal{J}_\perp)) \\ &= \Pr_{J \sim \mathcal{J}_\perp} (f_\perp \notin J \text{ and } J \setminus \{f_\perp\} \text{ represents } I \text{ (in } \mathcal{J})) \\ &= (1 - p_\perp) \cdot \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \cdot p_{\text{rep}} = \frac{(1 - p_0) - p_{\text{rep}}}{p_{\text{rep}}} \cdot \frac{P_{\mathcal{I}_\varphi}(\{I\})}{1 - p_0} \cdot p_{\text{rep}} = P_{\mathcal{I}_\varphi}(\{I\}). \end{aligned} \quad \lrcorner$$

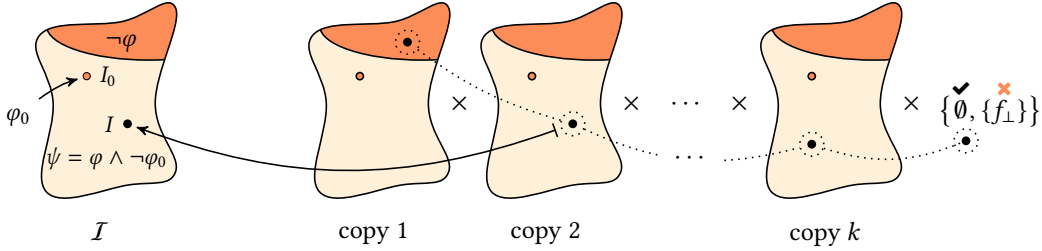


Figure 1: The above figure shows a sketch of the situation. On the left-hand side it shows the original PDB  $\mathcal{I}$  with its restriction to  $\varphi$  and the special instance  $I_0$ . On the right-hand side it shows the structure of the PDB  $\mathcal{J}_\perp$  we constructed. Every instance  $J \sim \mathcal{J}_\perp$  is essentially a tuple of  $k$  instances followed by a flag. The instance  $J$  represents  $I_0$  if either all of its parts are taken from the darker part of the copies, or it contains the flag fact  $f_\perp$ . Otherwise,  $J$  represents the instance corresponding to the first part residing in the lighter shaded part of the copies.



At this point, all that remains to show is that there exists an **FO**-view  $\Phi$  that maps each  $J \sim \mathcal{J}_\perp$  to the instance  $I \in \mathbb{I}$  (with  $I \models \varphi$ ) it represents. If  $\Phi$  has this property, then  $\Phi(\mathcal{J}_\perp) = \mathcal{I} \mid \varphi$ , proving our claim. Suppose that the relations of  $\mathcal{I}$  are  $R_1, \dots, R_N$ . Let  $\tilde{\varphi}$  be any **FO**-formula over the schema of  $\mathcal{I}$ . We assume that  $\tilde{\varphi}[i]$  is rewritten so that it contains neither  $\vee$ , nor universal quantification (this can be done in an equivalence preserving and domain-independent way, cf. [2, p. 82]). Then, for all  $i = 1, \dots, k$ , the formula  $\tilde{\varphi}[i]$  is defined from  $\tilde{\varphi}$  by structural induction as follows:

$$\tilde{\varphi}[i] := \begin{cases} R'_n(i, \mathbf{u}) & \text{if } \tilde{\varphi} = R_n(\mathbf{u}) \\ x = a & \text{if } \tilde{\varphi} = (x = a) \\ (x = y) \wedge (x \neq \perp \wedge \bigwedge_{j=1}^k x \neq j) & \text{if } \tilde{\varphi} = (x = y) \\ \neg \tilde{\varphi}'[i] \wedge \bigwedge_{x \in \text{free}(\varphi')} (x \neq \perp \wedge \bigwedge_{j=1}^k x \neq j) & \text{if } \tilde{\varphi} = \neg \tilde{\varphi}' \\ \tilde{\varphi}_1[i] \wedge \tilde{\varphi}_2[i] & \text{if } \tilde{\varphi} = \tilde{\varphi}_1 \wedge \tilde{\varphi}_2 \\ \exists x: \tilde{\varphi}'[i] & \text{if } \tilde{\varphi} = \exists x: \tilde{\varphi}'. \end{cases}$$

That is, in essence we change the atoms  $R_n(\mathbf{u})$  into  $R'_n(i, \mathbf{u})$  and guard all free variables in negated subformulae and in equalities between variables. Then, for all  $i = 1, \dots, k$  and all  $J \sim \mathcal{J}_\perp$  it holds that  $\tilde{\varphi}(J[i]) = \tilde{\varphi}[i](J)$ . This can easily be verified over the structure of  $\tilde{\varphi}$ .

Applying this construction to the **FO**-sentence  $\psi$  leaves us with **FO**-sentences  $\psi[i]$  over the schema of  $\mathcal{J}_\perp$  such that for all  $J \sim \mathcal{J}_\perp$  and all  $i = 1, \dots, k$  it holds that

$$J[i] \models \psi \Leftrightarrow J \models \psi[i].$$

Now for all  $n = 1, \dots, N$  we define the query

$$\Phi_n(\mathbf{x}) := \left[ \left( f_\perp \vee \bigwedge_{i^*=1}^k \neg \psi[i^*] \right) \wedge \left( \bigvee_{R_n(\mathbf{a}) \in I_0} \mathbf{x} = \mathbf{a} \right) \right] \wedge \left[ \left( \bigvee_{i^*=1}^k \neg f_\perp \wedge \psi[i^*] \wedge \bigwedge_{i=1}^{i^*-1} \neg \psi[i] \right) \wedge \left( R'_n(i^*, \mathbf{x}) \right) \right].$$

Then for all  $n = 1, \dots, N$  and all  $J \sim \mathcal{J}_\perp$ , it holds that  $\Phi_n(J)$  is equal to the restriction of  $I$  to the relation  $R_n$  where  $I$  is the instance from  $\mathcal{I}$  that is represented by  $J$ . By Theorem 4.5, it follows that  $\Phi(\mathcal{J}_\perp) = \mathcal{I} \mid \varphi$  where  $\Phi = \{\Phi_1, \dots, \Phi_N\}$ . That is,  $\mathcal{I} \mid \varphi \in \mathbf{FO}(\mathbf{TI})$ , concluding the proof of Theorem 4.1.  $\square$

**Remark 4.6.** 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  $\mathbf{FO}(\mathbf{TI}) \subseteq \mathbf{FO}(\mathbf{TI}) \mid \mathbf{FO} \subseteq \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO}) \subseteq \mathbf{FO}(\mathbf{TI})$ . 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 Tuple-Independent 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 Theorem 3.7 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 some constant  $c \in \mathbb{N}_+$  such that*

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

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

Note that  $(\dagger)$  looks quite similar to the requirement that the expected instance size of  $\mathcal{D}$  be finite. However, instead of weighting with the original probabilities of the possible worlds, in  $(\dagger)$  the probability of a world  $D$  is raised to the power of  $\frac{c}{|D|}$ . This makes  $(\dagger)$  more restrictive than the finite moments property.

The statement of the lemma was chosen to involve **FO(TI | FO)** rather than the equivalent **FO(TI)**, since in **FO(TI | FO)** we can also use conditioning which facilitates finding a representation. Before starting with the proof, let us introduce the main idea. So suppose that  $\mathcal{D} = (\mathbb{D}, P)$  is the probabilistic database we are given. We want to show that if there exists a constant  $c$  such that  $(\dagger)$  holds, then  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$ . For the moment, assume that  $c = 1$ , so that the condition  $(\dagger)$  becomes  $\sum_{D \in \mathbb{D} \setminus \{\emptyset\}} |D| \cdot P(\{D\})^{1/|D|} < \infty$ . We first want to construct a **TI**-PDB that serves as a base for the **FO(TI | FO)**-representation. For this, whenever  $D \in \mathbb{D}$  and  $f \in D$ , then we create a new fact that is a copy of  $f$ , tagged with an identifier for the instance  $D$ . In particular, this creates a copy of  $f$  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 construct the **FO**-condition in such a way that it filters out the possible worlds that are representations. Our **FO**-view then reconstructs the original instance from the tagged facts. The probabilities of the new facts have to be tuned in such a way that every possible world of the representation has the correct probability. The facts we constructed, together with their probabilities, have to adhere to the convergence condition from Theorem 2.5. Considering the representation mechanism, we end up with the condition  $(\dagger)$  (for  $c = 1$ ). The condition can be relaxed to the general form with a constant  $c$  by encoding  $c$  original facts into a single new fact.

*Proof of Theorem 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  $s_i := |D_i|$  denote the size of  $D_i$  and  $p_i := P(\{D_i\})$  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  $i = 0$ .) For simplicity, we assume that the schema of  $\mathcal{D}$  consists of a single  $r$ -ary relation symbol  $R$ . The proof can easily be generalized to arbitrary schemas.

Let  $c \in \mathbb{N}_+$  be some fixed constant (we will determine the value of  $c$  later). We now construct the **TI**-PDB  $\mathcal{I}$  used in our representation. The schema of  $\mathcal{I}$  consists of a single relation symbol

$\widehat{R}$  of arity  $3 + c \cdot r$ . We let  $\widehat{\mathbf{U}} := \mathbf{U} \cup \{\perp\}$  be the underlying universe of  $\mathcal{I}$ , which is the universe of  $\mathcal{D}$ , (disjointly) augmented with a new dummy symbol  $\perp$ .

For all  $i \in \mathbb{N}$ , suppose that  $D_i = \{R(\mathbf{a}_{i,1}), \dots, R(\mathbf{a}_{i,s_i})\}$  where  $\mathbf{a}_{i,j} \in \mathbf{U}^r$  for all  $1 \leq j \leq s_i$ . For  $j > s_i$ , we set  $\mathbf{a}_{i,j} := (\perp, \dots, \perp)$ . We define

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

Using this notation we define a set  $F$  of facts over the new schema  $\{\widehat{R}\}$  and the new universe  $\widehat{\mathbf{U}}$  by letting

$$F := \left\{ \widehat{R}(i, j, N_{i,j}, \mathbf{a}_{i,jc+1}, \dots, \mathbf{a}_{i,jc+c}) : i, j \in \mathbb{N} \text{ and } j \leq \max\left(\left\lceil \frac{s_i}{c} \right\rceil - 1, 0\right) \right\}$$

Essentially, we chop up every instance  $D_i$  into segments containing  $c$  facts each (where the last segment may contain  $< c$  facts). Every individual fact of  $F$  corresponds to such a segment of (up to)  $c$  facts in a database instance of  $\mathbf{D}$ . The structure of these facts is shown in Figure 2. The first entry of such a fact is the index of the original instance in which the encoded facts appear (*instance identifier*). The second entry 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 padded with dummy entries  $(\perp, \dots, \perp)$ .

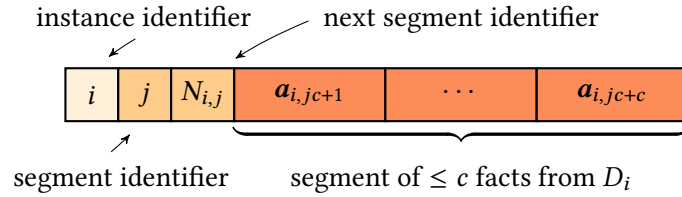


Figure 2: Structure of the new facts, and meaning of their individual components.

We now define the **TI**-PDB  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$  such that  $\text{facts}(\mathcal{I}) = F$ , giving every fact from  $F$  a positive probability (that we specify later) so that  $\mathcal{I}$  is well-defined. Let us first finalize the representation mechanism. For all  $i \in \mathbb{N}$  we let  $\widehat{D}_i \in \mathbb{I}$  be the instance from  $\mathcal{I}$  that contains precisely the facts that have instance identifier  $i$ . We say that an instance  $I \in \mathbb{I}$  represents  $D_i \in \mathbf{D}$  if 1. it contains all the facts from  $\widehat{D}_i$  (that is,  $\widehat{D}_i \subseteq I$ ), and 2. it does not contain all facts of some other  $\widehat{D}_j$  (that is,  $\widehat{D}_j \not\subseteq I$  for all  $j \neq i$ ). In general, we call an instance  $I \in \mathbb{I}$  a *representation*, if there exists some  $i \in \mathbb{N}$  such that  $I$  represents  $D_i$ .

Next, we define the marginal probabilities for  $\mathcal{I}$ . For this, first note that for  $i \in \mathbb{N}$ , the instances  $\widehat{D}_i$  are pairwise disjoint as their facts have different instance identifiers. Thus, by the construction of  $F$ , for all  $f \in F$ , there exists a unique  $i = i(f)$  such that  $f \in \widehat{D}_i$ . For the sizes  $\widehat{s}_i = |\widehat{D}_i|$  of these instances, it holds that  $\widehat{s}_i = \lceil \frac{s_i}{c} \rceil \geq 1$  for all  $i > 0$  and  $\widehat{s}_0 = 1$ . Recall that  $p_i$  is the probability of  $D_i$  in  $\mathcal{D}$  for all  $i \in \mathbb{N}$ . For every fact  $f \in F$ , we define the marginal probability  $q_f$  of  $f$  by setting

$$q_f := \left( \frac{p_{i(f)}}{1 + p_{i(f)}} \right)^{1/\widehat{s}_{i(f)}}.$$

Recall that everything we have done so far depends on the constant  $c$ .

**Claim 5.2.** *If  $c$  fulfills the condition that*

$$\sum_{i=1}^{\infty} \left\lceil \frac{s_i}{c} \right\rceil \cdot p_i^{\frac{1}{\lceil s_i/c \rceil}} < \infty \quad (\ddagger)$$

*then  $\mathcal{I}$  is a well-defined **TI**-PDB.*

*Proof.* Suppose that  $(\ddagger)$  holds for  $c$ . It holds that

$$\sum_{f \in F} q_f = \frac{p_0}{1+p_0} + \sum_{i=1}^{\infty} \left( \frac{p_i}{1+p_i} \right)^{\frac{1}{\lceil s_i/c \rceil}} \leq 1 + \sum_{i=1}^{\infty} p_i^{\frac{1}{\lceil s_i/c \rceil}} \leq 1 + \sum_{i=1}^{\infty} \left\lceil \frac{s_i}{c} \right\rceil \cdot p_i^{\frac{1}{\lceil s_i/c \rceil}} < \infty.$$

By Theorem 2.5, the **TI**-PDB  $\mathcal{I}$  is well-defined.  $\lrcorner$

Although  $(\ddagger)$  differs slightly from  $(\dagger)$ , we can work with  $(\ddagger)$  instead of  $(\dagger)$  by the following claim.

**Claim 5.3.** *The following statements are equivalent:*

1. *There exists  $c \in \mathbb{N}_+$  such that  $(\ddagger)$  holds.*
2. *There exists  $c \in \mathbb{N}_+$  such that  $(\dagger)$  holds.*

*Proof.* We start with the direction  $(1) \Rightarrow (2)$ . Suppose that  $c$  is a constant for which  $(\ddagger)$  holds. For all  $D_i$  with  $i > 0$  it holds that  $\frac{s_i}{c} \leq \left\lceil \frac{s_i}{c} \right\rceil$ . In particular,  $s_i \leq c \cdot \left\lceil \frac{s_i}{c} \right\rceil$  and  $\frac{c}{s_i} \geq 1/\left\lceil \frac{s_i}{c} \right\rceil$ . As  $p_i \in [0, 1]$ , it follows from the latter inequality that  $p_i^{c/s_i} \leq p_i^{1/\lceil s_i/c \rceil}$ . Hence,

$$\sum_{i=1}^{\infty} s_i \cdot p_i^{\frac{c}{s_i}} \leq c \cdot \sum_{i=1}^{\infty} \left\lceil \frac{s_i}{c} \right\rceil \cdot p_i^{\frac{1}{\lceil s_i/c \rceil}} < \infty,$$

so  $(\dagger)$  also holds for  $c$ .

For the other direction, suppose  $(\dagger)$  holds for the constant  $c$ . For all  $D_i$  with  $s_i > 2c$  we have

$$\left\lceil \frac{s_i}{2c} \right\rceil < \frac{s_i}{2c} + 1 = \frac{2c + s_i}{2c} < \frac{2s_i}{2c} = \frac{s_i}{c}.$$

Thus,  $1/\left\lceil \frac{s_i}{2c} \right\rceil \geq \frac{c}{s_i}$  and therefore  $p_i^{1/\lceil s_i/2c \rceil} \leq p_i^{c/s_i}$  whenever  $s_i > 2c$ . For the instances  $D_i$  with  $0 < s_i \leq 2c$ , it holds that  $\left\lceil \frac{s_i}{2c} \right\rceil = 1$  which yields

$$\sum_{i=1}^{\infty} \left\lceil \frac{s_i}{2c} \right\rceil \cdot p_i^{\frac{1}{\lceil s_i/2c \rceil}} \leq \sum_{\substack{i>0: \\ s_i \leq 2c}} p_i + \sum_{\substack{i>0: \\ s_i > 2c}} \frac{s_i}{c} p_i^{\frac{c}{s_i}} \leq 1 + \frac{1}{c} \cdot \sum_{i=1}^{\infty} s_i \cdot p_i^{\frac{c}{s_i}} < \infty,$$

so  $(\ddagger)$  holds for  $2c$ .  $\lrcorner$

Given Theorem 5.3, we fix the value of  $c$  such that  $(\ddagger)$  holds. Recall from Theorem 5.2 that  $\mathcal{I}$  is then well-defined.

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

$$\Pr_{I \sim \mathcal{I}} (\widehat{D}_i \subseteq I) = \prod_{f \in \widehat{D}_i} q_f = (q_i^{1/\widehat{s}_i})^{\widehat{s}_i} = q_i$$

as well as

$$\Pr_{I \sim \mathcal{I}} (\widehat{D}_0 \subseteq I) = q_{f_0} = q_0$$

where  $f_0$  is the unique fact in  $\widehat{D}_0$ . Moreover, note that  $0 < q_i < 1$  for all  $i \in \mathbb{N}$ . Since the sum over all  $q_i$  is finite because of  $\sum_{i=0}^{\infty} q_i \leq \sum_{i=0}^{\infty} p_i < \infty$ , it holds that

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

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

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

which yields

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

and, moreover,

$$\Pr_{I \sim \mathcal{I}} (I \text{ represents } D_i \mid I \text{ is a representation}) = \frac{\Pr_{I \sim \mathcal{I}} (I \text{ represents } D_i)}{\Pr_{I \sim \mathcal{I}} (I \text{ is a representation})} = p_i. \quad (12)$$

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

- Claim 5.4.**    1. *There exists an  $\mathbf{FO}$ -sentence  $\varphi$  such that for all  $I \in \mathbb{I}$  it holds that  $I \models \varphi$  if and only if  $I$  is a representation.*
2. *There exists an  $\mathbf{FO}$ -view  $\Phi$  that, for all  $I \in \mathbb{I}$  that are representations, maps  $I$  to the instance of  $\mathbb{D}$  it represents.*

*Proof.*    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 all facts of  $\widehat{D}_i$  but does not contain all facts of  $\widehat{D}_j$  for all  $j \neq i$ . One can check if  $\widehat{D}_i \subseteq I$  (and thus if  $\widehat{D}_j \not\subseteq I$  for  $j \neq i$ ) by checking the following:

- $I$  needs to contain a fact starting with instance identifier  $i$  and segment identifier 0; and
- whenever  $I$  contains a fact with instance identifier  $i$ , segment identifier  $j$  and next segment identifier  $j' \neq \perp$ , then  $I$  contains a fact with instance identifier  $i$  and segment identifier  $j'$  ( $j'$  will be  $j + 1$  by the definition of  $F$ ).

2. Recall that the facts of  $F$  contain up to  $c$  original facts in succession. These can be recovered by a union of  $c$  projections, under omitting the  $\perp$  entries.  $\square$

Letting  $\Phi$  and  $\varphi$  be as in Theorem 5.4, the equality Equation (12) implies that

$$\Pr_{I \sim I} (\Phi(I) = D_i \mid I \models \varphi) = p_i = P(\{D_i\})$$

for all  $i \in \mathbb{N}$  and hence,  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$ . This concludes the proof of Theorem 5.1.  $\square$

As already mentioned, due to Theorem 4.1 we directly obtain the following from Theorem 5.1.

**Theorem 5.5.** *Let  $\mathcal{D} = (\mathbb{D}, P)$  be a PDB. If there exists  $c \in \mathbb{N}_+$  such that  $(\dagger)$  holds, then  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ .*

Theorem 5.5 easily yields a representability result for the class of PDBs of bounded instance size: 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  $c$ . (Note that carrying out the construction from the proof of Theorem 5.1, every instance of the size-bounded PDB can be encoded by a single fact.)

**Corollary 5.6.** *Every PDB of bounded instance size is in  $\mathbf{FO}(\mathbf{TI})$ .*

*Proof.* Let  $\mathcal{D} = (\mathbb{D}, P)$  be a PDB and take  $c$  to be the bound on its instance sizes. Then, by Theorem 5.5,  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$  since

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

**Remark 5.7.** 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  $\mathbb{D} = \{D_1, D_2, \dots\}$ , 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 bounded by 1 although the domain and, in particular, the number of possible worlds is infinite.

The following example shows that the condition from Theorem 5.5 can also be applicable to some PDBs with unbounded instance size.

**Example 5.8.** Let  $x := \sum_{i=1}^{\infty} 2^{-i^2}$ . Then  $0 < x < \sum_{i=1}^{\infty} 2^{-i} = 1$ . Now define  $\mathcal{D} = (\mathbb{D}, P)$  with  $\mathbb{D} = \{D_1, D_2, \dots\}$  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  $(\frac{1}{x})^\alpha \leq \frac{1}{x}$  for all  $\alpha \in (0, 1]$  because  $\frac{1}{x} > 1$ . Then

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

Therefore,  $\mathcal{D}$  satisfies the condition of Theorem 5.5 for  $c = 1$ , so  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ . Yet,  $\mathcal{D}$  is of unbounded instance size.

Unfortunately, we have no full characterization of  $\mathbf{FO}(\mathbf{TI})$  yet. We know that the condition from Theorem 5.5 implies membership in  $\mathbf{FO}(\mathbf{TI})$  and that membership in  $\mathbf{FO}(\mathbf{TI})$  implies the finite moments property (Theorem 3.3). As we have seen in Theorem 3.16, the finite moments property does not imply membership in  $\mathbf{FO}(\mathbf{TI})$ . Our next example shows that membership in  $\mathbf{FO}(\mathbf{TI})$  does not imply that the condition from Theorem 5.5 is satisfied. That is, the converse of Theorem 5.5 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 5.9.** Consider the  $\mathbf{TI}$ -PDB  $\mathcal{I} = (\mathbb{I}, P)$  with fact set  $\text{facts}(\mathcal{I}) = \{R(i) : i \in \mathbb{N}_+\}$  and marginal probabilities

$$p_i = \Pr_{I \sim \mathcal{I}} (R(i) \in I) = \frac{1}{i^2 + 1} \in (0, 1).$$

First, note that  $\mathcal{I}$  is indeed well-defined as  $\sum_{i=1}^{\infty} \frac{1}{i^2 + 1} < \infty$ . Let  $Z := \prod_{i=1}^{\infty} (1 - p_i)$ . Since  $\sum_{i=1}^{\infty} p_i < \infty$ , it holds that  $0 < Z < 1$ . By the definition of tuple-independence, for all  $I \in \mathbb{I}$  it holds that

$$P(\{I\}) = \prod_{R(i) \in I} p_i \cdot \prod_{R(i) \notin I} (1 - p_i) = Z \cdot \prod_{R(i) \in I} \frac{p_i}{1 - p_i}.$$

Let  $I \neq \emptyset$ . Recall that the geometric mean of a finite set of numbers is at least the minimum of the numbers. Thus,

$$P(\{I\})^{\frac{1}{|I|}} = Z^{\frac{1}{|I|}} \cdot \left( \prod_{R(i) \in I} \frac{p_i}{1 - p_i} \right)^{\frac{1}{|I|}} \geq Z^{\frac{1}{|I|}} \cdot \min_{R(i) \in I} \frac{p_i}{1 - p_i}$$

As our sequence  $(p_i)_{i \in \mathbb{N}_+}$  is monotonically decreasing, it holds that  $\min_{R(i) \in I} p_i = p_{i(I)}$  where  $i(I) = \max\{i : R(i) \in I\}$ . Because the function  $x \mapsto \frac{x}{1-x}$  is monotonically increasing on  $(0, 1)$ , we get

$$\min_{R(i) \in I} \frac{p_i}{1 - p_i} = \frac{\min_{R(i) \in I} p_i}{1 - \min_{R(i) \in I} p_i} = \frac{p_{i(I)}}{1 - p_{i(I)}}.$$

Using  $Z^{\frac{1}{|I|}} \geq \min\{1, Z\}$ , we conclude

$$P(\{I\})^{\frac{1}{|I|}} \geq Z^{\frac{1}{|I|}} \min_{R(i) \in I} \frac{p_i}{1 - p_i} \geq \min\{1, Z\} \cdot \frac{p_{i(I)}}{1 - p_{i(I)}}.$$

Partitioning the instances according to their value of  $i(I)$ , we get the following for all  $c \in \mathbb{N}_+$ :

$$\begin{aligned} \sum_{I \neq \emptyset} |I| \cdot P(\{I\})^{\frac{c}{|I|}} &\geq \sum_{I \neq \emptyset} P(\{I\})^{\frac{c}{|I|}} \\ &= \sum_{i=1}^{\infty} \sum_{I: i(I)=i} \left( P(\{I\})^{\frac{1}{|I|}} \right)^c \\ &\geq \sum_{i=1}^{\infty} \sum_{I: i(I)=i} \left( \min\{1, Z\} \cdot \frac{p_i}{1 - p_i} \right)^c \\ &= \min\{1, Z\}^c \cdot \sum_{i=1}^{\infty} \left( \frac{p_i}{1 - p_i} \right)^c \cdot 2^{i-1}, \end{aligned}$$



because  $|I \in \mathbb{I}: i(I) = i| = 2^{i-1}$ . Note that  $\frac{p_i}{1-p_i} = \frac{1}{i^2}$ , since  $p_i = \frac{1}{i^2+1}$ . Thus,

$$\sum_{I \neq \emptyset} |I| \cdot P(\{I\})^{\frac{c}{|I|}} \geq \min\{1, Z\}^c \cdot \sum_{i=1}^{\infty} \left( \frac{p_i}{1-p_i} \right)^c \cdot 2^{i-1} = \frac{\min\{1, Z\}^c}{2} \cdot \sum_{i=1}^{\infty} \frac{2^i}{i^{2c}} = \infty,$$

since  $2^i$  grows faster than any polynomial. This shows that for  $\mathcal{I}$  there exists no  $c \in \mathbb{N}_+$  for which  $(\dagger)$  holds, so Theorem 5.5 is not applicable. Nevertheless,  $\mathcal{I}$  is in **TI**, so in particular it is in **FO(TI)**.

Thus, we still exhibit a proper gap between our conditions for containment in **FO(TI)**.

## 5.2 Block-Independent Disjoint Databases

In this section, we prove that every **BID**-PDB can be represented as an **FO** view over a **TI**-PDB. First, note that this does not follow from the previous section, as there are **BID**-PDBs that are not naturally tuple-independent and such that Theorem 5.5 does not apply to them.

**Example 5.10.** Consider the **BID**-PDB  $\mathcal{I} = (\mathbb{I}, P)$  with blocks  $B_1, B_2, \dots$  such that for all  $i$ , the block  $B_i$  contains exactly the two facts  $R_1(i)$  and  $R_2(i)$ , both with marginal probability  $p_i = \frac{1}{2(i^2+1)}$ . As  $\mathcal{I}$  is **BID**-PDB, it is also contained in **FO(BID)**. However, we now show that  $\mathcal{I}$  does not satisfy the condition from Theorem 5.5.

Let  $\mathcal{J} = (\mathbb{J}, P')$  be the **TI**-PDB from Theorem 5.9. Recall that the facts of  $\mathcal{J}$  were of the shape  $R(i)$  with  $i \in \mathbb{N}_+$ . We say an instance  $J$  of  $\mathcal{J}$  and an instance  $I$  of  $\mathcal{I}$  are *similar*, denoted  $J \simeq I$  if for all  $i \in \mathbb{N}$  it holds that

$$R(i) \in J \iff I \cap B_i = I \cap \{R_1(i), R_2(i)\} \neq \emptyset.$$

For every instance  $J \in \mathbb{J}$ , there are exactly  $2^{|J|}$  instances  $I \in \mathbb{I}$  with  $J \simeq I$ . Each such instance  $I$  has probability  $P(\{I\}) = 2^{-|J|} P'(\{J\})$ . Then it holds that

$$\begin{aligned} \sum_{I \neq \emptyset} |I| \cdot P(\{I\})^{\frac{c}{|I|}} &= \sum_{J \neq \emptyset} \sum_{I: J \simeq I} |I| \cdot P(\{I\})^{\frac{c}{|I|}} \\ &= \sum_{J \neq \emptyset} \sum_{I: J \simeq I} |I| \cdot \left( 2^{-|J|} \cdot P'(\{J\}) \right)^{\frac{c}{|I|}} \\ &= \sum_{J \neq \emptyset} 2^{|J|} \cdot |J| \cdot 2^{-c} \cdot P'(\{J\})^{\frac{c}{|J|}} \\ &\geq 2^{-c} \cdot \sum_{J \neq \emptyset} |J| \cdot P'(\{J\})^{\frac{c}{|J|}}. \end{aligned}$$

It follows from Theorem 5.9 that this sum diverges for all  $c \in \mathbb{N}_+$ . Therefore,  $\mathcal{I}$  is a **BID**-PDB for which there exists no  $c \in \mathbb{N}_+$  such that  $(\dagger)$  holds.

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

**Lemma 5.11.** **BID**  $\subseteq$  **FO(TI | FO)**.

The basic idea of the proof is as follows. We start with a **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 **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.

*Proof.* Let  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$  be a **BID**-PDB with blocks  $\{B_1, B_2, \dots\}$  such that block  $B_i$  consists of the facts  $\{f_{i,1}, f_{i,2}, \dots\}$ . 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  $p_{i,j} := \Pr_{I \sim \mathcal{I}}(f_{i,j} \in I)$ . By Theorem 2.7,  $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} p_{i,j} < \infty$ .

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  $B_i$ . for all  $\varepsilon \in (0, 1)$ , there are only finitely many residuals  $r_i$  with  $r_i < \varepsilon$  (otherwise, the sum over all fact probabilities in  $\mathcal{I}$  would diverge in contradiction to Theorem 2.7). We assume that the blocks of  $\mathcal{I}$  are indexed in increasing order of the  $r_i$ . That is,  $r_i \leq r_j$  whenever  $j \geq i$ . Let  $m$  be the non-negative integer with  $r_i = 0$  if and only if  $1 \leq i \leq m$ .

We construct a conditional, tuple-independent representation of  $\mathcal{I}$  by altering every relation in the schema of  $\mathcal{I}$  to contain an additional *block identifier* attribute. The facts of our new **TI**-PDB  $\mathcal{J} = (\mathbb{J}, P_{\mathcal{J}})$  are the facts from  $\mathcal{I}$ , augmented by the number of their block in the additional attribute. That is, for every fact  $f = R(\mathbf{a})$  from a block  $B_i$ , we construct a fact  $\tilde{f} = R'(\mathbf{a}, i)$ . Note that for  $f \neq f'$  we have  $\tilde{f} \neq \tilde{f}'$ . We let  $\tilde{B}_i = \{\tilde{f}_{i,j} : j \in \mathbb{N}_+\}$ . Then adding, respectively removing the block identifier  $i$  in the facts yields a one-to-one correspondence between  $B_i$  and  $\tilde{B}_i$ .

The marginal probabilities  $q_{i,j} = \Pr_{J \sim \mathcal{J}}(f_{i,j} \in J)$  of  $\tilde{f}_{i,j}$  in  $\mathcal{J}$  are defined in the following way:

$$q_{i,j} := \begin{cases} \frac{p_{i,j}}{1 + p_{i,j}} & \text{if } r_i = 0 \text{ and} \\ \frac{p_{i,j}}{r_i + p_{i,j}} & \text{if } r_i > 0. \end{cases}$$

We show that these marginal probabilities indeed span a well-defined **TI**-PDB. Consider  $\tilde{f}_{i,j}$  with  $i$  and  $j$  arbitrary. Then,

- if  $r_i > 0$ , it holds that  $q_{i,j} = \frac{p_{i,j}}{r_i + p_{i,j}} \leq \frac{p_{i,j}}{r_i} \leq \frac{p_{i,j}}{r_{m+1}}$ , and
- if  $r_i = 0$ , it holds that  $q_{i,j} = \frac{p_{i,j}}{1 + p_{i,j}} \leq \frac{p_{i,j}}{1} \leq \frac{p_{i,j}}{r_{m+1}}$ .

Since

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

it follows that  $\mathcal{J}$  is well-defined. The instances of interest in  $\mathcal{J}$  are those that obey the block structure of  $\mathcal{I}$ . The following claim asserts that this property is **FO**-definable.

**Claim 5.12.** *There exists an **FO**-sentence  $\varphi$  such that for all  $J \in \mathbb{J}$  it holds that  $J \models \varphi$  if and only if 1.  $J$  contains at most one fact with block identifier  $i$  for all  $i \in \mathbb{N}_+$ ; and 2.  $J$  contains exactly one fact with block identifier  $i$  for all  $i \leq m$ .*

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

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

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

Let  $\Phi$  be the view that projects the block identifier out of the facts of  $\mathbb{J}$ . The PDB  $\mathcal{J}$  together with the condition  $\varphi$  and the **FO**-view  $\Phi$  is our representation of  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$ . It is left to show that the probability distribution we obtain this way is the same as in the original PDB. That is, we need to show that for all  $I \in \mathbb{I}$  it holds that

$$\Pr_{J \sim \mathcal{J}} (\Phi(J) = I \mid J \models \varphi) = P_{\mathcal{I}}(\{I\}).$$

We first show that it suffices to check every block independently. Given an instance  $I$  of  $\mathcal{I}$ , we denote by  $I[i]$  the restriction of  $I$  to  $B_i$ . For the instances  $J$  of  $\mathcal{J}$ , we similarly let  $J[i]$  denote the restriction of  $J$  to  $\bar{B}_i$ . Since the blocks are independent in  $\mathcal{I}$ , for all  $I \in \mathbb{I}$  it holds that

$$P_{\mathcal{I}}(\{I\}) = \Pr_{I' \sim \mathcal{I}} (I'[i] = I[i] \text{ for all } i \in \mathbb{N}_+) = \prod_{i=1}^{\infty} \Pr_{I' \sim \mathcal{I}} (I'[i] = I[i]).$$

Let us now inspect the conditional representation. We denote by  $\varphi_i$  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  $B_i$ . It holds that  $J \sim \mathcal{J}$  satisfies  $\varphi$  if and only if  $J \models \varphi_i$  for all  $i \in \mathbb{N}_+$ .<sup>5</sup>

Let  $I \in \mathbb{I}$  be an instance with positive probability in  $\mathcal{I}$ . Note that for all  $J \in \mathbb{J}$  with  $J \not\models \varphi$ , the instance  $\Phi(J)$  can not be an instance of positive probability in  $\mathcal{I}$  since it either contains two distinct facts from the same block, or does not contain a fact from one of the blocks with residual 0. Thus, if  $\Phi(J) = I$  for any instance  $I$  of positive probability in  $\mathcal{I}$ , then  $J \models \varphi$ . Since the facts and therefore also the blocks are independent in  $\mathcal{J}$ , it holds that

$$\begin{aligned} \Pr_{J \sim \mathcal{J}} (\Phi(J) = I \mid J \models \varphi) &= \frac{\Pr_{J \sim \mathcal{J}} (\Phi(J) = I \text{ and } J \models \varphi)}{\Pr_{J \sim \mathcal{J}} (J \models \varphi)} \\ &= \frac{\Pr_{J \sim \mathcal{J}} (\Phi(J)[i] = I[i] \text{ for all } i \in \mathbb{N}_+)}{\Pr_{J \sim \mathcal{J}} (J[i] \models \varphi_i \text{ for all } i \in \mathbb{N}_+)} \\ &= \frac{\prod_{i=1}^{\infty} \Pr_{J \sim \mathcal{J}} (\Phi(J)[i] = I[i])}{\prod_{i=1}^{\infty} \Pr_{J \sim \mathcal{J}} (J[i] \models \varphi_i)}. \end{aligned}$$

<sup>5</sup>Although  $\varphi_i$  is expressible as an **FO**-sentence, this is not required here. We use the name  $\varphi_i$  and the notation  $\models$  for convenience.

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  $I \in \mathbb{I}$  and every block  $B_i$  it holds that

$$\frac{\Pr_{J \sim \mathcal{J}}(\Phi(J)[i] = I[i])}{\Pr_{J \sim \mathcal{J}}(J[i] \models \varphi_i)} = \Pr_{I' \sim I}(I'[i] = I[i]). \quad (13)$$

Thus, fix an arbitrary instance  $I \in \mathbb{I}$  with positive probability in  $\mathcal{I}$  and let  $i \geq 1$ . We denote  $Z_i := \prod_{\tilde{f}_{i,j} \in B_i} (1 - q_{i,j})$ . Recall that  $\tilde{B}_i$  is identical to  $B_i$  apart from the block identifier  $i$  that has been appended to the facts. Note that in  $\mathcal{J}$ , the probability of selecting only  $\tilde{f}_{i,j}$  among the facts of  $\tilde{B}_i$  is

$$q_{i,j} \cdot \prod_{\substack{\tilde{f}_{i,k} \in B_i \\ k \neq j}} (1 - q_{i,k}) = \frac{q_{i,j}}{1 - q_{i,j}} Z_i.$$

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

1. If  $r_i = 0$ , then the probability of having only fact  $\tilde{f}_{i,j}$  from  $\tilde{B}_i$  is

$$\frac{q_{i,j}}{1 - q_{i,j}} \cdot Z_i = \frac{p_{i,j}}{1 + p_{i,j}} \cdot \left(1 - \frac{p_{i,j}}{1 + p_{i,j}}\right)^{-1} \cdot Z_i = p_{i,j} \cdot Z_i.$$

Suppose that  $\tilde{f}_{i,k}$  is the unique fact from  $B_i$  in  $I$ . Then

$$\frac{\Pr_{J \sim \mathcal{J}}(\Phi(J)[i] = I[i])}{\Pr_{J \sim \mathcal{J}}(J[i] \models \varphi_i)} = \frac{p_{i,k} \cdot Z_i}{\sum_{j=1}^{\infty} p_{i,j} \cdot Z_i} = \frac{p_{i,k}}{\sum_{j=1}^{\infty} p_{i,j}} = p_{i,k} = \Pr_{I' \sim I}(I'[i] = I[i]).$$

2. If  $r_i > 0$ , then the probability of having no fact from  $\tilde{B}_i$  in  $\mathcal{J}$  is  $Z_i$ , and the probability of having only fact  $\tilde{f}_{i,j}$  from  $\tilde{B}_i$  is

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

Therefore, the probability of satisfying the condition  $\varphi_i$  is:

$$\Pr_{J \sim \mathcal{J}}(J[i] \models \varphi_i) = Z_i + \sum_{j=1}^{\infty} \frac{p_{i,j}}{r_i} \cdot Z_i = \frac{Z_i}{r_i} \cdot \left(r_i + \sum_{j=1}^{\infty} p_{i,j}\right) = \frac{Z_i}{r_i}.$$

We distinguish two subcases.

- a) In case there exists a fact  $\tilde{f}_{i,k}$  from  $B_i$  in  $I$ , it holds that

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

- b) In case no fact from  $B_i$  appears in  $I$ , it holds that

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

Since  $i$  is arbitrary, this holds for all  $i$ . Together, we have verified Equation (13). Thus,  $\mathcal{I}$  is an **FO**-view of an **FO**-conditioned **TI**-PDB, witnessed by  $\mathcal{J}$ ,  $\Phi$  and  $\varphi$ , concluding the proof of Theorem 5.11.  $\square$

From Theorem 5.11 and Theorem 4.1 we get the desired representability result for **BID**-PDBs.

**Theorem 5.13.**  $\mathbf{BID} \subseteq \mathbf{FO}(\mathbf{TI})$ .

**Remark 5.14.** Since

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

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

Similarly to Theorem 4.6,  $\mathbf{FO}(\mathbf{BID})$  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

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

The first equality above stems from the fact that

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

## 6 Seeking Logical Reasons

In the previous sections, we mainly investigated the question *whether* a given PDB is representable in a particular way without much focus on the “*why*”. The arguments concerned the probability distributions of probabilistic databases and, in particular, on convergence issues regarding the marginal probabilities of facts. For example, in Section 3 we showed probabilistic databases violating the finite moments property can not be in  $\mathbf{FO}(\mathbf{TI})$ . In essence, this is an arithmetical reason for non-representability that only takes the probabilities of possible worlds into account. There could, however, be purely logical reasons for non-representability in  $\mathbf{FO}(\mathbf{TI})$  that emerge when we forget about the probabilities and are just interested in the mere possible worlds. In this section, we aim to separate “arithmetical” arguments from “purely logical” ones.

Discarding the probabilities, we are left with sets of possible worlds such as they occur in the study of *incomplete databases* [35, 3]. Given a probabilistic database  $\mathcal{D} = (\mathbb{D}, P)$ , we focus on  $\text{worlds}(\mathcal{D})$ , the set of instances  $D \in \mathbb{D}$  with positive probability  $P(\{D\})$ . If  $\mathbf{D}$  is a class of PDBs, then  $\text{worlds}(\mathbf{D}) := \{\text{worlds}(\mathcal{D}) \mid \mathcal{D} \in \mathbf{D}\}$ . We now investigate what we can conclude about the possible representations of a probabilistic database based solely on its set of possible worlds. In Section 6.1, we show how this point of view can be used to establish the non-representability of infinite PDBs. In Section 6.2, we discuss the implications for the class  $\mathbf{FO}(\mathbf{TI})$ .

## 6.1 Proving Non-Representability

A tool that falls into the category of logical arguments is the monotonicity of certain classes of views: A view  $V: \mathbb{D}_1 \rightarrow \mathbb{D}_2$  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 [56, Proposition 2.17] the monotonicity of **UCQ**-views is exploited to show that some finite probabilistic databases can not be represented by a **UCQ**-view over a finite **TI**-PDB. This argument only needs to consider the sets of possible worlds and is oblivious to the concrete probability distributions. More precisely, they use the existence of a unique (inclusion-wise) maximal possible world in any finite **TI**-PDB and infer that a **UCQ**-view over such will also contain a maximal possible world:

**Proposition 6.1** ([56, proof of Proposition 2.17]). *Let  $\mathbf{V}$  be any class of monotone views. Then for every PDB  $\mathcal{D} \in \mathbf{V}(\mathbf{TI}_{\text{fin}})$  there exists  $D_{\text{max}} \in \text{worlds}(\mathcal{D})$  such that  $D \subseteq D_{\text{max}}$  for all  $D \in \text{worlds}(\mathcal{D})$ .*

It is then easy to construct probabilistic databases without a unique maximal possible world, and it directly follows that such PDBs cannot belong to **UCQ**(**TI**<sub>fin</sub>). This proof does not directly carry over to the infinite setting though, as infinite **TI**-PDBs in general may not have maximal possible worlds at all. However, a simple argument that works in both the finite and the infinite setting demonstrates how monotonicity can be used with respect to the set of possible worlds in order to conclude non-representability in general. For this, we start with some easy observations regarding the structure of the possible worlds of **TI**-PDBs.

**Observation 6.2.** *Let  $\mathcal{I}$  be a **TI**-PDB, and let  $p_f$  denote the marginal probability of fact  $f$ . Then  $\text{facts}(\mathcal{I})$  is partitioned into the sets  $F_{\text{always}} := \{f: p_f = 1\}$ ,  $F_{\text{sometimes}} := \{f: p_f \in (0, 1)\}$  and  $F_{\text{never}} := \{f: p_f = 0\}$  and it holds that  $\text{worlds}(\mathcal{I}) = \{F_{\text{always}} \cup F: F \subseteq F_{\text{sometimes}} \text{ and } |F| < \infty\}$ .*

We next inspect the effect of applying views over PDBs and observe that stripping PDBs of their probabilities commutes with applying views. That is, the diagram in Figure 3a commutes.

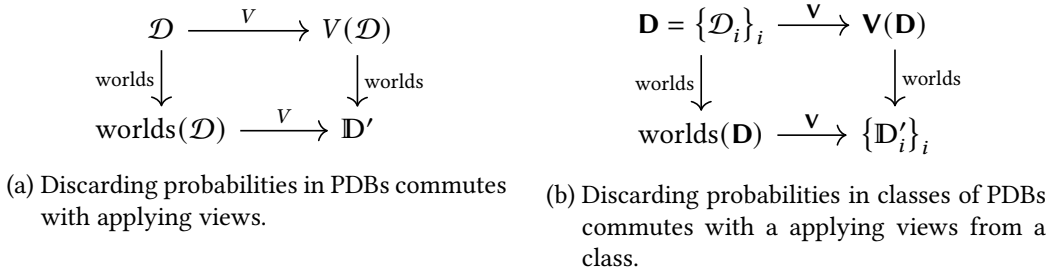


Figure 3: Commutative diagrams illustrating Theorem 6.3 and Theorem 6.4.

**Observation 6.3.** *Let  $\mathcal{D} = (\mathbb{D}_1, P)$  be a PDB, let  $\mathbb{D}_2$  be a set of database instances, and let  $V$  be a view  $V: \mathbb{D}_1 \rightarrow \mathbb{D}_2$ . Then, it holds that  $V(\text{worlds}(\mathcal{D})) = \text{worlds}(V(\mathcal{D}))$ .*

*Proof.* If  $D_2 \in V(\text{worlds}(\mathcal{D}))$ , then there exists  $D_1 \in \text{worlds}(\mathcal{D})$  such that  $V(D_1) = D_2$ . As  $P(\{D_1\}) > 0$ , this implies that  $P_{V(\mathcal{D})}(\{D_2\}) > 0$ . Thus,  $V(\text{worlds}(\mathcal{D})) \subseteq \text{worlds}(V(\mathcal{D}))$ .

If  $D_2 \in \text{worlds}(V(\mathcal{D}))$ , then  $P_{\mathcal{D}}(V^{-1}(\{D_2\})) = P_{V(\mathcal{D})}(\{D_2\}) > 0$ . Thus, there exists an instance  $D_1 \in V^{-1}(\{D_2\})$  with  $P_{\mathcal{D}}(D_1) > 0$ . In particular,  $D_2 \in V(\text{worlds}(\mathcal{D}))$ , which shows  $\text{worlds}(V(\mathcal{D})) \subseteq V(\text{worlds}(\mathcal{D}))$ .  $\square$

Theorem 6.3 directly implies the following.

**Proposition 6.4.** *Let  $\mathbf{D}$  be a class of PDBs and let  $\mathbf{V}$  be a class of views. Then, it holds that  $\mathbf{V}(\text{worlds}(\mathbf{D})) = \text{worlds}(\mathbf{V}(\mathbf{D}))$ . In particular,  $\text{worlds}(\mathcal{D}) \in \mathbf{V}(\text{worlds}(\mathbf{D}))$  holds for all  $\mathcal{D} \in \mathbf{V}(\mathbf{D})$ .*

The first part of Theorem 6.4 states that the diagram in Figure 3b commutes. The latter part of the proposition can be understood as a purely logical necessary condition for representability: if  $\text{worlds}(\mathcal{D}) \notin \mathbf{V}(\text{worlds}(\mathbf{D}))$ , then it follows immediately that  $\mathcal{D} \notin \mathbf{V}(\mathbf{D})$ . We now build on Theorem 6.4 and Theorem 6.2, and demonstrate how monotonicity can be used to conclude non-representability in the setting of infinite PDBs.

**Proposition 6.5.** *Let  $\mathbf{V}$  be a class of monotone views, and let  $\mathcal{D}$  be a PDB such that there exist two facts  $f_1$  and  $f_2$  of positive marginal probability that are mutually exclusive (that is,  $\Pr_{D \sim \mathcal{D}}(f_1, f_2 \in D) = 0$ ). Then  $\mathcal{D} \notin \mathbf{V}(\mathbf{TI})$ .*

*Proof.* Assume by contradiction that  $\mathcal{D} \in \mathbf{V}(\mathbf{TI})$  and let  $D_1, D_2 \in \text{worlds}(\mathcal{D})$  with  $f_1 \in D_1$  and  $f_2 \in D_2$ . By Theorem 6.4, we know that  $\text{worlds}(\mathcal{D}) \in \mathbf{V}(\text{worlds}(\mathbf{TI}))$ . Hence, there exists a set of instances  $\mathbb{I} \in \text{worlds}(\mathbf{TI})$  and a view  $V \in \mathbf{V}$  such that  $V(\mathbb{I}) = \text{worlds}(\mathcal{D})$ . Thus, let  $I_1, I_2 \in \mathbb{I}$  such that  $V(I_1) = D_1$  and  $V(I_2) = D_2$ . By Theorem 6.2, we also have  $I_1 \cup I_2 \in \mathbb{I}$  and, thus,  $V(I_1 \cup I_2) \in \text{worlds}(\mathcal{D})$ . Since  $V$  is monotone, it holds that

$$V(I_1 \cup I_2) \supseteq V(I_1) \cup V(I_2) = D_1 \cup D_2 \supseteq \{f_1, f_2\},$$

in contradiction to  $f_1$  and  $f_2$  being mutual exclusive.  $\square$

Theorem 6.5 immediately yields that **UCQ**(**TI**) does not contain any **BID**-PDBs that are not already tuple-independent.

## 6.2 First-Order Views

Theorem 6.5 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 **FO**(**TI**). More precisely, for any collection of possible worlds (over the same schema), we can attach probabilities to the worlds in such a way that we end up with a PDB in **FO**(**TI**).

**Lemma 6.6.** *Let  $\mathbb{ID}$  be any set of database instances over the same schema. Then there exists a PDB  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$  such that  $\text{worlds}(\mathcal{D}) = \mathbb{ID}$ .*

*Proof.* Let  $\mathbb{ID} = \{D_1, D_2, \dots\}$  and define

$$z_i := \begin{cases} (2^{-i} \cdot |D_i|^{-1})^{|D_i|} & \text{if } D_i \neq \emptyset \text{ and} \\ 1 & \text{otherwise.} \end{cases}$$



Note that  $z_i \in (0, 1]$ . It holds that

$$0 < \sum_{i=1}^{\infty} z_i \leq 1 + \sum_{i, D_i \neq \emptyset} \left( \frac{1}{2^i \cdot |D_i|} \right)^{|D_i|} \leq 1 + \sum_{i, D_i \neq \emptyset} \frac{1}{2^i \cdot |D_i|} \leq 1 + \sum_{i, D_i \neq \emptyset} \frac{1}{2^i} \leq 2.$$

Thus,  $\frac{1}{Z} \in [\frac{1}{2}, \infty)$  for  $Z := \sum_{i=1}^{\infty} z_i$ . We now define probability distribution  $P: \mathbb{D} \rightarrow [0, 1]$  by letting  $P(\{D_i\}) := \frac{z_i}{Z}$  for all  $i \in \mathbb{N}_+$ . Then it holds that

$$\sum_{D \in \mathbb{D} \setminus \emptyset} |D| \cdot P(\{D\})^{1/|D|} = \sum_{i, D_i \neq \emptyset} |D_i| \cdot \left( \frac{z_i}{Z} \right)^{1/|D_i|} = \sum_{i, D_i \neq \emptyset} \frac{1}{2^i} \cdot \left( \frac{1}{Z} \right)^{1/|D_i|} \leq \sum_{i=1}^{\infty} \frac{1}{2^i} \cdot \max\{1, \frac{1}{Z}\} < \infty.$$

This means that  $\mathcal{D} = (\mathbb{D}, P)$  satisfies the condition from Theorem 5.5, and thus,  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ .  $\square$

Theorem 6.6 shows that in order to prove that a PDB is not in  $\mathbf{FO}(\mathbf{TI})$ , it is not enough to only consider the set of its possible worlds. Any such proof *must* take probabilities of the possible worlds into account.

Similar to the terminology for PDBs, we call a set  $\mathbb{D}$  of database instances *of unbounded instance size*, if for all  $n \in \mathbb{N}$  there exists some  $D \in \mathbb{D}$  with  $|D| \geq n$ . Recall that we already know from Theorem 5.6, that if the underlying set of possible worlds of a PDB is of bounded instance size, then the PDB is in  $\mathbf{FO}(\mathbf{TI})$ , regardless of the probabilities. Theorem 6.6 says that for any tentative set of possible worlds of unbounded instance size, we can always find a probability distribution such that the resulting PDB is in  $\mathbf{FO}(\mathbf{TI})$ . The following result, however, states that in this setting we can also always find probability distributions such that the resulting PDB will *not* be in  $\mathbf{FO}(\mathbf{TI})$ .

**Lemma 6.7.** *Let  $\mathbb{D}$  be a set of instances over the same schema such that  $\mathbb{D}$  is of unbounded instance size. Then there exists a PDB  $\mathcal{D}$  with  $E(|\cdot|) = \infty$  such that  $\text{worlds}(\mathcal{D}) = \mathbb{D}$ .*

*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, in particular,  $|D_{i_k}| \geq k$ . For all  $k \in \mathbb{N}_+$ , we define

$$P(\{D_{i_k}\}) := \frac{3}{\pi^2 \cdot k^2}.$$

Then  $\sum_{k=1}^{\infty} P(\{D_{i_k}\}) = \frac{1}{2}$ . We assign positive probabilities to the instances in  $\mathbb{D} \setminus \{D_{i_k} : k \in \mathbb{N}_+\}$  such that they also add up to  $\frac{1}{2}$ . Then  $P: \mathbb{D} \rightarrow [0, 1]$  is a probability distribution and  $\mathcal{D} = (\mathbb{D}, P)$  is a probabilistic database. However, the expected instance size of  $\mathcal{D}$  is infinite:

$$E_{\mathcal{D}}(|\cdot|) = \sum_{D \in \mathbb{D}} |D| \cdot P(\{D\}) \geq \sum_{k=1}^{\infty} |D_{i_k}| \cdot P(\{D_{i_k}\}) \geq \sum_{k=1}^{\infty} k \cdot \frac{3}{\pi^2 \cdot k^2} = \frac{3}{\pi^2} \cdot \sum_{k=1}^{\infty} \frac{1}{k} = \infty. \quad \square$$

The following theorem summarizes our results regarding the class  $\mathbf{FO}(\mathbf{TI})$  that emerged from the inspection of the underlying sets of possible worlds.

**Theorem 6.8.** *Let  $\mathbb{D}$  be a set of database instances over the same schema.*

1. If  $\mathbb{D}$  has bounded instance size, then  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$  for every  $\mathcal{D}$  with  $\text{worlds}(\mathcal{D}) = \mathbb{D}$ .
2. If  $\mathbb{D}$  has unbounded instance size, then there exist PDBs  $\mathcal{D}_1 \in \mathbf{FO}(\mathbf{TI})$  and  $\mathcal{D}_2 \notin \mathbf{FO}(\mathbf{TI})$  such that  $\text{worlds}(\mathcal{D}_1) = \text{worlds}(\mathcal{D}_2) = \mathbb{D}$ .

*Proof.* The first statement directly follows from Theorem 5.6. For unbounded instance size, Theorem 6.6 shows the existence of  $\mathcal{D}_1$  and Theorem 6.7 shows the existence of  $\mathcal{D}_2$ , where the latter is not in  $\mathbf{FO}(\mathbf{TI})$  due to Theorem 3.7.  $\square$

In short, assume we are given a PDB  $\mathcal{D}$  and want to determine whether  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ . If  $\text{worlds}(\mathcal{D})$  has bounded instance size, then  $\mathcal{D} \in \mathbf{FO}(\mathbf{TI})$ . Otherwise, we have to investigate the instance probabilities to settle the question.

## 7 Views Using Fragments of First Order-Logic

### 7.1 The Situation for Finite Probabilistic Databases

In this section, we review the relative expressive power of classes of views over finite PDBs. Recall that  $\mathbf{FO}(\mathbf{TI}_{\text{fin}}) = \mathbf{PDB}_{\text{fin}}$ , as shown in [56, Proposition 2.16], but  $\mathbf{UCQ}(\mathbf{TI}_{\text{fin}}) \subsetneq \mathbf{PDB}_{\text{fin}}$ , as shown in [56, Proposition 2.17]. The following two examples show that the classes  $\mathbf{UCQ}(\mathbf{TI}_{\text{fin}})$ ,  $\mathbf{CQ}(\mathbf{TI}_{\text{fin}})$  and  $\mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}})$  are incomparable to  $\mathbf{BID}_{\text{fin}}$  (with respect to  $\subseteq$ ).

**Example 7.1.** Consider the  $\mathbf{BID}$ -PDB  $\mathcal{I}$  that consists of only a single block containing two facts  $f$  and  $f'$ , each with marginal probability  $\frac{1}{2}$ . Then  $\text{worlds}(\mathcal{I}) = \{\{f\}, \{f'\}\}$  and both worlds  $\{f\}$  and  $\{f'\}$  are maximal. It follows from Theorem 6.1 that  $\mathcal{I} \notin \mathbf{V}(\mathbf{TI}_{\text{fin}})$  for any class  $\mathbf{V}$  of monotone views. In particular,  $\mathcal{I} \notin \mathbf{TI}_{\text{fin}}$ .

**Example 7.2.** Consider the  $\mathbf{TI}$ -PDB  $\mathcal{I} = (\mathbb{I}, P)$  with  $\text{facts}(\mathcal{I}) = \{R(1, 1), R(1, 2), R(2, 2), S(1), S(2)\}$ , where the marginal probability of all  $R$ -facts is 1 and the marginal probability of both  $S$ -facts is  $\frac{1}{2}$ . Now consider the  $\mathbf{sjfCQ}$ -view  $\Phi(x) := \exists y: R(x, y) \wedge S(y)$ . Then  $\Phi(\mathcal{I}) \in \mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}})$ . Note that the PDB  $\mathcal{I}$  has possible worlds and images according to  $\Phi$  as shown in Table 1. Thus,

Table 1: Possible worlds of the PDB  $\mathcal{I}$ .

$I$	$\Phi(I)$
$\{R(1, 1), R(1, 2), R(2, 2)\}$	$\emptyset$
$\{R(1, 1), R(1, 2), R(2, 2), S(1)\}$	$\{R_\Phi(1)\}$
$\{R(1, 1), R(1, 2), R(2, 2), S(2)\}$	$\{R_\Phi(1), R_\Phi(2)\}$
$\{R(1, 1), R(1, 2), R(2, 2), S(1), S(2)\}$	$\{R_\Phi(1), R_\Phi(2)\}$

$\Phi(\mathcal{I})$  has exactly three possible worlds: the empty instance  $\emptyset$  with probability  $\frac{1}{4}$ , the instance  $\{R_\Phi(1)\}$ , also with probability  $\frac{1}{4}$ , and the instance  $\{R_\Phi(1), R_\Phi(2)\}$  with probability  $\frac{1}{2}$ . Since the facts  $R_\Phi(1)$  and  $R_\Phi(2)$  are neither independent nor mutually exclusive in  $\Phi(\mathcal{I})$ , it holds that  $\Phi(\mathcal{I}) \notin \mathbf{BID}_{\text{fin}}$ .

Note that the latter example also shows that  $\mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}}) \supsetneq \mathbf{TI}_{\text{fin}}$ . Remarkably, it turns out that the classes  $\mathbf{CQ}(\mathbf{TI}_{\text{fin}})$  and  $\mathbf{UCQ}(\mathbf{TI}_{\text{fin}})$  collapse to  $\mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}})$ . This is a consequence of the following, more general insight.

**Proposition 7.3.** *Let  $\mathcal{D} = V(\mathcal{I})$  where  $\mathcal{I} \in \mathbf{TI}_{\text{fin}}$  and  $V$  is a monotone view. Then  $\mathcal{D} \in \mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}})$ .*

*Proof.* We show this for the case where  $V = \{Q\}$  is a single query with output relation symbol  $R_Q = R$ . This generalizes directly to general monotone views. As earlier, for  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$ , we let  $F_{\text{always}}(\mathcal{I})$  and  $F_{\text{sometimes}}(\mathcal{I})$  denote the sets of facts that appear with probability 1 and  $> 0$ , respectively. Suppose that  $F_{\text{sometimes}}(\mathcal{I}) = \{f_1, \dots, f_n\}$  (with  $f_i$  pairwise different). We construct a  $\mathbf{TI}_{\text{fin}}$ -PDB  $\mathcal{J} = (\mathbb{J}, P_{\mathcal{J}})$  over the same universe as  $\mathcal{I}$ , augmented (disjointly) by  $\{0, 1\}$ . The schema of  $\mathcal{J}$  consists of  $n$  new unary relation symbols  $S_1, \dots, S_n$ , and a separate new  $n + r$ -ary relation symbol  $S$ , where  $r$  is the arity of the query  $Q$ . The marginal probabilities of the facts in  $\mathcal{J}$  are given as follows:

- The fact  $S_i(0)$  has marginal probability 1 for all  $i = 1, \dots, n$ .
- The fact  $S_i(1)$  has marginal probability  $P_{\mathcal{I}}(f_i)$  for all  $i = 1, \dots, n$ .
- For all  $a_1, \dots, a_n \in \{0, 1\}$  and all facts  $R(b_1, \dots, b_r) \in Q(F_{\text{always}}(\mathcal{I}) \cup \{f_i : a_i \neq 0\})$ , the fact  $S(a_1, \dots, a_n, b_1, \dots, b_r)$  has marginal probability 1.
- All other facts have probability 0.

Consider the following  $\mathbf{sjfCQ}$ -query  $\Phi$  with output relation symbol  $R_{\Phi} = R$  defined by

$$\Phi(y_1, \dots, y_r) = \exists x_1 \dots \exists x_n : S_1(x_1) \wedge \dots \wedge S_n(x_n) \wedge S(x_1, \dots, x_n, y_1, \dots, y_r).$$

We claim that  $\Phi(\mathcal{J}) = Q(\mathcal{I})$ . We show this by constructing a bijection  $\beta$  between the worlds of  $\mathcal{I}$  and the worlds of  $\mathcal{J}$  with the properties that 1.  $P_{\mathcal{I}}(I) = P_{\mathcal{J}}(\beta(I))$ ; and 2.  $Q(I) = \Phi(J)$ .

Let  $I$  be any possible world of  $\mathcal{I}$ . Then  $I = F_{\text{always}}(\mathcal{I}) \cup \{f_{i_1}, \dots, f_{i_k}\}$  for some distinct facts  $f_{i_1}, \dots, f_{i_k} \in F_{\text{sometimes}}(\mathcal{I})$ . We let  $\beta(I)$  be the instance of  $\mathcal{J}$  that contains all facts of marginal probability 1, along with (precisely) the facts  $S_{i_1}(1), \dots, S_{i_k}(1)$ . Clearly,  $\beta$  is both injective and surjective, i. e., a bijection. Moreover, we have that

$$P_{\mathcal{I}}(\{I\}) = \prod_{j=1}^k P_{\mathcal{I}}(f_{i_j}) = \prod_{j=1}^k P_{\mathcal{J}}(S_{i_j}(1)) = P_{\mathcal{J}}(\{J\}).$$

Therefore, it only remains to show that for all  $I \in \text{worlds}(\mathcal{I})$ , it holds that  $Q(I) = \Phi(\beta(I))$ .

Let  $I = F_{\text{always}}(\mathcal{I}) \cup \{f_{i_1}, \dots, f_{i_k}\}$  as before and suppose that  $R(b_1, \dots, b_r) \in Q(I)$ . By construction, the instance  $\beta(I)$  contains the facts  $S_{i_1}(1), \dots, S_{i_k}(1)$ . Moreover, as  $R(b_1, \dots, b_r) \in Q(I)$ , by definition, the fact  $S(a_1, \dots, a_n, b_1, \dots, b_r)$  where

$$a_i = \begin{cases} 1 & \text{if } i \in \{i_1, \dots, i_k\} \\ 0 & \text{otherwise.} \end{cases}$$

for all  $i = 1, \dots, n$  is one of the facts with marginal probability 1 in  $\mathcal{J}$ . Together this shows that  $J \models \Phi(b_1, \dots, b_r)$ , i.e.  $R(b_1, \dots, b_r) \in \Phi(J)$ .

For the other direction, suppose that  $R(b_1, \dots, b_r) \in \Phi(\beta(I))$ . This means that there exist  $a_1, \dots, a_n$  such that  $S_1(a_1), \dots, S_n(a_n) \in \beta(I)$  and  $R(b_1, \dots, b_r) \in Q(F_{\text{always}}(\mathcal{I}) \cup \{f_j : a_j \neq 0\})$ . Note that

$$\{f_j : a_j \neq 0\} \subseteq \{f_j : S_j(1) \in \beta(I)\}.$$

As  $Q$  is monotonous, we have that

$$Q\left(F_{\text{always}}(\mathcal{I}) \cup \{f_j : a_j \neq 0\}\right) \subseteq Q\left(F_{\text{always}}(\mathcal{I}) \cup \{f_j : S_j(1) \in \beta(I)\}\right) = Q(I),$$

which finally yields  $R(b_1, \dots, b_r) \in Q(I)$ .  $\square$

**Corollary 7.4.**  $\mathbf{sjfCQ}(\mathbf{TI}_{\text{fin}}) = \mathbf{CQ}(\mathbf{TI}_{\text{fin}}) = \mathbf{UCQ}(\mathbf{TI}_{\text{fin}})$ .

For finite **BID**-PDBs, the classes  $\mathbf{sjfCQ}(\mathbf{BID}_{\text{fin}})$ ,  $\mathbf{CQ}(\mathbf{BID}_{\text{fin}})$  and  $\mathbf{UCQ}(\mathbf{BID}_{\text{fin}})$ : in [56, Proposition 2.18] it is shown that  $\mathbf{PDB}_{\text{fin}} = \mathbf{CQ}(\mathbf{BID}_{\text{fin}})$ . In fact, the conjunctive query used in the proof is self-join free, showing that  $\mathbf{PDB}_{\text{fin}} = \mathbf{sjfCQ}(\mathbf{BID}_{\text{fin}})$ .

**Proposition 7.5.** *The classes of views over finite probabilistic databases have the relative expressive power shown in the Hasse diagram of Figure 4.*

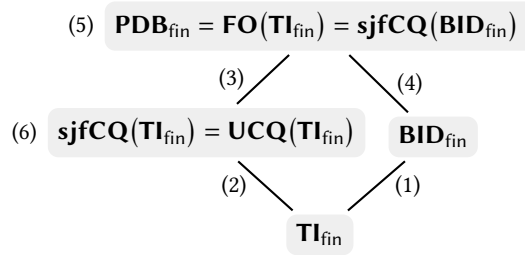


Figure 4: Hasse diagram for the relative expressive power of views over finite probabilistic databases.

*Proof.* Consider Figure 4. The inclusions depicted therein (shown as solid lines) are trivial. We now give arguments for (1) throughout (4) that these inclusions are proper and that the remaining classes (5) and (6) collapse.

1. This is witnessed by Theorem 7.1.
2. This is witnessed by Theorem 7.2.
3. This is shown in [56, Proposition 2.17].
4. This is shown in [56, Chapter 2.7], another example is given by Theorem 7.2.
5. This is shown in [56, Propositions 2.16 and 2.18].
6. This is the statement of Theorem 7.4.  $\square$

## 7.2 The Situation for Countably Infinite Probabilistic Databases

In the previous section, we discussed finite **TI**- and **BID**-PDBs and views in the **UCQ**-, **CQ**- and **sjfCQ**-fragment of first-order logic. We now want to investigate the corresponding classes in the countably infinite setting. Interestingly, the different classes that collapse in the finite setting do no longer collapse in the infinite setting, yielding pairwise different classes of probabilistic databases. A reason for is intuitively that in a representation, we can no longer hard-code the complete structure of possible worlds using facts of probability one. This was possible in the finite setting, and is essentially what is done in the proofs of the respective statements [56, Propositions 2.16 and 2.18].

The tools we have seen and developed so far are ineffective for a possible separation of the classes **UCQ(TI)**, **CQ(TI)** and **sjfCQ(TI)** and the corresponding classes for **BID**-PDBs as they only concern the class **FO(TI)** and we already know from Theorems 5.13 and 6.5 that **UCQ(TI)** is strictly weaker than **FO(TI)**. Thus, we have to develop new methods.

The separations we find in this section rely on the varying level of capability for the respective classes to represent symmetries. For our arguments, we thus focus on a special class of probabilistic databases that facilitates such discussions. A *graph database*  $D$  is a database instance over a single binary relation  $E$  with universe  $\mathbb{U}^2$ . It is called

- *simple*, if it contains no fact  $E(a, a)$  and
- *undirected*, if  $E(a, b) \in D$  implies  $E(b, a) \in D$ .

A *probabilistic graph* is a PDB whose possible worlds are graph databases.

**Definition 7.6.** Consider the class  $\mathcal{G}$  of probabilistic databases  $\mathcal{D}$  over ordered, countably infinite universes  $(\mathbb{U}, <)$  satisfying all of the following properties:

- every possible world of  $\mathcal{D}$  is a simple and undirected graph database,
- it is *edge-independent*, that is, for any sequence of pairwise distinct two-element subsets  $\{a_1, b_1\}, \dots, \{a_k, b_k\}$  of  $\mathbb{U}$ , the facts  $E(a_1, b_1), \dots, E(a_k, b_k)$  are stochastically independent in  $\mathcal{D}$ , that is,

$$\Pr_{D \sim \mathcal{D}} (E(a_i, b_i) \in D \text{ for all } i = 1, \dots, k) = \prod_{i=1}^k p_{a_i, b_i}.$$

and

- it is *unbounded*, that is, for every  $n \in \mathbb{N}$  there exists  $D \in \text{worlds}(\mathcal{D})$  with  $|D| > n$ .

If  $\mathcal{D}$  is a PDB in  $\mathcal{G}$ , we let  $p_{a,b}$  denote the marginal probability of the fact  $E(a, b)$ , that is,  $p_{a,b} = \Pr_{D \sim \mathcal{D}} (E(a, b) \in D)$ . Since every possible world of  $\mathcal{D}$  is an undirected (and simple) graph database, we have  $p_{a,b} = p_{b,a}$  for all  $a$  and  $b$ . Note that  $\mathcal{D}$  is uniquely determined by  $(p_{a,b})_{a \neq b}$ .

**Remark 7.7.** Intuitively, every PDB  $\mathcal{D}$  in  $\mathcal{G}$  represents a random graph with vertex set  $\mathbb{U}$  and edges  $\{a, b\} \subseteq \binom{\mathbb{U}}{2}$  where all edges are independent, and the probability of the edge  $\{a, b\}$  being present is precisely  $p_{a,b}$ . Thus, these PDBs are tightly connected to a particular model that arises in the study of random graphs [27, Chapter 10.1].

Naturally, despite not being tuple-independent, the PDBs of  $\mathcal{G}$  are strongly related to **TI**-PDBs.

**Proposition 7.8.** *Let  $(p_{a,b})_{a < b}$  be a family of numbers in  $[0, 1]$  and let  $p_{b,a} = p_{a,b}$  for all  $a, b$ . Then  $(p_{a,b})_{a \neq b}$  defines a probabilistic graph in  $\mathcal{G}$  if and only if  $(p_{a,b})_{a < b}$ , taken as marginal probabilities for the facts  $\{E(a, b)\}_{a < b}$ , defines a **TI**-PDB.*

In particular,  $(p_{a,b})_{a \neq b}$  defines a probabilistic graph in  $\mathcal{G}$  if and only if  $\sum_{a \neq b} p_{a,b}$  is finite.

*Proof.* First, let  $\mathcal{D} \in \mathcal{G}$ . We construct a new PDB  $\mathcal{D}'$  by restricting  $\mathcal{D}$  to the facts  $\{E(a, b)\}_{a < b}$ . In detail, for every world  $D \in \text{worlds}(\mathcal{D})$  we only keep the facts  $E(a, b)$  with  $a < b$ . Note that this transformation is a one-to-one-correspondence between  $\text{worlds}(\mathcal{D})$  and  $\text{worlds}(\mathcal{D}')$ . Therefore,  $\mathcal{D}'$  is well-defined. Moreover, all the facts in  $\mathcal{D}'$  are independent and have the same marginal probabilities as before, rendering  $\mathcal{D}'$  a **TI**-PDB with marginal probabilities  $(p_{a,b})_{a < b}$ .

Second, let  $\mathcal{I}$  be a **TI**-PDB over  $\{E(a, b)\}_{a < b}$  with marginal probabilities  $(p_{a,b})_{a < b}$ . Inverting the construction from the first step, we define  $\mathcal{D}$  by adding  $E(b, a)$  to every possible world that contains  $E(a, b)$  for every  $a < b$ .  $\square$

It is quite easy to see that every PDB in  $\mathcal{G}$  admits a **UCQ(TI)**-representation.

**Lemma 7.9.** *If  $\mathcal{D} \in \mathcal{G}$ , then  $\mathcal{D} \in \text{UCQ(TI)}$ .*

*Proof.* Let  $\mathcal{D} = (\mathbb{D}, P_{\mathcal{D}})$  be a probabilistic graph from  $\mathcal{G}$  over  $\mathbb{U}$ , and let  $\mathcal{I}$  be the **TI**-PDB constructed from  $\mathcal{D}$  as in Theorem 7.8. We show that  $\mathcal{D}$  can be reconstructed from  $\mathcal{I}$  using the **UCQ**-view  $\Phi(x, y) := E(x, y) \vee E(y, x)$ . That is, we claim that  $\Phi(\mathcal{I}) = \mathcal{D}$ . To see this, consider any possible world  $D \in \text{worlds}(\mathcal{D})$ . Note that

$$P_{\mathcal{D}}(D) = \prod_{\substack{a < b \text{ and} \\ E(a, b) \in D}} p_{a,b} \cdot \prod_{\substack{a < b \text{ and} \\ E(a, b) \notin D}} (1 - p_{a,b}). \quad (14)$$

The preimage of  $D$  under  $\Phi$  consists of only the single instance

$$I = \{E(a, b) : a < b \text{ and } E(a, b) \in D\} \in \mathbb{I}.$$

By construction, the probability of this instance in  $\mathcal{I}$  is equal to (14) as well. That is,  $\Phi$  is a probability preserving one-to-one correspondence between the worlds of  $\mathcal{I}$  and the worlds of  $\mathcal{D}$ , so  $\Phi(\mathcal{I}) = \mathcal{D}$ .  $\square$

The next lemma gives an example of the other extreme: *no* PDB in  $\mathcal{G}$  admits a **sjfCQ(BID)**-presentation.

**Lemma 7.10.** *If  $\mathcal{D} \in \mathcal{G}$ , then  $\mathcal{D} \notin \text{sjfCQ(BID)}$ .*

The idea behind the proof of this lemma is as follows. We start by assuming that we have a **sjfCQ(BID)**-representation of some  $\mathcal{D} \in \mathcal{G}$ . The property we exploit is that facts in the underlying **BID**-PDB can only be either independent or mutually exclusive. We pick a particularly chosen instance from the **BID**-PDB and show that we can modify it in such a way that the resulting instance still has positive probability, but its image necessarily contains exactly one of the edges  $E(a, b)$  or  $E(b, a)$  for some  $(a, b)$ . This then contradicts the assumption that we started with a representation of a PDB from  $\mathcal{G}$ .

*Proof of Theorem 7.10.* Let  $\mathcal{D} = (\mathbb{D}, P_{\mathcal{D}})$  be a probabilistic graph in  $\mathcal{G}$  over  $\mathbb{U}$ . Assume that there exists a **BID**-PDB  $\mathcal{I} = (\mathbb{I}, P_{\mathcal{I}})$  and a self-join free **CQ**-view of the following shape

$$\Phi(x, y) := \exists z_1 \dots \exists z_m : R_1(t_1) \wedge \dots \wedge R_n(t_n)$$

such that  $\mathcal{D} = \Phi(\mathcal{I})$ . This can be assumed since equality atoms can be removed using substitution and, as the PDBs in  $\mathcal{G}$  are simple and undirected,  $\Phi$  can not be of the shape  $\Phi(x, x)$  or  $\Phi(x, a)$ . Without loss of generality, we can assume that all the variables  $z_1, \dots, z_m$  appear among  $R_1(t_1) \wedge \dots \wedge R_n(t_n)$ . In the following, we let

$$f_i[a; b; \mathbf{c}] := R_i(t_i[x/a; y/b; z_1/c_1; \dots; z_m/c_m])$$

for all  $a, b \in \mathbb{U}$  and all  $\mathbf{c} = (c_1, \dots, c_m) \in \mathbb{U}^m$ .

Observe that for all  $I \in \mathbb{I}$  it holds that

$$I \models \Phi(a, b) \iff \text{there exists } \mathbf{c} \text{ such that for all } i = 1, \dots, n \text{ it holds that } f_i[a; b; \mathbf{c}] \in I; \quad (15)$$

and

$$I \models \Phi(b, a) \iff \text{there exists } \mathbf{c}' \text{ such that for all } i = 1, \dots, n \text{ it holds that } f_i[b; a; \mathbf{c}'] \in I. \quad (16)$$

From now on, let  $i \in \{1, \dots, n\}$  be fixed such that  $R_i(t_i)$  contains at least one of the variables  $x$  or  $y$ . Without loss of generality, we assume that  $R_i(t_i)$  contains  $x$  (the argument for  $y$  follows analogously).

Let  $(a, b), (a', b') \in \mathbb{U}^2$ . We say that a block  $B$  from  $\mathcal{I}$  is *full*, if the marginal probabilities of the facts in  $B$  add up to 1. We say that  $B$  is  $(a, b)$ -*saturated* if  $B$  is a full block such that every fact of  $B$  is of the shape  $f_i[a; b; \mathbf{c}]$  for some  $\mathbf{c} \in \mathbb{U}^m$ . Note that if  $(a, b)$  and  $(a', b')$  have  $a \neq a'$ , then no block  $B$  is *both*  $(a, b)$ - and  $(a', b')$ -saturated, since  $f_i$  contains the variable  $x$  which yields  $f_i[a; b; \mathbf{c}] \neq f_i[a'; b'; \mathbf{c}']$  for all  $\mathbf{c}, \mathbf{c}' \in \mathbb{U}^m$ .

Since  $\mathcal{D}$  is unbounded, there exists a family  $T$  of infinitely many pairs  $(a, b)$  with pairwise different values of  $a$  such that  $E(a, b)$  appears among the instances of  $\text{worlds}(\mathcal{D})$ . Since  $\mathcal{I}$  contains at most finitely many full blocks, by the pigeonhole-principle, there exists a pair  $(a^*, b^*) \in T$  with  $a^* \neq b^*$  such that no block of  $B$  is  $(a^*, b^*)$ -saturated.

Let  $D$  be one of the instances in  $\text{worlds}(\mathcal{D})$  that contain both  $E(a^*, b^*)$  and  $E(b^*, a^*)$ , and let  $I \in \text{worlds}(\mathcal{I})$  be an arbitrary preimage of  $D$  under  $\Phi$ . Then both  $I \models \Phi(a^*, b^*)$  and  $I \models \Phi(b^*, a^*)$ . In particular,  $I$  contains facts  $f_i[a^*; b^*; \mathbf{c}]$  and  $f_i[b^*; a^*; \mathbf{c}']$  for some (possibly multiple)  $\mathbf{c}, \mathbf{c}' \in \mathbb{U}^m$ .

Let  $F(a^*, b^*) := \{f_i[a^*; b^*; \mathbf{c}] : \mathbf{c} \in \mathbb{U}^m\}$ . We can partition  $I$  into

$$I = I' \cup I_{=1}(a^*, b^*) \cup I_{<1}(a^*, b^*) \quad (17)$$

such that

- $I' := I \setminus F(a^*, b^*)$ ,
- $I_{=1}(a^*, b^*) := I \cap F(a^*, b^*) \cap \bigcap_{B \text{ full block}} B$ , and
- $I_{<1}(a^*, b^*) := I \cap F(a^*, b^*) \cap \bigcap_{B \text{ not full}} B$ .

Recall that no block of  $I$  is  $(a^*, b^*)$ -saturated. Hence, if  $f \in I_{=1}(a^*, b^*)$ , then there exists  $f'$  in the same block as  $f$  such that  $f' \notin F(a^*, b^*)$ . We simply drop all facts from  $I_{<1}(a^*, b^*)$  and transform  $I_{=1}(a^*, b^*)$  into a new set of facts  $I''$  by replacing every fact  $f$  in  $I_{=1}(a^*, b^*)$  with such an  $f'$ . This yields a new set of facts

$$J := I' \cup I'' \in \text{worlds}(\mathcal{I}),$$

Now,  $J \not\models \Phi(a^*, b^*)$ , since  $\Phi$  is self-join free and  $J$  does not contain any of the facts  $F(a^*, b^*)$ . However,  $I \cap F(b^*, a^*) \subseteq J \cap F(b^*, a^*)$  (where  $F(b^*, a^*)$  is defined analogously to  $F(a^*, b^*)$ ), since  $F(a^*, b^*) \cap F(b^*, a^*) = \emptyset$ . This latter disjointness stems from the fact that  $f_i$  contains the variable  $x$  and the property that  $\Phi$  is self-join free. Therefore, using the monotonicity of  $\Phi$ , it follows from (16) that  $J \models \Phi(b^*, a^*)$ .

Together we have that  $(b^*, a^*) \in \Phi(J)$  but  $(a^*, b^*) \notin \Phi(J)$ , contradicting  $\Phi(J) \in \text{worlds}(\mathcal{D})$ .  $\square$

**Remark 7.11.** The above proof is similar to the discussions of Section 6 in the sense that it only concerns the sets of possible worlds and is, apart from this, oblivious to probabilities.

Note that as **sjfCQ(BID)** contains **sjfCQ(TI)**, Theorem 7.10 in particular shows that **sjfCQ(TI)** contains none of the PDBs from  $\mathcal{G}$ . Recall that, on the contrary, **UCQ(TI)** contains all the PDBs from  $\mathcal{G}$ . Naturally, the class **CQ(TI)** should be investigated next. In fact, when it comes to the representation of PDBs from  $\mathcal{G}$ , the class **CQ(TI)** proves to constitute a middle ground between the previous two extreme cases.

**Lemma 7.12.** *Let  $\mathcal{D}_1$  be a probabilistic graph in  $\mathcal{G}$  with edge set  $E = \{a_i, b_i\}_{i \in \mathbb{N}}$  where the marginal probabilities of the edges  $E(a_i, b_i)$  are given by  $p_{a_i, b_i} = p_{b_i, a_i} = \frac{1}{i^4}$ . Then  $\mathcal{D}_1 \in \mathbf{CQ(TI)}$ .*

*Proof.* First, note that  $\sum_{i=1}^{\infty} \frac{1}{i^4} < \infty$ , so  $\mathcal{D}_1$  is well-defined. We will now construct a **TI**-PDB  $\mathcal{I}$  and a **CQ**  $\Phi$  such that  $\mathcal{D}_1 = \Phi(\mathcal{I})$ . For this, we define  $\mathcal{I}$  to be the **TI**-PDB over the fact set  $\{E(a_i, b_i) : i = 1, 2, \dots\} \cup \{E(b_i, a_i) : i = 1, 2, \dots\}$  such that the marginal probability of  $E(a_i, b_i)$  and  $E(b_i, a_i)$ , respectively, is given as  $\frac{1}{i^2}$ . As  $\sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$ , this makes  $\mathcal{I}$  a well-defined PDB. Consider the **CQ**

$$\Phi(x, y) = E(x, y) \wedge E(y, x).$$

We claim that  $\Phi(\mathcal{I}) = \mathcal{D}_1$ . For this, it suffices to establish that  $\Phi(\mathcal{I})$  has the correct marginal probabilities and that  $\Phi(\mathcal{I}) \in \mathcal{G}$ . For all  $I \in \text{worlds}(\mathcal{I})$  and all  $i = 1, 2, \dots$ , it holds that  $I \models \Phi(a_i, b_i)$  if and only if  $I$  contains both  $E(a_i, b_i)$  and  $E(b_i, a_i)$ . The probability of this event is precisely  $\frac{1}{i^2} \cdot \frac{1}{i^2} = \frac{1}{i^4}$ , which is the marginal probability of  $E(a_i, b_i)$  in  $\mathcal{D}_1$ .

It remains to show that  $\Phi(\mathcal{I}) \in \mathcal{G}$ . It is clear that every possible world of  $\Phi(\mathcal{I})$  is an undirected and simple graph database. Because the presence of any double edge  $\{E(a_i, b_i), E(b_i, a_i)\}$  in  $\Phi(\mathcal{I})$  is determined solely by the presence of the facts  $E(a_i, b_i)$  and  $E(b_i, a_i)$  in  $\mathcal{I}$ , the PDB  $\Phi(\mathcal{I})$  is also edge-independent. Moreover,  $\Phi(\mathcal{I})$  is unbounded, since there are infinitely many facts of positive marginal probability and the edges are independent. Together,  $\Phi(\mathcal{I}) = \mathcal{D}_1$ .  $\square$

Our current status is that **UCQ(TI)** contains *all* PDBs from  $\mathcal{G}$ , that **sjfCQ(TI)** contains *no* PDB from  $\mathcal{G}$ , and that **CQ(TI)** contains *some* PDBs of  $\mathcal{G}$ . Our next lemma yields, in particular,



that **CQ(TI)** is strictly weaker than **UCQ(TI)** by giving an example of a PDB in  $\mathcal{G}$  that is not contained in **CQ(BID)**.

**Lemma 7.13.** *Let  $\mathcal{D}_2$  be the probabilistic graph representing a countably infinite set of pairwise disjoint edges, where the  $i$ -th edge  $\{a_i, b_i\}$  has probability  $p_{a_i, b_i} = p_{b_i, a_i} = \frac{1}{i^2}$ . Then  $\mathcal{D}_2 \notin \mathbf{CQ}(\mathbf{BID})$ .*

The structure of this counterexample is strikingly similar to the  $\mathcal{G}$ -PDB we investigated in Theorem 7.12. We comment on this after the proof.

*Proof of Theorem 7.13.* We assume otherwise and let  $\mathcal{I}$  be a **BID**-PDB and  $\Phi$  a **CQ**-view

$$\Phi(x, y) = \exists z_1 \dots \exists z_m: R_1(t_1) \wedge \dots \wedge R_n(t_n)$$

such that  $\mathcal{D}_2 = \Phi(\mathcal{I})$ . That we can assume  $\Phi$  to be of the above shape has the same reason as in the proof of Theorem 7.10. We also reuse the notation from there and let  $f_j[a; b; c]$  denote the fact originating from  $R_j(t_j)$  by replacing  $x, y, z$  with  $a, b$  and  $c$ , respectively, for  $j = 1, \dots, n$ . Note that the equivalences from (15) and (16) still hold. Again, we pick an index  $j \in \{1, \dots, n\}$  such that  $x$  or  $y$  or both appear in  $R_j(t_j)$ .

Let

$$F(a, b) := \{f_j[a; b; c] : c \in \mathbb{U}^m\}$$

for all  $a, b \in \mathbb{U}$ . Because  $\mathcal{I}$  is a **BID**-PDB, for all sets  $F$  of facts (in particular, for the sets  $F(a, b)$ ), it holds that

$$\Pr_{I \sim \mathcal{I}} (I \cap F = \emptyset) = \prod_{B \text{ block of } \mathcal{I}} \left(1 - \sum_{f \in B \cap F} p_f\right).$$

Let  $i \in \mathbb{N}_+$  be arbitrary. Since  $f_j$  contains  $x$  or  $y$ , we have  $F(a_i, b_i) \cap F(b_i, a_i) = \emptyset$ . Moreover, for any distinct  $i_1, i_2 \in \mathbb{N}_+$ , it holds that

$$\left(F(a_{i_1}, b_{i_1}) \cup F(b_{i_1}, a_{i_1})\right) \cap \left(F(a_{i_2}, b_{i_2}) \cup F(b_{i_2}, a_{i_2})\right) = \emptyset, \quad (18)$$

since the edges of  $\mathcal{D}_2$  are pairwise disjoint. We now want to show that  $\sum_{f \in F(a_i, b_i) \cup F(b_i, a_i)} p_f \geq \frac{1}{i}$  which, together with (18), leads to

$$\sum_{f \in \text{facts}(\mathcal{I})} p_f \geq \sum_{i=1}^{\infty} \sum_{f \in F(a_i, b_i) \cup F(b_i, a_i)} p_f \geq \sum_{i=1}^{\infty} \frac{1}{i} = \infty.$$

This is a contradiction to the assumption that  $\mathcal{I}$  is a well-defined **BID**-PDB via Theorem 2.7.

To prove  $\sum_{f \in F(a_i, b_i) \cup F(b_i, a_i)} p_f \geq \frac{1}{i}$ , we first observe that

$$\begin{aligned} \frac{1}{i^2} &= \Pr_{D \sim \mathcal{D}} (E(a_i, b_i) \in D \text{ and } E(b_i, a_i) \in D) = \Pr_{I \sim \mathcal{I}} (I \models \Phi(a_i, b_i) \text{ and } I \models \Phi(b_i, a_i)) \\ &\leq \Pr_{I \sim \mathcal{I}} (I \cap F(a_i, b_i) \neq \emptyset \text{ and } I \cap F(b_i, a_i) \neq \emptyset) \end{aligned}$$

Suppose that the blocks of  $\mathcal{I}$  are  $B_1, B_2, \dots$ . For  $k = 1, 2, \dots$ , we define events  $X_k, Y_k$  as follows:

$$X_k := \{I \in \text{worlds}(\mathcal{I}) : I \cap B_k \cap F(a_i, b_i) \neq \emptyset\} \quad \text{and} \quad Y_k := \{I \in \text{worlds}(\mathcal{I}) : I \cap B_k \cap F(b_i, a_i) \neq \emptyset\}.$$

Because  $\mathcal{I}$  is **BID**, all events  $X_k$  are independent, and so are all  $Y_k$ . Moreover, the two events  $X_k$  and  $Y_k$  are mutually exclusive for all  $k$ . We let  $x_k$  and  $y_k$  denote the probabilities of  $X_k$  and  $Y_k$  in  $\mathcal{I}$ , respectively. As the events  $X_k$ , respectively  $Y_k$ , only concern facts in the block  $B_k$ , we have

$$x_k = \sum_{f \in F(a_i, b_i) \cap B_k} p_f \quad \text{and} \quad y_k = \sum_{f \in F(b_i, a_i) \cap B_k} p_f.$$

Now, since

$$\sum_{f \in F(a_i, b_i) \cup F(b_i, a_i)} p_f = \sum_{k=1}^{\infty} x_k + \sum_{k=1}^{\infty} y_k,$$

it remains to show that  $\sum_{k=1}^{\infty} x_k + \sum_{k=1}^{\infty} y_k \geq \frac{1}{i}$ . For this, we proceed using the inclusion-exclusion-formula to get

$$\begin{aligned} & \Pr_{I \sim \mathcal{I}} (I \cap F(a_i, b_i) \neq \emptyset \text{ and } I \cap F(b_i, a_i) \neq \emptyset) \\ &= 1 - \Pr_{I \sim \mathcal{I}} (I \cap F(a_i, b_i) = \emptyset \text{ or } I \cap F(b_i, a_i) = \emptyset) \\ &= 1 - \Pr_{I \sim \mathcal{I}} (I \cap F(a_i, b_i) = \emptyset) - \Pr_{I \sim \mathcal{I}} (I \cap F(b_i, a_i) = \emptyset) + \Pr_{I \sim \mathcal{I}} (I \cap F(a_i, b_i) = I \cap F(b_i, a_i) = \emptyset) \\ &= 1 - \prod_{k=1}^{\infty} (1 - x_k) - \prod_{k=1}^{\infty} (1 - y_k) + \prod_{k=1}^{\infty} (1 - x_k - y_k) \\ &\leq 1 - \prod_{k=1}^{\infty} (1 - x_k) - \prod_{k=1}^{\infty} (1 - y_k) + \prod_{k=1}^{\infty} (1 - x_k - y_k + x_k y_k) \\ &= 1 - \prod_{k=1}^{\infty} (1 - x_k) - \prod_{k=1}^{\infty} (1 - y_k) + \prod_{k=1}^{\infty} (1 - x_k) \prod_{k=1}^{\infty} (1 - y_k) \\ &= \left(1 - \prod_{k=1}^{\infty} (1 - x_k)\right) \cdot \left(1 - \prod_{k=1}^{\infty} (1 - y_k)\right). \end{aligned}$$

There are two cases. If one of  $\sum_{k=1}^{\infty} x_k$  or  $\sum_{k=1}^{\infty} y_k$  is larger than or equal to 1, then we are done. In the other case, when both sums are smaller than 1, we can apply a Weierstrass inequality (cf. [13, Chapter VI]) to get

$$\left(1 - \prod_{k=1}^{\infty} (1 - x_k)\right) \cdot \left(1 - \prod_{k=1}^{\infty} (1 - y_k)\right) \leq \left(\sum_{k=1}^{\infty} x_k\right) \cdot \left(\sum_{k=1}^{\infty} y_k\right).$$

Together with the inequalities from before, we get

$$\frac{1}{i^2} \leq \left(\sum_{k=1}^{\infty} x_k\right) \cdot \left(\sum_{k=1}^{\infty} y_k\right). \quad (19)$$

Therefore, one of  $\sum_{k=1}^{\infty} x_k$  and  $\sum_{k=1}^{\infty} y_k$  is at least  $\frac{1}{i}$ , so in particular

$$\sum_{k=1}^{\infty} x_k + \sum_{k=1}^{\infty} y_k \geq \frac{1}{i}. \quad \square$$

Let us now reconsider Theorems 7.12 and 7.13 in direct comparison. First, we note that, the view from the proof of Theorem 7.12, given Theorem 7.10 *had* to make use of a self-join. Looking deeper into the proof, we can see that this self-join essentially brought about a squaring of the marginal probabilities when applying the view to the underlying **TI**-PDB. This is why the marginal probabilities in Theorem 7.12 were chosen to be  $\frac{1}{i^4}$ , since taking the root leaves us with  $\frac{1}{i^2}$ , leading to a convergent series. Note that instead, the marginal probabilities of the edges in the example from Theorem 7.13 were chosen to be  $\frac{1}{i^2}$ , where taking the root leaves us with  $\frac{1}{i}$ , leading to a divergent series. At the beginning of the proof, it is not clear why we would end up with the consideration of this root. However, our arguments led us to the related situation of (19). All in all, the salient point is that in Theorem 7.12 the series of marginal probabilities converges fast enough for our argument to go through, whereas the series from Theorem 7.13 converges too slowly to admit a representation in the desired class.

To complete the picture, we still need to determine whether the classes **UCQ(BID)** and **FO(BID)** are separated or equal. This is done in the next proposition.

**Proposition 7.14.** **UCQ(BID)  $\subsetneq$  FO(BID).**

*Proof.* We show that there exists a **FO(TI)**-PDB  $\mathcal{D}$  such that  $\mathcal{D} \notin \mathbf{UCQ(BID)}$ . Consider the database schema with a single unary relation  $R$  over a countably infinite domain  $\mathbb{U}$ . We let  $\mathbb{D} = \{D_1, D_2, \dots\}$  be a set of pairwise disjoint fact sets over this schema such that for all  $n \in \mathbb{N}_+$ , the set  $D_n$  has size  $|D_n| \geq n$ . An easy example is  $D_n = \{R(a_i) : i = n^2, n^2+1, \dots, (n+1)^2-1\}$ . We know from Theorem 6.8, that there exists a PDB  $\mathcal{D} \in \mathbf{FO(TI)} = \mathbf{FO(BID)}$  with  $\text{worlds}(\mathcal{D}) = \mathbb{D}$ .

We now show that  $\mathcal{D} \notin \mathbf{UCQ(BID)}$ . Assume otherwise and let  $\mathcal{I}$  be a **BID**-PDB and  $\Phi$  be a **UCQ**-view such that  $\mathcal{D} = \Phi(\mathcal{I})$ . Recall that a full block of  $\mathcal{I}$  is a block in which all marginal probabilities sum to 1. We call an instance  $I$  from  $\mathcal{I}$  *minimal*, if it only contains facts from full blocks. Note that if  $I$  is an instance of positive probability in  $\mathcal{I}$ , then it is a superset of a minimal instance, and that all minimal instances themselves have positive probability. Moreover, all minimal instances have the same cardinality, say  $m$ , which is equal to the number of full blocks in  $\mathcal{I}$ . As  $\Phi$  is (in particular) first-order,  $|\Phi(I)| \leq q(|I|)$  for some fixed polynomial  $q$  (we saw that in the proof of Theorem 3.5). Therefore,  $|\Phi(I)| \leq q(m)$  for all minimal instances  $I$ . Let  $n > q(m)$  and pick an arbitrary instance  $I$  from the preimage of  $D_n$  under  $\Phi$ . Let  $I'$  be the restriction of  $I$  to the facts from the full blocks of  $\mathcal{I}$ . Then  $I'$  is a minimal instance with  $I' \subseteq I$ . Since  $\Phi$  is monotone, it holds that  $\Phi(I') \subseteq \Phi(I) = D_n$ . Note that  $\Phi(I') \neq \emptyset$ , since  $I'$  has positive probability, but  $\mathcal{D}$  does not contain the empty instance. Since  $\mathcal{D}$  is domain-disjoint, this implies  $\Phi(I') = \Phi(I) = D_n$ . Then, however,  $|\Phi(I')| \leq q(m) < n \leq |D_n|$  contradicts  $\Phi(I') = D_n$ .  $\square$

Combining all the previous results from this subsection, we are now finally ready to state the main result of our investigation of the relative expressive power of views over PDBs.

**Theorem 7.15.** *The classes of views of probabilistic databases have exactly the relative expressive power as shown in the Hasse diagram of Figure 5a.*

*Proof.* Consider the diagram shown in Figure 5b. The thick lines indicate the trivial inclusions that arise from **TI**  $\subseteq$  **BID**  $\subseteq$  **PDB** and  $\{\text{Id}\} \subseteq \mathbf{sjfCQ} \subseteq \mathbf{CQ} \subseteq \mathbf{UCQ} \subseteq \mathbf{FO}$  where  $\text{Id}$  denotes

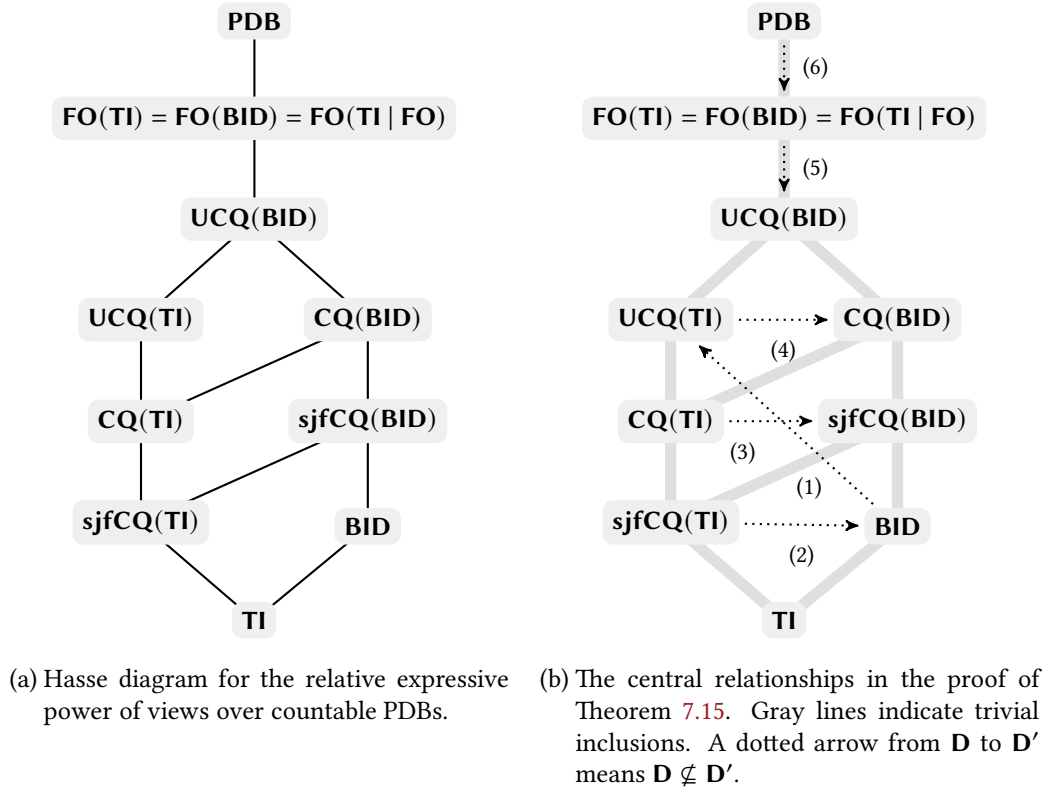


Figure 5: Diagrams for the proof of Theorem 7.15.

the identity view. It remains to show that these are proper inclusions and that no other inclusions hold. For this purpose, consider the dotted arrows. Once they are shown, all remaining relationships are entailed by transitivity as follows:

- If  $\mathbf{D}_1 \not\subseteq \mathbf{D}_2$  and  $\mathbf{D}'_2 \subseteq \mathbf{D}_2$ , then  $\mathbf{D}_1 \not\subseteq \mathbf{D}'_2$ .
- If  $\mathbf{D}_1 \not\subseteq \mathbf{D}_2$  and  $\mathbf{D}_1 \subseteq \mathbf{D}'_1$ , then  $\mathbf{D}'_1 \not\subseteq \mathbf{D}_2$ .

The relationships of the dotted arrows follow from previous results as follows.

1. This follows from Theorem 6.5.
2. This is witnessed by Theorem 7.2.
3. Theorem 7.10 states that no PDB from  $\mathcal{G}$  are in  $\mathbf{sjfCQ}(\mathbf{BID})$ . However, Theorem 7.12 presents an example of a PDB from  $\mathcal{G}$  that is in  $\mathbf{CQ}(\mathbf{TI})$ .
4. Theorem 7.9 states that all PDBs from  $\mathcal{G}$  are in  $\mathbf{UCQ}(\mathbf{TI})$ , yet Theorem 7.13 gives an example of a PDB from  $\mathcal{G}$  that is not in  $\mathbf{CQ}(\mathbf{BID})$ .
5. See Theorem 7.14.

6. See Theorem 3.1 (which is a restatement of [31, Proposition 4.9]).  $\square$

Concluding this section, we point out that apart from Theorems 7.12 and 7.13 the arguments we found for the separations are of a purely logical nature. In Theorems 7.12 and 7.13, however, the argument depended on our choice of probabilities that allowed us to find a well-defined representation in the first case, and enabled us to rule out the existence of such in the second case. It is not clear whether this separation can also be obtained using an argument that only involves the structure of the possible worlds rather than their probabilities.

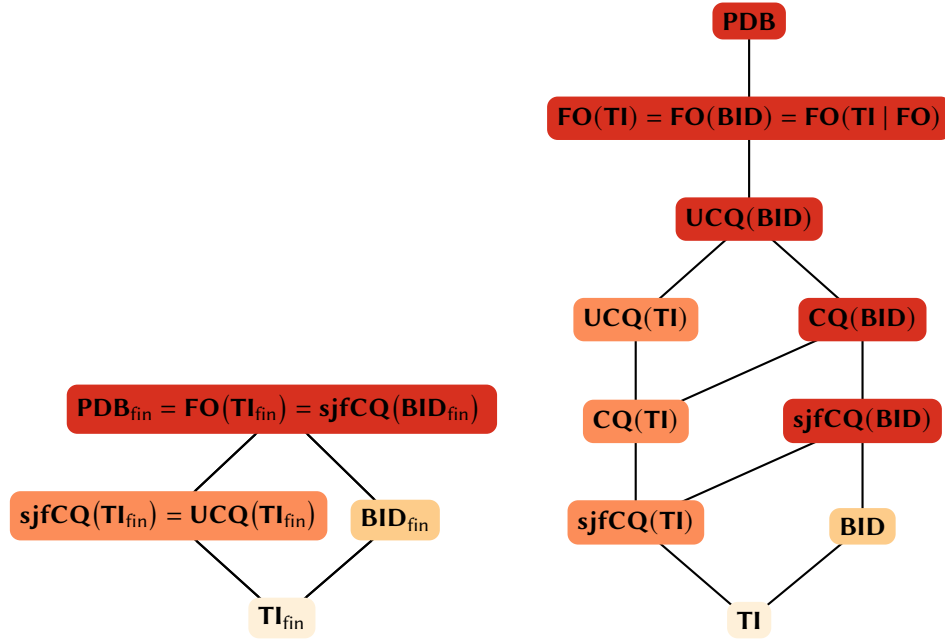
## 8 Concluding Remarks

This article initiates research on representations of infinite probabilistic databases. Our focus is on studying representability by first-order views over tuple-independent PDBs. Although it is known that  $\mathbf{FO}(\mathbf{TI}) \subsetneq \mathbf{PDB}$ , we show that  $\mathbf{FO}(\mathbf{TI})$  is quite robust, as  $\mathbf{FO}(\mathbf{TI}) = \mathbf{FO}(\mathbf{BID}) = \mathbf{FO}(\mathbf{TI} \mid \mathbf{FO})$ , and it can represent any bounded size PDB. In addition, while the finite moments property is necessary for membership in  $\mathbf{FO}(\mathbf{TI})$ , we show that it is not sufficient. We also study how restricting the expressive power of the views (using fragments of first-order logic) affects the PDBs that can be represented. We show that the different classes obtained using **sjfCQ**-, **CQ**- and **UCQ**-views over **TI**- and **BID**-PDBs are distinct (even though many of them collapse for finite PDBs). We establish the concrete relationships between these classes, as shown in Figure 6. In our analysis, we pay attention to the causes for (non-)representability. These can be twofold in general: arithmetical or purely logical. The class  $\mathbf{FO}(\mathbf{TI})$ , however, (unlike some of its fragments) eludes purely logical arguments for non-representability. When considering logical reasons, the mentioned relationships to just sets of possible worlds (in the context of incomplete databases) extend existing work to infinite PDBs [10, 28].

Our characterization of  $\mathbf{FO}(\mathbf{TI})$  is only partial, as there is a gap between our necessary and sufficient conditions for membership. Thus, a natural direction for future work is obtaining a complete characterization of this class or the other classes of PDBs that appeared in this paper. This is especially challenging since, even in the finite setting, we know of no algorithm to decide whether a given finite PDB is a  $\mathbf{UCQ}(\mathbf{TI}_{\text{fin}})$ -PDB. Regarding the causes for (non-)representability, most of our separations have been purely logical. It remains open whether such a separation of  $\mathbf{CQ}(\mathbf{TI})$  and  $\mathbf{UCQ}(\mathbf{TI})$  exists. Further possible extensions of our work include conditioning with respect to fragments of  $\mathbf{FO}$  or using first-order views with inbuilt relations (e.g., over ordered universes). Further research could also investigate which of our results can be extended to uncountable PDBs. Other interesting insights might be obtained by relating our work to recent notions of measuring uncertainty in incomplete databases [45, 17]. Regarding query evaluation in infinite PDBs, beyond preliminary results on **TI**-PDBs [31], an abstract investigation remains open.

## 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.



(a) Hasse diagram for the relative expressive power of views over finite PDBs. (b) Hasse diagram for the relative expressive power of views over countable PDBs.

Figure 6: Hasse diagrams of PDB classes with independence assumptions in the finite and the countable setting. Edges and their upwards closure mean proper inclusion. All remaining pairs of classes are incomparable. Areas of the same shade indicate which classes from the countable setting collapse in the finite setting.

## 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] Serge Abiteboul, Paris Kanellakis, and Gösta Grahne. On the Representation and Querying of Sets of Possible Worlds. *Theoretical Computer Science*, 78(1):159–187, 1991. doi:10.1016/0304-3975(51)90007-2.
- [4] 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.

- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] Philip J. Boland. A Biographical Glimpse of William Sealy Gosset. *The American Statistician*, 38(3):179–183, 1984. URL: <http://www.jstor.org/stable/2683648>.
- [12] 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.
- [13] T. J. I’A. Bromwich. *An Introduction to the Theory of Infinite Series*. Macmillan and Company Ltd., London, United Kingdom, 2nd, revised edition, 1926.
- [14] 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.
- [15] Nofar Carmeli, Martin Grohe, Peter Lindner, and Christoph Standke. Tuple-Independent Representations of Infinite Probabilistic Databases. In *PODS’21: Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Virtual Event, China, June 20-25, 2021*, pages 388–401. ACM, 2021. doi:10.1145/3452021.3458315.



- [16] 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.
- [17] 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.
- [18] Nilesch 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.
- [19] Nilesch 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.
- [20] Nilesch 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.
- [21] Nilesch 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Alan Frieze and Michal Karonski. *Introduction to Random Graphs*. Cambridge University Press, Cambridge, UK, 2016. doi:10.1017/CB09781316339831.
  - [28] 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.
  - [29] 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.
  - [30] 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.
  - [31] 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 \[cs.DB\]](https://arxiv.org/abs/1807.00607). doi:10.1145/3294052.3319681.
  - [32] Martin Grohe and Peter Lindner. Independence in Infinite Probabilistic Databases, 2021. To appear. URL: <https://arxiv.org/abs/2011.00096>.
  - [33] Martin Grohe and Peter Lindner. Infinite Probabilistic Databases. *Logical Methods in Computer Science*, Volume 18, Issue 1, 2022. URL: <https://lmcs.episciences.org/9137>, doi:10.46298/lmcs-18(1:34)2022.
  - [34] 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.
  - [35] Tomasz Imieliński and Witold Lipski. Incomplete Information in Relational Databases. *Journal of the ACM*, 31(4):761–791, 1984. doi:10.1145/1634.1886.
  - [36] 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.
  - [37] Abhay Jha and Dan Suciu. Probabilistic Databases with MarkoViews. *Proceedings of the VLDB Endowment*, 5(11):1160–1171, 2012. doi:10.14778/2350229.2350236.
  - [38] 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.
- [39] 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.
  - [40] 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.
  - [41] Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer-Verlag London, London, UK, 2nd edition, 2014. Translation from the German language edition. doi:10.1007/978-1-4471-5361-0.
  - [42] Christoph Koch and Dan Olteanu. Conditioning Probabilistic Databases. *Proceedings of the VLDB Endowment*, 1(1):313–325, 2008. doi:10.14778/1453856.1453894.
  - [43] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models—Principles and Techniques*. MIT Press, 2009.
  - [44] 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.
  - [45] 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.
  - [46] 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.
  - [47] 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.
  - [48] Christopher Ré. Applications of Probabilistic Constraints. Technical Report TR2007-03-03, University of Washington, Seattle, WA, USA, 2007.
  - [49] 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.

- [50] 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.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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:10.1561/19000000052.
- [58] 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.
- [59] 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.