Journal of Computer and System Sciences 117 (2021) 154–164

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Counting subset repairs with functional dependencies

Ester Livshits [a,*], Benny Kimelfeld [a], Jef Wijsen [b]

[a] Technion, Haifa, Israel
[b] University of Mons, Belgium

ARTICLE INFO

Article history:
Received 10 July 2019
Received in revised form 9 September 2020
Accepted 28 October 2020
Available online 26 November 2020

Keywords:
Inconsistent databases
Repair
Subset repair
Repair counting
Functional dependencies
Conflict graph

ABSTRACT

We study the problem of counting the repairs of an inconsistent database in the case where constraints are Functional Dependencies (FDs). A repair is then a maximal independent set of the conflict graph, wherein nodes represent facts and edges represent violations. We establish a dichotomy in data complexity for the complete space of FDs: when the FD set has, up to equivalence, what we call a "left-hand-side chain," the repairs can be counted in polynomial time; otherwise, the problem is ♯**P**-complete. Moreover, the property of having a left-hand-side chain up to equivalence coincides with the condition that the conflict graph of every inconsistent database for the schema is P_4-free, and it is polynomial-time decidable.

1. Introduction

Database inconsistency arises for different reasons and in different applications. For example, in common applications of Big Data, information is obtained from imprecise sources (e.g., social encyclopedias, social networks, and sensors attached to appliances) via imprecise procedures (e.g., natural-language and signal processing). It may also arise when integrating conflicting data from different (possibly consistent) sources. Arenas, Bertossi and Chomicki [1] introduced a principled approach to managing inconsistency, via the notions of *repairs* and *consistent query answering*. Informally, a *repair* of an inconsistent database I is a consistent database J that differs from I in a "minimal" way, where *minimality* refers to the *symmetric difference*. In the case of anti-monotonic integrity constraints (e.g., functional dependencies), a repair is an inclusion-maximal consistent subinstance (not properly contained in any consistent subinstance), and is referred to as a *subset repair* [2].

Various computational problems around database repairs have been extensively investigated [3]. Most studied is the problem of computing the *consistent answers* of a query on an inconsistent database; these are the tuples that are obtained in every possible repair [1,4]. Another well-studied question is that of *repair checking* [2]: given instances I and J, determine whether J is a repair of I. Depending on the type of repairs and the type of integrity constraints, these problems may vary from tractable to highly intractable complexity classes.

In this work, we study the complexity of computing the number of subset repairs, when the constraints are Functional Dependencies (FDs). Although FDs have been studied for almost five decades and database repairs have been studied for over two decades, not much is known about this problem. Maslowski and Wijsen [5,6] and later Calautti et al. [7] have studied the problem of counting repairs that satisfy a Boolean Conjunctive Query (CQ), in the case of key constraints. They established dichotomy results, classifying CQs into those where counting can be done in polynomial time, and those where

* Corresponding author.
E-mail addresses: esterliv@cs.technion.ac.il (E. Livshits), bennyk@cs.technion.ac.il (B. Kimelfeld), jef.wijsen@umons.ac.be (J. Wijsen).

https://doi.org/10.1016/j.jcss.2020.10.001
0022-0000/

counting is ♯**P**-complete.[1] The challenge in their work is due to the query. If we ignore the query and simply count the repairs, then we get the straightforward case of counting subset repairs under key constraints.

Another fundamental problem where repair counting arises is that of *measuring database inconsistency*, which has been studied extensively by the Knowledge Representation (KR) and Logic communities [8–13], and has been recently acknowledged by the database community [14–16]. Inconsistency measures can be used for estimating the extent to which a database is trustworthy, and the effort required to clean it. One of the well studied measures is the number of repairs. This measure is sometimes denoted as I_{M} [11,12] and sometimes as I_{mc} [13]. In particular, under this measure, the problem we study is that of estimating the level of inconsistency of the database.

We study the *data complexity* of repair counting, where the schema, which is comprised of the relational signature and the set of FDs, is considered fixed, and every schema defines a separate computational problem. In particular, the complexity of repair counting may be different for different schemas. As we explain below, it suffices to consider single-relation schemas; our results generalize to multi-relation schemas in a straightforward manner.

The main result of this manuscript is a dichotomy in data complexity that classifies FD sets into those for which the number of subset repairs can be computed in polynomial time, and those for which the problem is ♯**P**-complete. In particular, we introduce the definition of a *left-hand-side chain* (or *lhs chain*, for short) that captures *precisely* the tractable cases of counting subset repairs. We say that a set of FDs has an lhs chain if the FDs in the set can be arranged in an order such that the left-hand side of each FD is contained in the left-hand side of every FD that appears later in the order. Our dichotomy is as follows. If the set of FDs is equivalent to an FD set with an lhs chain, then the repairs can be counted in polynomial time (and, in fact, even under combined complexity, where both the schema and the database are given as input). Conversely, if the set of FDs is not equivalent to any FD set with an lhs chain, then the problem is ♯**P**-complete. We also show that if an FD set is equivalent to an FD set with an lhs chain, then it has a single minimal cover, and this minimal cover has an lhs chain; hence, it is decidable in polynomial time to which side of the dichotomy a given schema belongs.

The dichotomy easily generalizes to multi-relation schemas, since we consider only FDs and, hence, conflicts are always within the same relation. In particular, the problem is solvable in polynomial time if its restriction to every single relation is solvable in polynomial time (and then the number of repairs is the product of the number of repairs of each relation), and is ♯**P**-complete otherwise. Hence, in the remainder of the manuscript, we continue with the assumption of a single relation.

Observe that repair counting is the same as the problem of counting the maximal independent sets of the *conflict graph*, which is the graph that has the facts of the database as nodes and an edge between every inconsistent pair of facts. Counting the maximal independent sets of a graph is ♯**P**-complete [17]. Special tractable cases include the class of P_4*-free* graphs [18] (also known as *complement reducible graphs* or *cographs*); that is, the graphs that do not have any induced subgraph that is a simple path of length four. We prove that the property of being equivalent to an FD schema with an lhs chain coincides with the property that every conflict graph over the schema is P_4-free. This explains the tractability side of our dichotomy. In fact, our dichotomy implies that when a set of FDs allows for polynomial-time repair counting, it is *precisely due to the tractability of counting independent sets of cographs*.

This manuscript contains the full version of a result published in a conference publication of the authors [19]. We have added in this manuscript all the proofs and intermediate results that were excluded from the conference paper. In particular, Sections 4, 5, and 6 are new and contain the full proof of our main result—the dichotomy in the complexity of counting subset repairs (Theorem 3.2).

The rest of the manuscript is organized as follows. In the next section, we introduce some basic terminology that will be used throughout the manuscript. In Section 3, we present the problem that we study, as well as our main result. We prove the main result in Sections 4, 5, and 6. We summarize our results and discuss directions for future work in Section 7.

2. Preliminaries

We first present some basic terminology and notation that we use throughout the manuscript.

2.1. Relation schemas

We denote by **S** a *relation schema* $R(A_1, \ldots, A_k)$ where R is a *relation symbol* and $(A_1, \ldots, A_k)$ is a sequence of distinct *attributes*. We refer to k as the *arity* of the schema. A *relation* r over **S** is a finite set of tuples $(c_1, \ldots, c_k)$ where each c_i is a constant. We refer to each tuple as a *fact* of r. For a fact f and an attribute A_i, we denote by $f[A_i]$ the value of f in the attribute A_i (i.e., if $f = (c_1, \ldots, c_k)$, then $f[A_i] = c_i$). We may omit stating the schema **S** of a relation r when it is clear from the context or irrelevant.

[1] To be precise, for general key constraints they classified the CQs without self joins, and for key constraints where the key consists of a single attribute they classified all CQs.

Table 1
Specific FD schemas.

FD Schema	Relation Schema	FDs
$(\mathbf{S}_{2k}, \Delta_{2k})$	$R(A, B)$	$A \to B,\ B \to A$
$(\mathbf{S}_{ch}, \Delta_{ch})$	$R(A, B, C)$	$\emptyset \to A,\ B \to C$
$(\mathbf{S}_{2fd}, \Delta_{2fd})$	$R(A, B, C, D)$	$A \to B,\ C \to D$

fact	dept	time	plate	state	make
f_1	HPD	1500	AA11	CA	Acura
f_2	NYPD	1800	AA11	CA	Acura
f_3	PPD	1900	AA11	CA	Honda
f_4	LAPD	2000	AA11	CA	Honda
f_5	LAPD	1000	AA11	CA	Honda
f_6	LAPD	1600	AA11	CA	Mazda
f_7	HPD	1600	AA11	CA	Mazda
f_8	HPD	1100	AA11	CA	Mazda
f_9	NYPD	1500	AA11	CA	Nissan

Fig. 1. Inconsistent relation over the schema $(\mathbf{S}, \Delta)$ of Example 2.1.

2.2. Functional dependencies

A *Functional Dependency* (*FD* for short) over a relation schema $\mathbf{S}$ is an expression of the form $X \to Y$, where X and Y are sets of attributes of $\mathbf{S}$. We may also write X and Y by simply concatenating the attribute symbols; for example, we may write $AB \to C$ instead of $\{A, B\} \to \{C\}$ for the relation schema $R(A, B, C)$. An FD $X \to Y$ is *trivial* if $Y \subseteq X$, and otherwise it is *nontrivial*.

A relation r *satisfies* an FD $X \to Y$ if for every two facts f and g in r, if f and g agree on (i.e., have the same constants in the position of) the attributes of X, then they also agree on the attributes of Y. We say that r satisfies a set Δ of FDs if r satisfies every FD in Δ; otherwise, we say that r *violates* Δ. Two FD sets over the same schema are *equivalent* if every relation that satisfies one also satisfies the other. For example, $\{A \to BC, C \to A\}$ and $\{A \to C, C \to AB\}$ are equivalent. An FD $X \to Y$ is *entailed* by Δ (denoted by $\Delta \models X \to Y$) if for every relation r over the schema, if r satisfies Δ, then it also satisfies $X \to Y$. The *closure* of an attribute set X w.r.t. an FD set Δ, denoted by $X^{+,\Delta}$, is the set of attributes A such that $\Delta \models X \to A$.

Let Δ be a set of FDs. We say that Δ is *minimal* [20] if it satisfies the following properties:

1. FDs in Δ contain a single attribute on their right-hand side; that is, they have the form $X \to A$ with A an attribute.
2. No FD in Δ is redundant; that is, no $X \to A$ in Δ satisfies $(\Delta \setminus \{X \to A\}) \models X \to A$.
3. There is no redundant attribute in Δ; that is, no FD $XB \to A$ with $B \notin X$ in Δ satisfies $\Delta \models X \to A$.

A *minimal cover* [21] of an FD set Δ is a minimal set Δ_m of FDs that is equivalent to Δ.

An *FD schema* is a pair $(\mathbf{S}, \Delta)$, where $\mathbf{S}$ is a relation schema and Δ is a set of FDs over $\mathbf{S}$. Two FD schemas $(\mathbf{S}, \Delta)$ and $(\mathbf{S}', \Delta')$ are *equivalent* if $\mathbf{S} = \mathbf{S}'$ and Δ is equivalent to Δ'. For example, Table 1 depicts specific FD schemas that we refer to throughout the paper. In $(\mathbf{S}_{2k}, \Delta_{2k})$ the subscript "2k" stands for "two keys" as it includes two key constraints (one where A is the key and one where B is the key). In $(\mathbf{S}_{ch}, \Delta_{ch})$ the subscript "ch" stands for "chain" and we introduce this term later in Section 3. In $(\mathbf{S}_{2fd}, \Delta_{2fd})$ the subscript "2fd" refers to simply "two FDs," and its two FDs are disjoint in the attributes they involve.

Example 2.1. Fig. 1 depicts an inconsistent relation that stores information about traffic camera records, as established by integrating several data sources. For instance, the fact f_1 states that an Acura car with a California plate number New York has been recorded by the Huston Police Department at time 1500.

In the corresponding FD schema $(\mathbf{S}, \Delta)$, the relation schema $\mathbf{S}$ is

$$\mathit{Image}(\mathsf{dept}, \mathsf{time}, \mathsf{plate}, \mathsf{state}, \mathsf{make})$$

and Δ consists of the two FDs: $\mathsf{plate\ state} \to \mathsf{make}$ (the brand of the car is determined by the state and plate number), and $\mathsf{time\ plate\ state} \to \mathsf{dept}$ (a car cannot be recorded by two police departments at the same time). Note, however, that the relation of Fig. 1 violates the FDs. In particular, all records mention the same plate and state, but there is no agreement on the brand. Moreover, f_6 and f_7 mention recordings of the same car at the same time in Los Angeles and Huston, respectively. □

2.3. Conflict graphs

Conventionally, inconsistent databases are databases that violate integrity constraints [1]. Here, we use the abstraction of a *conflict graph* that can represent inconsistencies for various types of integrity constraints, including FDs. The translation

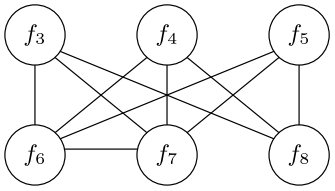

Fig. 2. The conflict graph of the subset $\{f_3, f_4, f_5, f_6, f_7, f_8\}$ of Fig. 1.

from the logical constraints to the conflict graph for FDs can be done in polynomial time even under *combined complexity* (i.e., when the schema, the constraints, and the relation are all given as input).

All graphs used in this paper are finite and undirected. The set of nodes of a graph g is denoted by $\mathsf{N}(g)$, and its set of edges is denoted by $\mathsf{E}(g)$; here, every edge $e \in \mathsf{E}(g)$ is a pair $\{u, v\}$ of distinct nodes. Two nodes of a graph g are *neighbors* if they are connected by an edge in $\mathsf{E}(g)$. An *independent set* of a graph g is a set U of nodes that does not include any edge; that is, $U \subseteq \mathsf{N}(g)$ and $e \not\subseteq U$ for all $e \in \mathsf{E}(g)$. An independent set U of a graph g is *maximal* if U is not strictly contained in any other independent set of g.

For an FD schema $(\mathbf{S}, \Delta)$ and a relation r over $\mathbf{S}$, the *conflict graph* $\mathcal{G}^r_\Delta$ is the graph over the facts of r that has an edge between every two facts that violate one or more FDs in Δ. An independent set of $\mathcal{G}^r_\Delta$ is also called a *consistent subset* (of r w.r.t. $\mathcal{G}^r_\Delta$), and a *maximal* independent set of $\mathcal{G}^r_\Delta$ is also called a *subset repair* (of r w.r.t. $\mathcal{G}^r_\Delta$) [1,2]. We denote by $\mathrm{SRep}(\mathcal{G}^r_\Delta)$ the set of all subset repairs of r w.r.t. $\mathcal{G}^r_\Delta$.

Example 2.2. Fig. 2 depicts the graph $\mathcal{G}^{r'}_\Delta$, where Δ is given in Example 2.1, and r' consists of a subset of the facts of Fig. 1. (We do not include all the facts to avoid clutter.) Observe that there is an edge between f_3 and f_6 since they jointly violate the FD plate state $\rightarrow$ make, and there is an edge between f_6 and f_7 since they violate the FD time plate state $\rightarrow$ dept. The relation of Fig. 1 has five subset repairs: *(a)* $\{f_1, f_2\}$, *(b)* $\{f_3, f_4, f_5\}$, *(c)* $\{f_6, f_8\}$, *(d)* $\{f_7, f_8\}$, and *(e)* $\{f_9\}$. The reader can verify that, indeed, each subset repair corresponds to a maximal independent set of the graph, and that no other subset repairs exist. □

3. Problem definition and main result

We investigate the problem of counting the subset repairs for fixed FD schemas $(\mathbf{S}, \Delta)$. Formally, the problem $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$ is that of computing the cardinality $|\mathrm{SRep}(\mathcal{G}^r_\Delta)|$ for a given relation r over $\mathbf{S}$. For counting problems, a basic tractable class is **FP** (functions computable in polynomial time), and a common measure of intractability is $\sharp$**P**-hardness (or $\sharp$**P**-completeness). $\sharp$**P** is the class of functions that count the number of witnesses for an NP problem (e.g., the number of satisfying assignments for a given formula in propositional logic). Hardness for $\sharp$**P** is defined by means of polynomial-time Turing reductions. Using an oracle to a $\sharp$**P**-hard function, one can solve in polynomial time every problem in the polynomial hierarchy [22].

Our main result in this manuscript is a dichotomy in data complexity for all problems $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$, classifying FD schemas into those for which the problem is in **FP** and those for which the problem is $\sharp$**P**-complete. To present our dichotomy, we need the following definition.

Definition 3.1 *(Left-hand-side chain).* An FD schema $(\mathbf{S}, \Delta)$ has a left-hand-side chain (*lhs chain* for short) if for every two FDs $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ in Δ, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$. □

Note that if $(\mathbf{S}, \Delta)$ has an lhs chain, then the FDs of Δ can be arranged in an order $X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n$ such that $X_i \subseteq X_j$ for all $i < j$. As an example, every FD schema with at most one FD has an lhs chain. As another example, the FD schema $(\mathbf{S}_{\mathrm{ch}}, \Delta_{\mathrm{ch}})$ of Table 1 has an lhs chain. However, the FD schema $(\mathbf{S}_{2\mathrm{k}}, \Delta_{2\mathrm{k}})$ in that table does not have an lhs chain, since there is no containment among $\{A\}$ and $\{B\}$. Our main result establishes that having an lhs chain captures precisely the cases where $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$ can be computed in polynomial time.

Theorem 3.2. *Let $(\mathbf{S}, \Delta)$ be an FD schema. If $(\mathbf{S}, \Delta)$ is equivalent to an FD schema with a left-hand-side chain, then $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$ is in* **FP**. *Otherwise, $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$ is $\sharp$**P**-complete.*

We now illustrate the application of Theorem 3.2 for classifying FD schemas into tractable and intractable ones (w.r.t. counting repairs).

Example 3.3. Consider again the FD schema $(\mathbf{S}, \Delta)$ of Example 2.1, and note that the FD set Δ has an lhs chain. Consequently, from Theorem 3.2 we conclude that $\#\mathrm{SRep}\langle\mathbf{S}, \Delta\rangle$ can be solved in polynomial time for the FD schema of our running example. Observe that we could equivalently write Δ with redundancy, as consisting of the following FDs.

(a) A P_4-free graph. (b) A graph that is not P_4-free.

Fig. 3. Illustration of P_4-freeness.

plate state → make

time plate state → dept

dept plate state → make

In this case, $(\mathbf{S}, \Delta)$ would not have an lhs chain, but would rather be *equivalent* to an FD schema with an lhs chain.

On the other hand, consider the FD schema $(\mathbf{S}_{2k}, \Delta_{2k})$ of Table 1. Here, the FD set Δ does not have an lhs chain, and it is not equivalent to any FD set with an lhs chain; hence, Theorem 3.2 implies that $\#\text{SRep}\langle \mathbf{S}_{2k}, \Delta_{2k}\rangle$ cannot be solved in polynomial time (under standard complexity assumptions). We get a similar lower bound for the FD schema $(\mathbf{S}_{2fd}, \Delta_{2fd})$. □

The proof of Theorem 3.2 is based on the following result, providing different characterizations of the classification criterion.

Theorem 3.4. *Let $(\mathbf{S}, \Delta)$ be an FD schema. The following statements are equivalent:*

1. *$(\mathbf{S}, \Delta)$ is, up to equivalence, an FD schema with a left-hand-side chain;*
2. *Δ has a single minimal cover Δ_m, and Δ_m has a left-hand-side chain;*
3. *$\mathcal{G}^r_\Delta$ is P_4-free for every relation r over $\mathbf{S}$.*

The following dichotomy is an immediate corollary of Theorems 3.2 and 3.4.

Corollary 3.5. *For every FD schema $(\mathbf{S}, \Delta)$, the problem $\#\text{SRep}\langle \mathbf{S}, \Delta\rangle$ is either in* **FP** *or* ♯**P**-*complete. Moreover, it is decidable in polynomial time in the size of $(\mathbf{S}, \Delta)$ which of the two cases applies.*

The equivalence between statements 1 and 2 of Theorem 3.4 shows that it is decidable in polynomial time whether a given FD schema is equivalent to an FD schema with an lhs chain. The equivalence between statements 1 and 3 proves the tractability side of Theorem 3.2, as we explain next.

Recall that $\#\text{SRep}\langle \mathbf{S}, \Delta\rangle$ is, in fact, the problem of counting the maximal independent sets of the conflict graph. In general, computing the number of maximal independent sets of a graph is ♯**P**-complete [17]. A known island of tractability is the class of P_4-*free graphs*, namely the graphs where no induced subgraph is a four-node path. A P_4-free graph is also called *cograph*, and the class of P_4-free graphs is characterized as the one generated from single nodes by repeatedly applying disjoint unions and complementation [18]. As an example, the graph depicted in Fig. 3a is P_4-free. The graph induced by $\{v_3, v_4, v_5, v_6\}$, for example, is not a path of length four. In contrast, the graph of Fig. 3b is not P_4-free, since the subgraph induced by $\{v_3, v_4, v_6, v_7\}$ is a path of length four.

We now illustrate the equivalence $1 \leftrightarrow 3$ of Theorem 3.4 in the following example.

Example 3.6. Consider the conflict graph of Fig. 2. The reader can verify that this graph is P_4-free, by inspecting every four nodes (though there are more efficient ways of verifying P_4-freeness [18]). For example, the nodes along the path $f_6 - f_7 - f_5 - f_8$ do not induce a P_4, since there is an edge between f_6 and f_5. Theorem 3.4, implies that every conflict graph of the FDs of our running example (Example 2.1) is P_4-free.

Next, consider the FD schema $(\mathbf{S}_{2k}, \Delta_{2k})$ of Table 1. Recall that Theorem 3.2 implies that $\#\text{SRep}\langle \mathbf{S}_{2k}, \Delta_{2k}\rangle$ is ♯**P**-complete. And, indeed, this can be evidenced by the relation $\{R(0,1), R(0,2), R(1,2), R(1,3)\}$ that has a P_4 conflict graph. As for the FD schema $(\mathbf{S}_{2fd}, \Delta_{2fd})$, the reader can easily verify that the relation $\{R(0,0,0,0), R(0,1,1,1), R(1,1,1,2), R(1,2,2,2)\}$ has a P_4 conflict graph. □

Corneil et al. [18] have shown the following.

Proposition 3.7. [18] *The maximal independent sets can be counted in polynomial time for P_4-free graphs.*

The tractability side of Theorem 3.2 immediately follows from Proposition 3.7 and the implication $1 \rightarrow 3$ of Theorem 3.4. Moreover, the fact that the translation from FDs to a conflict graph can be done in polynomial time under *combined complexity* implies that if the classification criterion of Theorem 3.2 is satisfied, the number of subset repairs can be computed in polynomial time even under combined complexity.

In the remainder of this manuscript, we prove the rest of Theorem 3.2. In particular, in Section 4, we prove that the classification criterion can be checked in polynomial time, by showing the equivalence $1 \leftrightarrow 2$ of Theorem 3.4. Then, in Section 5, we prove the tractable side of the theorem, by showing the equivalence $1 \leftrightarrow 3$ of Theorem 3.4. Finally, we prove the hardness side of Theorem 3.2 in Section 6.

4. Minimal covers of lhs-chain schemas

In this section, we prove the equivalence $1 \leftrightarrow 2$ of Theorem 3.4, which implies that we can decide in polynomial time whether a given FD schema $(\mathbf{S}, \Delta)$ belongs to the tractable or intractable side of Theorem 3.2.

Our proof consists of two parts. First, we prove that a minimal set Δ of FDs with an lhs chain has a unique representation (that is, the only minimal cover of Δ is Δ itself). Then, we prove that an FD set Δ that is equivalent to an FD set with an lhs chain has a minimal cover with an lhs chain. The combination of these two results implies that an FD set Δ that is equivalent to an FD set with an lhs chain has a single minimal cover Δ_m, and this minimal cover has an lhs chain.

In the proof, we use the standard decision procedure for logical implication of FDs [23]. According to this decision procedure, a set Δ of FD entails an FD $X \rightarrow A$ if and only if there exists a sequence $X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n$ of FDs in Δ such that:

- $A \in Y_n$,
- for every $i \in \{1, \ldots, n\}$ we have that $X_i \subseteq \left[X \cup \left(\bigcup_{j=1}^{i-1}(X_j \cup Y_j)\right)\right]$.

Hereon, we refer to a sequence of FDs that is used by this decision procedure to decide $\Delta \models X \rightarrow A$ simply as a *sequence of FDs that implies* $\Delta \models X \rightarrow A$.

We start by proving the following simple lemma.

Lemma 4.1. *Let Δ be an FD set with an lhs chain and let $X \rightarrow A$ be an FD entailed by Δ. Then, there is a sequence $X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n$ of FDs in Δ that implies $\Delta \models X \rightarrow A$, such that $X_1 \subseteq \cdots \subseteq X_n$.*

Proof. Let $X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n$ be a sequence of FDs in Δ that implies $\Delta \models X \rightarrow A$ and has a maximal lhs-chain prefix. That is, for some $1 \leq k \leq n$, it holds that $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_k$, and there is no sequence $Z_1 \rightarrow W_1, \ldots, Z_m \rightarrow W_m$ of FDs in Δ that also implies $\Delta \models X \rightarrow A$ such that $Z_1 \subseteq Z_2 \subseteq \cdots \subseteq Z_{k+1}$. We claim that $k = n$; that is, there exists a sequence of FDs in Δ that implies $\Delta \models X \rightarrow A$ and forms an lhs chain.

Assume, towards a contradiction, that for some $1 \leq i \leq n$, it holds that $X_i \not\subseteq X_{i+1}$. Then, since Δ has an lhs chain, we have that $X_{i+1} \subseteq X_i$, which, combined with the fact that $X_i \subseteq \left[X \cup \left(\bigcup_{j=1}^{i-1}(X_j \cup Y_j)\right)\right]$, implies that $X_{i+1} \subseteq \left[X \cup \left(\bigcup_{j=1}^{i-1}(X_j \cup Y_j)\right)\right]$; hence, we can swap the FDs $X_i \rightarrow Y_i$ and $X_{i+1} \rightarrow Y_{i+1}$ in the sequence, and obtain a different sequence that satisfies all the conditions required by the standard decision procedure for logical implication of FDs. Moreover, this new sequence has a longer lhs-chain prefix $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_{i+1} \subseteq X_i$, which is a contradiction to our assumption. Note that if $i = n - 1$, then after swapping $X_{n-1} \rightarrow Y_{n-1}$ and $X_n \rightarrow Y_n$ we can simply remove the FD $X_{n-1} \rightarrow Y_{n-1}$ from the sequence, as we have that $A \in Y_n$. □

Next, we prove that a minimal set of FDs with an lhs chain has a single representation.

Lemma 4.2. *Let Δ be a minimal set of FDs with an lhs chain. Every minimal cover Δ_m of Δ satisfies $\Delta_m = \Delta$.*

Proof. Let Δ_m be a minimal cover of Δ. We will prove that $\Delta_m \subseteq \Delta$, and since it cannot be the case that $\Delta_m \subsetneq \Delta$ (as Δ is minimal), we will conclude that $\Delta_m = \Delta$.

Let $X \rightarrow A$ be an FD in Δ_m. Clearly, $A \notin X$. Since Δ and Δ_m are equivalent, we have that $\Delta \models X \rightarrow A$. Let $X_1 \rightarrow A_1, \ldots, X_n \rightarrow A_n$ be a sequence of FDs in Δ that implies $\Delta \models X \rightarrow A$, such that $X_1 \subseteq \cdots \subseteq X_n$. Such a sequence exists according to Lemma 4.1. We now show that $X_n = X$ and $A_n = A$ (thus, $X \rightarrow A \in \Delta$), which will conclude our proof.

Assume, by way of contradiction, that there is $B \in X$ such that $B \notin X_n$. Since the sequence $X_1 \rightarrow A_1, \ldots, X_n \rightarrow A_n$ is an lhs chain, it holds that $B \notin X_i$ for every $1 \leq i \leq n$. Thus, this sequence actually shows that $\Delta \models (X \setminus \{B\}) \rightarrow A$, and a minimal cover of Δ cannot contain the FD $X \rightarrow A$, which is a contradiction to the fact that $X \rightarrow A \in \Delta_m$; hence, we have that $X \subseteq X_n$. Finally, we have that $X_n \rightarrow A \in \Delta$ and $\Delta \models X \rightarrow A$, and because Δ is minimal (hence, does not have redundant attributes), we conclude that $X_n = X$ and $X \rightarrow A \in \Delta$. □

Finally, we prove that an FD set Δ that is equivalent to an FD set with an lhs chain has a minimal cover with an lhs chain.

Lemma 4.3. *Let Δ be an FD set that is equivalent to an FD set with an lhs chain. Then, Δ has a minimal cover with an lhs chain.*

Proof. It is easily verified that we can assume, without loss of generality, that *(1)* Δ itself has an lhs chain, and that *(2)* all FDs in Δ have a single attribute on the right-hand side. We can also assume, without loss of generality, that all redundant FDs have been removed from Δ, because a set of FDs with an lhs chain still has an lhs chain after removal of one or more FDs. Let $X_1 \to A_1, \ldots, X_n \to A_n$ be an lhs chain of Δ. Let $Y = \{A_1, \ldots, A_n\}$. Let $\Delta_m = \{X_1 \setminus Y \to A_1, \ldots, X_n \setminus Y \to A_n\}$, which clearly has an lhs chain. It suffices to show that Δ_m is a minimal cover of Δ. It is obvious that for all $1 \leq j \leq n$, we have $\Delta_m \models X_j \to A_j$. Conversely, for any integer j such that $1 \leq j \leq n$, we claim that $X_1 \to A_1, \ldots, X_j \to A_j$ is a sequence of FDs that implies $\Delta \models X_j \setminus Y \to A_j$. To prove this claim, it clearly suffices to show that for every $i \in \{1, \ldots j\}$, we have that $X_i \subseteq (X_j \setminus Y) \cup \{A_1, \ldots, A_{i-1}\}$. Assume for the sake of contradiction that $X_i \not\subseteq (X_j \setminus Y) \cup \{A_1, \ldots, A_{i-1}\}$ for some $i \in \{1, \ldots, j\}$. Since $X_i \subseteq X_j$, there must be some $k > i$ such that $A_k \in X_i$. But then, since $X_i \subseteq X_k$, we have that $A_k \in X_k$, contradicting that the FD $X_k \to A_k$ is not trivial.

Since Δ_m has a single attribute on the right-hand side of every FD and does not contain redundant FDs, it is only left to prove that Δ_m has no redundant attributes. Assume towards a contradiction that there is $1 \leq j \leq n$ and $B \in X_j \setminus Y$ such that $\Delta_m \models X_j \setminus YB \to A_j$. Then, there exists a shortest sequence of FDs (call it σ) that implies $\Delta_m \models X_j \setminus YB \to A_j$. Since Δ_m contains $X_j \setminus Y \to A_j$, we can assume, without loss of generality, that this sequence σ contains no FD that comes after $X_j \setminus Y \to A_j$ in the lhs chain of Δ_m. Therefore, the last FD in σ must be equal to $X_i \setminus Y \to A_i$ for some $i \in \{1, \ldots, j\}$ such that $A_i = A_j$. Since $B \notin Y$, there is no FD in Δ_m whose right-hand side is B. Therefore, since $B \in X_j \setminus Y$, the FD $X_j \setminus Y \to A_j$ cannot occur in σ. It follows $i < j$. Then, since $X_i \subseteq X_j$ and Δ contains $X_i \to A_j$, the FD $X_j \to A_j$ is redundant in Δ, contradicting our assumption that Δ contains no redundant FDs. This concludes the proof. □

Lemmas 4.2 and 4.3 imply that an FD set Δ that has an lhs chain (up to equivalence) has a single minimal cover that has an lhs chain; hence, in the remainder of this manuscript, we refer to *the* minimal cover of Δ.

5. Conflict graphs of lhs-chain schemas

In this section, we prove the equivalence $1 \leftrightarrow 3$ of Theorem 3.4. Note that we use only the implication $1 \to 3$ in the proof of Theorem 3.2; however, the result shown in this section is stronger, as it implies that P_4-free graphs are, in fact, a characterization of FD schemas with an lhs chain, that does not depend on any complexity assumptions.

We start by proving the implication $1 \to 3$, and then we prove the implication $3 \to 1$.

5.1. Left-hand-side chain implies P_4-freeness

We start by proving that if $(\mathbf{S}, \Delta)$ is equivalent to some FD schema with an lhs chain, then $\mathcal{G}^r_\Delta$ is P_4-free for every relation r of $\mathbf{S}$. Let Δ_m be the minimal cover of Δ. According to the result of the previous section, the FD set Δ_m has an lhs chain. Let r be a relation over $\mathbf{S}$. Clearly, it holds that $\mathcal{G}^r_\Delta$ and $\mathcal{G}^r_{\Delta_m}$ are the same graph; thus, it is sufficient to prove that $\mathcal{G}^r_{\Delta_m}$ is P_4-free. To show this, we will assume that $\mathcal{G}^r_{\Delta_m}$ is not P_4-free and derive a contradiction.

If $\mathcal{G}^r_{\Delta_m}$ is not P_4-free, then there are four nodes v_1, v_2, v_3, v_4 in the graph, such that $\{v_1, v_2\}, \{v_2, v_3\}, \{v_3, v_4\} \in E(\mathcal{G}^r_{\Delta_m})$, but $\{v_1, v_3\}, \{v_1, v_4\}, \{v_2, v_4\} \notin E(\mathcal{G}^r_{\Delta_m})$. Hence, there are four facts f_1, f_2, f_3, f_4 in r, such that $\{f_1, f_2\}$, $\{f_2, f_3\}$ and $\{f_3, f_4\}$ violate Δ_m, but $\{f_1, f_3\}$, $\{f_1, f_4\}$ and $\{f_2, f_4\}$ do not violate Δ_m.

Let us assume that $\{f_1, f_2\}$ violates the FD $X \to A$, $\{f_2, f_3\}$ violates the FD $X' \to A'$ and $\{f_3, f_4\}$ violates the FD $X'' \to A''$. Since Δ_m has an lhs chain, one set among X, X', X'' is included in the other two, which leads to three possibilities:

- $X \subseteq X'$ and $X \subseteq X''$. In this case, all of these facts agree on the attributes of X. However, f_1 and f_2 do not agree on A. Then, f_4 cannot agree with both f_1 and f_2 on this attribute; thus, it holds that either $\{f_1, f_4\}$ or $\{f_2, f_4\}$ violates the FD $X \to A$.
- $X' \subseteq X$ and $X' \subseteq X''$. In this case, all of these facts agree on the attributes of X'. However, f_2 and f_3 do not agree on A'. Since $\{f_1, f_3\}$ does not violate Δ_m, it holds that f_1 and f_3 agree on A'. Moreover, since $\{f_2, f_4\}$ does not violate Δ_m, we have that f_2 and f_4 agree on A'. Thus, f_1 and f_4 do not agree on A', and $\{f_1, f_4\}$ violates the FD $X' \to A'$.
- $X'' \subseteq X$ and $X'' \subseteq X'$. This case and the first one are symmetrical.

Note that in all three cases we get a contradiction (as v_1, v_2, v_3, v_4 do not induce a path of length four), and this concludes our proof.

5.2. P_4-freeness implies left-hand-side chain

Now, we prove that if $(\mathbf{S}, \Delta)$ is not equivalent to any FD schema with an lhs chain, then there exists a relation r of $\mathbf{S}$, such that $\mathcal{G}^r_\Delta$ is not P_4-free. Clearly, since $(\mathbf{S}, \Delta)$ is not equivalent to any FD schema with an lhs chain, a minimal cover of

fact	A	B	C	D
f_1	a	c	c	c
f_2	a	d	d	d
f_3	f	g	d	g
f_4	f	h	h	h

Fig. 4. The relation r constructed for the schema $(\mathbf{S}_{\text{2fd}}, \Delta_{\text{2fd}})$ in the proof of Lemma 5.1.

Δ cannot have an lhs chain. Moreover, if we look at a minimal cover Δ_m of Δ, then for each relation r of $\mathbf{S}$ the graphs $\mathcal{G}^r_\Delta$ and $\mathcal{G}^r_{\Delta_m}$ are the same. Thus, it is sufficient to prove that there exists a relation r for which $\mathcal{G}^r_{\Delta_m}$ is not P_4-free.

Lemma 5.1. *Let $(\mathbf{S}, \Delta)$ be an FD schema, such that Δ is a minimal set of FDs that does not have an lhs chain. Then, there exists a relation r of $\mathbf{S}$, such that $\mathcal{G}^r_\Delta$ is not P_4-free.*

Proof. Since Δ does not have an lhs chain, there exist two FDs, $X \to B$ and $X' \to B'$, in Δ, such that $X \not\subseteq X'$ and $X' \not\subseteq X$. We build a relation r over $\mathbf{S}$, such that $\mathcal{G}^r_\Delta$ is not P_4-free in the following way. The relation r will contain the following four facts f_1, f_2, f_3, f_4, defined using the constants $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}, \mathtt{h}$.

- $f_1[A] = \mathtt{a}$ for all $A \in X \setminus (X \cap X')^{+,\Delta}$, $f_1[A] = \mathtt{b}$ for all $A \in (X \cap X')^{+,\Delta}$, and $f_1[A] = \mathtt{c}$ otherwise.
- $f_2[A] = \mathtt{a}$ for all $A \in X \setminus (X \cap X')^{+,\Delta}$, $f_2[A] = \mathtt{b}$ for all $A \in (X \cap X')^{+,\Delta}$, $f_2[A] = \mathtt{d}$ for all $A \in X' \setminus (X \cap X')^{+,\Delta}$, and $f_2[A] = \mathtt{e}$ otherwise.
- $f_3[A] = \mathtt{f}$ for all $A \in X \setminus (X \cap X')^{+,\Delta}$, $f_3[A] = \mathtt{b}$ for all $A \in (X \cap X')^{+,\Delta}$, $f_3[A] = \mathtt{d}$ for all $A \in X' \setminus (X \cap X')^{+,\Delta}$, and $f_3[A] = \mathtt{g}$ otherwise.
- $f_4[A] = \mathtt{f}$ for all $A \in X \setminus (X \cap X')^{+,\Delta}$, $f_4[A] = \mathtt{b}$ for all $A \in (X \cap X')^{+,\Delta}$, and $f_4[A] = \mathtt{h}$ otherwise.

Since Δ is minimal we have that $B \notin X$. Moreover, since $X \not\subseteq X'$, we have that $(X \cap X') \subsetneq X$, and, again, because Δ is minimal, it holds that $\Delta \not\models (X \cap X' \to B)$. It follows that $B \notin X \cup (X \cap X')^{+,\Delta}$; hence, $\{f_1, f_2\} \not\models (X \to B)$, and similarly, $\{f_3, f_4\} \not\models (X \to B)$. By symmetrical reasoning, we have that $B' \notin X' \cup (X \cap X')^{+,\Delta}$; thus, $\{f_2, f_3\} \not\models (X' \to B')$.

Finally, the facts f_1 and f_3 only agree on the attributes of $(X \cap X')^{+,\Delta}$. Assume, by way of contradiction, that $\{f_1, f_3\}$ violates an FD $Y \to C$ in Δ. Hence, it holds that $Y \subseteq (X \cap X')^{+,\Delta}$, but $C \notin (X \cap X')^{+,\Delta}$. Then, since $\Delta \models (X \cap X') \to Y$ and $\Delta \models Y \to C$, we have that $\Delta \models (X \cap X') \to C$, and $C \in (X \cap X')^{+,\Delta}$, which is a contradiction. We can similarly show that $\{f_1, f_4\} \models \Delta$ and $\{f_2, f_4\} \models \Delta$, and that concludes our proof. □

The relation constructed in the proof of Lemma 5.1 for the FD schema $(\mathbf{S}_{\text{2fd}}, \Delta_{\text{2fd}})$ is illustrated in Fig. 4 (note that Δ_{2fd} is minimal). The reader can verify that, indeed, the conflict graph $\mathcal{G}^r_{\Delta_{\text{2fd}}}$ is not P_4-free.

6. Proof of hardness

Finally, we prove the hardness side of Theorem 3.2. We begin with the simplest FD schema that does not have an lhs chain, namely $(\mathbf{S}_{\text{2k}}, \Delta_{\text{2k}})$ of Table 1. Consider a relation r over $\mathbf{S}_{\text{2k}}$. If r is viewed as the set of edges of a bipartite graph g (where the constants of r correspond to the nodes of g, assuming, without the loss of generality, that the sets of constants of each attribute are disjoint), then counting the subset repairs of r is the problem of counting the *maximal matchings* of g, which is ♯**P**-complete [24]. Hence, we get the following.

Proposition 6.1. *The problem $\#\text{SRep}\langle \mathbf{S}_{\text{2k}}, \Delta_{\text{2k}} \rangle$ is ♯**P**-complete.*

Next, we use the concept of a *fact-wise reduction* [25]. Let $(\mathbf{S}, \Delta)$ and $(\mathbf{S}', \Delta')$ be two FD schemas. A *mapping* from $\mathbf{S}$ to $\mathbf{S}'$ is a function μ that maps facts over $\mathbf{S}$ to facts over $\mathbf{S}'$. (We say that f is a fact *over* $\mathbf{S}$ if f is a fact of some relation r over $\mathbf{S}$.) We extend a mapping μ to map relations r over $\mathbf{S}$ to relations over $\mathbf{S}'$ by defining $\mu(r)$ to be $\{\mu(f) \mid f \in r\}$. A *fact-wise reduction* from $(\mathbf{S}, \Delta)$ to $(\mathbf{S}', \Delta')$ is a mapping Π from $\mathbf{S}$ to $\mathbf{S}'$ with the following properties.

1. Π is injective; that is, for all facts f and g over $\mathbf{S}$, if $\Pi(f) = \Pi(g)$ then $f = g$.
2. Π preserves consistency and inconsistency; that is, for all facts f and g over $\mathbf{S}$, $\{f, g\}$ satisfies Δ if and only if $\{\Pi(f), \Pi(g)\}$ satisfies Δ'.
3. Π is computable in polynomial time.

The following lemma is straightforward.

Lemma 6.2. *Let $(\mathbf{S}, \Delta)$ and $(\mathbf{S}', \Delta')$ be FD schemas, and suppose that there is a fact-wise reduction from $(\mathbf{S}, \Delta)$ to $(\mathbf{S}', \Delta')$. If $\#\text{SRep}\langle \mathbf{S}, \Delta \rangle$ is ♯**P**-hard, then $\#\text{SRep}\langle \mathbf{S}', \Delta' \rangle$ is ♯**P**-hard as well.*

Hence, we complete the proof by showing that there is a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to any FD schema $(\mathbf{S}, \Delta)$ that is not equivalent to an FD schema with an lhs chain. Then, Proposition 6.1 and Lemma 6.2 imply that computing $\#\text{SRep}\langle \mathbf{S}, \Delta\rangle$ is $\sharp\mathbf{P}$-complete for all such schemas. Note that the existence of a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to any FD schema $(\mathbf{S}, \Delta)$ that is not equivalent to an FD schema with an lhs chain is a stronger result than what we need to prove the hardness side of Theorem 3.2 (instead, we could construct a reduction from $\#\text{SRep}\langle \mathbf{S}_{2k}, \Delta_{2k}\rangle$ to $\#\text{SRep}\langle \mathbf{S}, \Delta\rangle$), but it is of independent interest, since fact-wise reductions constitute general tools for proving dichotomy results.

We start by proving the claim for minimal sets of FDs.

Lemma 6.3. *Let $(\mathbf{S}, \Delta)$ be an FD schema, such that Δ is a minimal set of FDs that does not have an lhs chain. Then, there is a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta)$.*

Proof. Since Δ does not have an lhs chain, there are two FDs $X \to A$ and $X' \to A'$, in Δ, such that $X \not\subseteq X'$ and $X' \not\subseteq X$. We define a fact-wise reduction $\Pi : \mathbf{S}_{2k} \to \mathbf{S}$, using the FDs $X \to A$ and $X' \to A'$ and the constant $\odot \in \mathsf{Const}$. Let $f = (a, b)$ be a fact over $\mathbf{S}_{2k}$ and let $\{A_1, \ldots, A_n\}$ be the set of attributes in $\mathbf{S}$. We define Π as follows:

$$\Pi(f)[A_k] \stackrel{\text{def}}{=} \begin{cases} \odot & A_k \in (X \cap X')^{+,\Delta} \\ a & A_k \in X \setminus (X \cap X')^{+,\Delta} \\ b & A_k \in X' \setminus (X \cap X')^{+,\Delta} \\ \langle a, b\rangle & \text{otherwise} \end{cases}$$

It is left to show that Π is a fact-wise reduction. To do so, we prove that Π is well defined, injective and preserves consistency and inconsistency.

Π is well defined. This is straightforward from the definition.

Π is injective. Let f, f' be two facts, such that $f = (a, b)$ and $f' = (a', b')$. Let us denote $\Pi(f) = (x_1, \ldots, x_n)$ and $\Pi(f') = (x'_1, \ldots, x'_n)$, and assume that $\Pi(f) = \Pi(f')$. Note that $X \setminus (X \cap X')^{+,\Delta}$ is not empty, as otherwise, the fact that $\Delta \models (X \cap X') \to X$ and $\Delta \models X \to A$ would imply that $\Delta \models (X \cap X') \to A$, which is a contradiction to the fact that Δ is minimal (recall that $(X \cap X') \subsetneq X$ since $X \not\subseteq X'$). Similarly, we have that $X' \setminus (X \cap X')^{+,\Delta}$ is not empty. Therefore, there are l and p such that $x_l = a$, $x_p = b$. Hence, $\Pi(f) = \Pi(f')$ implies that $x_l = x'_l$ and $x_p = x'_p$. We obtain that $a = a'$ and $b = b'$, which implies $f = f'$.

Π preserves consistency. Let $f = (a, b)$ and $f' = (a', b')$ be two distinct facts. We contend that the set $\{f, f'\}$ is consistent w.r.t. Δ_{2k} if and only if the set $\{\Pi(f), \Pi(f')\}$ is consistent w.r.t. Δ.

The "if" direction. Assume $\{f, f'\}$ is inconsistent w.r.t. Δ_{2k}. We prove that $\{\Pi(f), \Pi(f')\}$ is inconsistent w.r.t. Δ. Let us denote $\Pi(f) = (x_1, \ldots, x_n)$ and $\Pi(f') = (x'_1, \ldots, x'_n)$. If $\{f, f'\}$ is inconsistent w.r.t. Δ_{2k}, then either $a = a'$ and $b \neq b'$ or $a \neq a'$ and $b = b'$. By the definition of Π, for every attribute $A_k \in X$ it either holds that $x_k = a$ or $x_k = \odot$. It also holds that $x_k = b$ or $x_k = \langle a, b\rangle$ for $A_k = A$ since Δ is minimal and does not contain trivial FDs $X \to A$ or redundant attributes (thus, $\Delta \not\models (X \cap X') \to A$). Therefore, in the first case ($a = a'$ and $b \neq b'$), $\Pi(f)$ and $\Pi(f')$ agree on the attributes of X, but do not agree on the value of A, and $X \to A$ does not hold. Similarly, for every attribute $A_k \in X'$ it either holds that $x_k = b$ or $x_k = \odot$, and for $A_k = A'$ it either holds that $x_k = a$ or $x_k = \langle a, b\rangle$. Thus, in the second case ($a \neq a'$ and $b = b'$), the FD $X' \to A'$ does not hold. This leads us to the conclusion that $\{\Pi(f), \Pi(f')\}$ is inconsistent w.r.t. Δ.

The "only if" direction. Assume that $\{f, f'\}$ is consistent w.r.t. Δ_{2k}. We prove that $\{\Pi(f), \Pi(f')\}$ is consistent w.r.t. Δ. Note that $a \neq a'$ since otherwise the FD $A \to B$ implies that $b = b'$ and thus $f = f'$. Similarly, $b \neq b'$ due to the FD $B \to A$. Thus, $\Pi(f)$ and $\Pi(f')$ do not agree on the attributes on the left-hand side of any FD in Δ that contains an attribute $A_k \notin (X \cap X')^{+,\Delta}$. Hence, all these FDs are satisfied. Now, assume, by way of contradiction, that there is an FD $Y \to B$ in Δ such that $Y \subseteq (X \cap X')^{+,\Delta}$ and $\{\Pi(f), \Pi(f')\} \not\models Y \to B$. It holds that $\Delta \models (X \cap X') \to Y$ and $\Delta \models Y \to B$; hence, we have that $\Delta \models (X \cap X') \to B$. It follows that $B \in (X \cap X')^{+,\Delta}$ and $\Pi(f)[B] = \Pi(f')[B]$, which is a contradiction to the fact that $\Pi(f)$ and $\Pi(f')$ jointly violate $Y \to B$. □

At this point, we know that for an FD schema $(\mathbf{S}, \Delta)$ where Δ is a minimal FD set that does not have an lhs chain, there is a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta)$. It is left to show that this also holds for Δ that is not a minimal FD set and is not equivalent to an FD set with an lhs chain.

Lemma 6.4. *Let $(\mathbf{S}, \Delta)$ be an FD schema that is not equivalent to any FD schema with an lhs chain. Then, there is a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta)$.*

Proof. Let Δ_m be a minimal cover of Δ. Clearly, Δ_m does not have an lhs chain. Moreover, it is straightforward that if there exists a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta_m)$, there exists a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta)$ (the exact same reduction). Thus, it is sufficient to prove that there exists a fact-wise reduction from $(\mathbf{S}_{2k}, \Delta_{2k})$ to $(\mathbf{S}, \Delta_m)$. Lemma 6.3 implies that this is indeed the case. □

7. Conclusions

We investigated the complexity of the problem of counting subset repairs. We focused on FD constraints and established a dichotomy in data complexity, partitioning FD schemas into polynomial-time counting and ♯**P**-complete counting. We showed that the tractable FD schemas are the ones having an lhs chain, or equivalently, those guaranteeing P_4-free conflict graphs. Moreover, we showed that it is possible to decide in polynomial time whether or not an FD schema is on the tractable side of the dichotomy.

For future work, we highlight two main directions. The first is *approximate* counting of repairs. In particular, does the classification hold if we allow for approximate rather than exact repair counting? The second is that of counting repairs in more general repair frameworks that support additional types of integrity constraints (e.g., conditional FDs [26], equality-generating dependencies, and denial constraints [27]) and repairing operations (e.g., tuple addition and cell updates [28–30]).

CRediT authorship contribution statement

Ester Livshits: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Benny Kimelfeld:** Conceptualization, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Jef Wijsen:** Conceptualization, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Israel Science Foundation (ISF) Grant 1295/15.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jcss.2020.10.001.

References

[1] M. Arenas, L.E. Bertossi, J. Chomicki, Consistent query answers in inconsistent databases, in: PODS, ACM, 1999, pp. 68–79.
[2] F.N. Afrati, P.G. Kolaitis, Repair checking in inconsistent databases: algorithms and complexity, in: ICDT, 2009, pp. 31–41.
[3] L.E. Bertossi, Database Repairing and Consistent Query Answering, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011.
[4] P. Koutris, J. Wijsen, The data complexity of consistent query answering for self-join-free conjunctive queries under primary key constraints, in: PODS, ACM, 2015, pp. 17–29.
[5] D. Maslowski, J. Wijsen, A dichotomy in the complexity of counting database repairs, J. Comput. Syst. Sci. 79 (2013) 958–983.
[6] D. Maslowski, J. Wijsen, Counting database repairs that satisfy conjunctive queries with self-joins, in: ICDT, 2014, pp. 155–164.
[7] M. Calautti, M. Console, A. Pieris, Counting database repairs under primary keys revisited, in: PODS, 2019, pp. 104–118.
[8] S. Konieczny, J. Lang, P. Marquis, Quantifying information and contradiction in propositional logic through test actions, in: IJCAI, Morgan Kaufmann, 2003, pp. 106–111.
[9] J. Grant, A. Hunter, Measuring inconsistency in knowledgebases, J. Intell. Inf. Syst. 27 (2006) 159–184.
[10] A. Hunter, S. Konieczny, On the measure of conflicts: Shapley inconsistency values, Artif. Intell. 174 (2010) 1007–1026.
[11] J. Grant, A. Hunter, Measuring consistency gain and information loss in stepwise inconsistency resolution, in: Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 11th European Conference, Proceedings, ECSQARU 2011, Belfast, UK, June 29-July 1, 2011, 2011, pp. 362–373.
[12] J. Grant, A. Hunter, Analysing inconsistent information using distance-based measures, Int. J. Approx. Reason. 89 (2017) 3–26.
[13] M. Thimm, On the compliance of rationality postulates for inconsistency measures: a more or less complete picture, Künstl. Intell. 31 (2017) 31–39.
[14] E. Livshits, I.F. Ilyas, B. Kimelfeld, S. Roy, Principles of progress indicators for database repairing, CoRR, arXiv:1904.06492, 2019.
[15] L.E. Bertossi, Repair-based degrees of database inconsistency, in: LPNMR, 2019, pp. 195–209.
[16] F. Parisi, J. Grant, Inconsistency measures for relational databases, CoRR, arXiv:1904.03403, 2019.
[17] J.S. Provan, M.O. Ball, The complexity of counting cuts and of computing the probability that a graph is connected, SIAM J. Comput. 12 (1983) 777–788.
[18] D. Corneil, H. Lerchs, L. Burlingham, Complement reducible graphs, Discrete Appl. Math. 3 (1981) 163–174.
[19] E. Livshits, B. Kimelfeld, Counting and enumerating (preferred) database repairs, in: PODS, 2017, pp. 289–301.
[20] J.D. Ullman, Principles of Database and Knowledge-Base Systems, Volume I, Principles of Computer Science Series, vol. 14, Computer Science Press, 1988.
[21] D. Maier, Minimum covers in relational database model, J. ACM 27 (1980) 664–674.

[22] S. Toda, M. Ogiwara, Counting classes are at least as hard as the polynomial-time hierarchy, SIAM J. Comput. 21 (1992).
[23] S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley, 1995.
[24] S.P. Vadhan, The complexity of counting in sparse, regular, and planar graphs, SIAM J. Comput. 31 (2001) 398–427.
[25] B. Kimelfeld, A dichotomy in the complexity of deletion propagation with functional dependencies, in: PODS, 2012, pp. 191–202.
[26] P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, Conditional functional dependencies for data cleaning, in: ICDE, IEEE, 2007, pp. 746–755.
[27] T. Gaasterland, P. Godfrey, J. Minker, An overview of cooperative answering, J. Intell. Inf. Syst. 1 (1992) 123–157.
[28] J. Wijsen, Database repairing using updates, ACM Trans. Database Syst. 30 (2005) 722–768.
[29] F. Geerts, G. Mecca, P. Papotti, D. Santoro, The LLUNATIC data-cleaning framework, Proc. VLDB Endow. 6 (2013) 625–636.
[30] E. Livshits, B. Kimelfeld, S. Roy, Computing optimal repairs for functional dependencies, in: PODS, 2018, pp. 225–237.