Independence in Database Relations

Juha Kontinen[1], Sebastian Link[2], and Jouko Väänänen[1,3]

[1] Department of Mathematics and Statistics, University of Helsinki, Finland
[2] Department of Computer Science, The University of Auckland, New Zealand
[3] Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands

Abstract. We investigate the implication problem for independence atoms $X \perp Y$ of disjoint attribute sets X and Y on database schemata. A relation satisfies $X \perp Y$ if for every X-value and every Y-value that occurs in the relation there is some tuple in the relation in which the X-value occurs together with the Y-value. We establish an axiomatization by a finite set of Horn rules, and derive an algorithm for deciding the implication problem in low-degree polynomial time in the input. We show how to construct Armstrong relations which satisfy an arbitrarily given set of independence atoms and violate every independence atom not implied by the given set. Our results establish independence atoms as an efficient subclass of embedded multivalued data dependencies which are not axiomatizable by a finite set of Horn rules, and whose implication problem is undecidable.

1 Introduction

Independence and conditional independence are fundamental concepts in areas as diverse as artificial intelligence, probability theory, social choice theory, and statistics [2,9,17]. Recently, independence logic has been introduced as an extension of classical first-order logic by independence atoms [8]. In databases, conditional independence is better known as the class of embedded multivalued data dependencies. Their associated implication problem is known to be not axiomatizable by a finite set of Horn rules, and undecidable [11,16]. Multivalued data dependencies [3] form an efficient subclass of embedded multivalued data dependencies whose implication problem has been axiomatized by a finite set of Horn rules [1] and can be decided in almost linear time [5]. They form the basis for Fagin's fourth normal form proposal to avoid data redundancy in database relations and guarantee the absence of processing difficulties [3].

In this paper we investigate an efficient subclass of embedded multivalued data dependencies which we call—in accordance with [8]— independence atoms. Intuitively, a relation r satisfies the independence atom $X \perp Y$ between two disjoint sets X and Y of attributes, if for all tuples $t_1, t_2 \in r$ there is some tuple $t \in r$ which matches the values of t_1 on all attributes in X and matches the values of t_2 on all attributes in Y. In other words, in relations that satisfy $X \perp Y$, the occurrence of X-values is independent of the occurrence of Y-values.

L. Libkin, U. Kohlenbach, and R. de Queiroz (Eds.): WoLLIC 2013, LNCS 8071, pp. 179–193, 2013.

Example 1. Consider a simple database schema that stores information about the enrolment of students into a fixed course. In fact, the schema records for each enrolled student, the year in which they completed a prerequisite course. More formally, we have the schema ENROL={*S(tudent)*, *P(rerequisite)*, *Y(ear)*}. Intuitively, every student must have completed every prerequisite in some year. For this reason, for any value in the *Student* column and every value in the *Prerequisite* column there is some year for when this student has completed that prerequisite. That is, the values in the *Student* column are independent of the values in the *Prerequisite* column. A snapshot relation r over ENROL may be:

Student	*Prerequisite*	*Year*
Turing	Math201	1932
Gödel	Math201	1925
Turing	Phys220	1932
Gödel	Phys220	1925

illustrating the independence of *Student* from *Prerequisite*.

Primarily, we propose the use of independence atoms to restrict the set of possible relations to those considered semantically meaningful for the given application domain. In this sense, updates would only be allowed if they result in a relation that satisfies all the independence atoms declared on the schema. For efficient updates it is therefore important to eliminate redundant independence atoms from those that need to be validated whenever updates occur. Naturally, this leads us to the implication problem: given a set $\Sigma \cup \{\varphi\}$ of independence atoms, does every relation that satisfies all elements in Σ also satisfy φ? If that is true, then φ is redundant since validating that the updated relation satisfies all elements in Σ guarantees that it also satisfies φ. If it is false, then it must also be validated that φ is satisfied after the update. In our example, $S \perp P$ implies $P \perp S$. Hence, we do not need to validate $P \perp S$ explicitly, since it is already validated implicitly by validating $S \perp P$ explicitly. Efficient solutions to the implication problem for classes of embedded data dependencies are particularly important for efficient updates. Indeed, validating independence atoms is rather costly due to the high amount of redundancy they cause relations to exhibit. However, in contrast to full dependencies, such as multivalued dependencies, it is yet unknown how to effectively avoid data redundancy caused by embedded dependencies. For multivalued dependencies, larger database schemata can be decomposed into smaller schemata that satisfy Fagin's Fourth Normal Form condition, which characterizes the absence of data redundancy under such dependencies [3]. For now, therefore, the primary practical impact of solving the implication problem efficiently is an effective way of avoiding redundant independence atoms.

Besides increased consistency and integrity, independence atoms can be exploited for other important data processing tasks. In query optimization, for example, they can be used to avoid expensive database operations to return query answers more efficiently. For instance, knowing that the independence atom $S \perp P$ is implied by the atoms that relations are validated against, the query on the right can be used instead of the query on the left - both returning

all combinations of students and prerequisites that occur in the relation. The right query does not use an expensive join unlike the left query.

```
SELECT E.Student, E'.Prerequisite
FROM Enrol AS E, Enrol AS E'
```

```
SELECT E.Student, E.Prerequiste
FROM Enrol AS E
```

Contributions. Motivated by these benefits we investigate the implication problem of independence atoms in database relations. Our work is inspired by and based on [7] where, however, the basic setup is one of random variables. In particular, we show that the results of [7] can be transferred from the context of random variables to the context of databases. Our first contribution is an axiomatization of the implication problem by a finite set of Horn rules. In particular, for each atom φ which cannot be inferred by our inference rules from the atoms in Σ, we construct a finite relation r_φ that satisfies Σ and violates φ. This also shows that finite and unrestricted implication problem coincide for the class of independence atoms. Exploiting our axiomatization we establish an algorithm that decides the implication problem in $\mathcal{O}(|\Sigma| \cdot ||\varphi||^2 + |\Sigma| \cdot ||\Sigma \cup \{\varphi\}||)$ time, where $|\Sigma|$ denotes the number of atoms in Σ, and $||\Sigma||$ denotes the number of attributes in Σ. Finally, we show that the implication problem of independence atoms can be reduced to the model checking problem on a single relation. For that purpose, we show how to construct for an arbitrarily given set Σ of independence atoms a relation that satisfies all the elements of Σ and violates every independence atom not implied by Σ. In the literature such relations are known as Armstrong relations [4]. Hence, checking whether φ is implied by Σ amounts to checking whether an Armstrong relation for Σ satisfies φ. Inspecting Armstrong relations is likely to increase the number of data dependencies that business analysts discover to be meaningful for a given application domain [12].

Organization. We summarize related work in Section 2, providing further motivation for the study of independence atoms and relating them to existing work in probability theory and artificial intelligence. We introduce independence atoms and their associated implication problem in Section 3. Axiomatic and algorithmic characterizations of the implication problem are established in Sections 4 and 5, respectively. The construction of Armstrong relations is shown in Section 6. We conclude in Section 7 where we also comment on future work.

2 Related Work

Approximately 100 different classes of relational data dependencies have been studied in the research literature [22]. The expressivity of embedded multivalued dependencies results in the non-axiomatizability of its implication problem by a finite set of Horn rules [16] and its undecidability [11]. Multivalued dependencies [3] form an efficient sub-class of embedded multivalued dependencies, whose implication problem has been characterized by a finite axiomatization of Horn rules [1], by an almost linear time algorithm [5], and by a fragment of Boolean propositional logic [19]. These results have recently been generalized to multivalued

dependencies over SQL databases [10]. Multivalued dependencies are very special embedded dependencies, called full dependencies, as their attributes cover the full set of attributes of the underlying relation schema. Our results show that independence atoms form another efficient sub-class of embedded multivalued dependencies. In contrast to multivalued dependencies, independence atoms are not full dependencies. Given the vast amount of literature on data dependencies, given that independence is a natural concept in many areas, and given the outlined benefits of independence atoms, it is surprising that their investigation in the database literature is rather limited [15,20][1].

In [20] Sagiv and Walecka introduce the class of *subset dependencies* which are generalizations of embedded multivalued dependencies. A subset dependency $Z(X) \subseteq Z(Y)$ for attribute sets X, Y, Z of R, where both X and Y are disjoint from Z, is satisfied by some relation r over R, if for all tuples $t_1, t_2 \in r$ that agree on all attributes in X there is some tuple $t_3 \in r$ that agrees with t_1 on all attributes in Y and that agrees with t_2 on all attributes in Z. In particular, the independence atom $Y \perp Z$ is satisfied by r if and only if the subset dependency $Z(\emptyset) \subseteq Z(Y)$ is satisfied by r. The authors establish a finite axiomatization for the class of Z-subset dependencies, which, for all relation schemata R and some fixed set $Z \subseteq R$, consists of the subset dependencies $Z(X) \subseteq Z(Y)$. It follows from the definitions that Z-subset dependencies and independence atoms are different classes of embedded multivalued dependencies.

In [15], Paredaens investigates, among three other classes of data dependencies, so called *crosses* $X \times Y$ which are equivalent to independence atoms $X \perp Y$ where both X and Y are non-empty. Paredaens establishes a finite axiomatization for crosses, including the symmetry and decomposition rules, and a somewhat convoluted version of the exchange rule. If empty attribute sets are excluded, then our axiomatization is equivalent to that for crosses, as one would expect. The motivation for our axiomatization comes from its strong analogy to the Geiger-Paz-Pearl axioms for independence atoms over probability distributions [7], in particular the simplicity of the exchange rule. Indeed, Paradaens' version of the exchange rule can be derived from the symmetry, decomposition and our exchange rule. Based on our motivation from the introduction we were also interested in algorithmic solutions to the implication problem, and Armstrong relations, which Paradaens did not aim to address.

As indicated, our study is further motivated by the existing studies of conditional independence in artificial intelligence and statistics. Here, the implication problem of conditional independence atoms is known to be not axiomatizable by a finite set of Horn rules, and to be different from that of embedded multivalued dependencies [21]. The implication problem of saturated conditional independence atoms is axiomatizable by a finite set of Horn rules, equivalent to the implication problem of multivalued dependencies, and thus equivalent to that of a fragment in Boolean propositional logic and decidable in almost linear time [2,6]. In contrast to databases, independence atoms have been investigated in

[1] We would like to thank the anonymous reviewers who pointed us to the paper by Sagiv and Walecka [20], which pointed us to the paper by Paredaens [15].

probability theory. Indeed, their implication problem over discrete probability measures has been axiomatized by a finite set of Horn rules, and can be decided in low-degree polynomial time [7]. Furthermore, probability distributions can be constructed that satisfy a given set of probabilistic independence atoms and violate all those probabilistic independence atoms not implied by the given set [7]. Therefore, our paper establishes results for independence atoms over database relations that correspond to those known for independence atoms over probability distributions. They further show that reasoning about probabilistic independence atoms does not require probabilities at all.

3 Independence Atoms

In this section we first summarize basic concepts from the relational model of data, and then introduce the syntax and semantics of independence atoms, as well as their associated implication problem.

Let $\mathfrak{A} = \{A_1, A_2, \ldots\}$ be a (countably) infinite set of symbols, called *attributes*. A *relation schema* is a finite set $R = \{A_1, \ldots, A_n\}$ of attributes from $\mathfrak{A}$. Each attribute A of a relation schema is associated with a domain $dom(A)$ which represents the set of possible values that can occur in the column named A. A *tuple* over R is a function $t : R \to \bigcup_{A \in R} dom(A)$ with $t(A) \in dom(A)$ for all $A \in R$. For $X \subseteq R$ let $t(X)$ denote the restriction of the tuple t over R on X, and $dom(X) = \prod_{A \in X} dom(A)$ the Cartesian product of the domains of attributes in X. A *relation* r over R is a finite set of tuples over R. Let $r(X) = \{t(X) \mid t \in r\}$ denote the *projection* of the relation r over R on $X \subseteq R$. For attribute sets X and Y we often write XY for their set union $X \cup Y$. For disjoint subsets $X, Y \subseteq R$, $r_1 \subseteq dom(X)$ and $r_2 \subseteq dom(Y)$ let $r_1 \times r_2 = \{t \in dom(XY) \mid \exists t_1 \in r_1, t_2 \in r_2(t(X) = t_1(X) \wedge t(Y) = t_2(Y))\}$ denote the *Cartesian product* of r_1 and r_2.

3.1 Syntax and Semantics

Intuitively, an attribute set X is independent of a disjoint attribute set Y, if X-values occur independently of Y-values. That is, the independence holds in a relation, if every X-value that occurs in the relation occurs together with every Y-value that occurs in the relation. Therefore, we arrive at the following concept, in analogy with the similar concept in so-called team semantics [8]:

Definition 1. *An* independence atom *over relation schema R is an expression $X \perp Y$ where X and Y are two disjoint subsets of R. A relation r over R is said to* satisfy *the independence atom $X \perp Y$ over R if and only if for all $t_1, t_2 \in r$ there is some $t \in r$ such that $t(X) = t_1(X)$ and $t(Y) = t_2(Y)$. If r does not satisfy $X \perp Y$, then we also say that r* violates *$X \perp Y$.*

The semantics of independence atoms can be stated explicitly as that of an embedded dependency. In the context of the attribute set XY, the concept represented by X is independent of the concept represented by Y.

Proposition 1. *Let r be a relation, and $X \perp Y$ an independence atom over relation schema R. Then r satisfies $X \perp Y$ if and only if $r(XY) = r(X) \times r(Y)$.* □

Proposition 1 captures the equivalence of independence atoms to crosses [15]. Rissanen shows in [18] that a relation r over relation schema R which satisfies a functional dependency $X \to Y$ is the lossless join of its projections $r(XY)$ and $r(X(R - XY))$. Proposition 1 shows that for a relation r which satisfies the independence atom $X \perp Y$, the projection $r(XY)$ is the lossless Cartesian product of the projections $r(X)$ and $r(Y)$.

We illustrate the semantics of independence atoms on our running example.

Example 2. The projection of relation r from Example 1 on *Student* and *Prerequisite* is the Cartesian product of the projection on *Student* and the projection on *Prerequisite*. Hence, r satisfies *Student* $\perp$ *Prerequisite*. However, the projection of r on *Student* and *Year* is not the Cartesian product of the projection on *Student* and the projection on *Year*. Thus, r violates *Student* $\perp$ *Year*.

3.2 The Implication Problem

Data dependencies are usually defined as semantic constraints that restrict the possible relations of a schema to those considered meaningful for a given application domain. Relations that satisfy all those data dependencies that express "business rules" of the domain are considered meaningful, relations that violate some business rule are considered meaningless. For efficient data processing it is therefore important to minimize the time that it takes to validate whether a relation satisfies the given set of data dependencies. Indeed, relations that satisfy a given set of data dependencies do not need to be tested whether they satisfy any data dependency that is implied by the given set. Therefore, it is essential to efficiently decide the implication problem of data dependencies. We will now define this problem for independence atoms.

For a set $\Sigma \cup \{\varphi\}$ of independence atoms we say that Σ *implies* φ, or that φ *is implied* by Σ, written $\Sigma \models \varphi$, if every relation that satisfies every element in Σ also satisfies φ. For a set Σ of independence atoms over some fixed relation schema R, we let $\Sigma^* = \{\varphi \mid \Sigma \models \varphi\}$ be the *semantic closure* of Σ, i.e., the set of all independence atoms implied by Σ. In order to determine the implied independence atoms we use a syntactic approach by applying inference rules. These inference rules have the form $\frac{\text{premise}}{\text{conclusion}}$ and inference rules without any premise are called axioms. An inference rule is called *sound*, if the independence atoms in the premise of the rule imply the independence atom in the conclusion of the rule. We let $\Sigma \vdash_{\mathfrak{R}} \varphi$ denote the *inference* of φ from Σ by the set $\mathfrak{R}$ of inference rules. That is, there is some sequence $\gamma = [\sigma_1, \ldots, \sigma_n]$ of independence atoms such that $\sigma_n = \varphi$ and every σ_i is an element of Σ or results from an application of an inference rule in $\mathfrak{R}$ to some elements in $\{\sigma_1, \ldots, \sigma_{i-1}\}$. For Σ, let $\Sigma^+_{\mathfrak{R}} = \{\varphi \mid \Sigma \vdash_{\mathfrak{R}} \varphi\}$ be its *syntactic closure* under inferences by $\mathfrak{R}$. A set $\mathfrak{R}$ of inference rules is said to be *sound* (*complete*) for the implication of independence atoms, if for every R and for every set Σ of independence atoms

over R we have $\Sigma^+_{\mathfrak{R}} \subseteq \Sigma^*$ ($\Sigma^* \subseteq \Sigma^+_{\mathfrak{R}}$). The (finite) set $\mathfrak{R}$ is said to be a (finite) *axiomatization* for the implication of independence atoms if $\mathfrak{R}$ is both sound and complete. The implication problem of independence atoms is defined as follows.

PROBLEM:	Implication problem for independence atoms
INPUT:	Relation schema R,
	Set $\Sigma \cup \{\varphi\}$ of independence atoms over R
OUTPUT:	Yes, if $\Sigma \models \varphi$; No, otherwise

We illustrate the implication problem on our running example.

Example 3. Continuing Example 1, the set Σ of independence atoms consisted of *Student* $\perp$ *Prerequisite*. If φ denotes the independence atom *Student* $\perp$ *Year*, then the relation r from Example 1 shows that $\Sigma \not\models \varphi$. An example of an independence atom that is implied by Σ is *Prerequisite* $\perp$ *Student*.

4 Axiomatic Characterization

In this section we establish an axiomatic characterization of the implication problem for independence atoms over relations. In fact, we show that the set $\mathfrak{I}$ of inference rules from Table 1 is sound and complete. The set $\mathfrak{I}$ of inference rules is the same set used in [7] to axiomatize implication among independence atoms in the framework of random variables. It is remarkable that the same axioms have found their way also to the study of concurrency [13] and secrecy [14]. If the attribute sets X, Y and Z are interpreted as sets of vectors in a vector space, the rules $\mathfrak{I}$ would govern the concept of linear (as well as algebraic) independence as was noted already in [23,24].

Table 1. Axiomatization $\mathfrak{I}$ of Independence in Database Relations

$\dfrac{}{X \perp \emptyset}$ (trivial independence, $\mathcal{T}$)	$\dfrac{X \perp Y}{Y \perp X}$ (symmetry, $\mathcal{S}$)
$\dfrac{X \perp YZ}{X \perp Y}$ (decomposition, $\mathcal{D}$)	$\dfrac{X \perp Y \quad XY \perp Z}{X \perp YZ}$ (exchange, $\mathcal{E}$)

Using Definition 1 it is not difficult to show the soundness of the inference rules in $\mathfrak{I}$ for the implication of independence atoms. As an example, we prove the soundness of the exchange rule $\mathcal{E}$. Let r be a relation that satisfies the independence atoms $X \perp Y$ and $XY \perp Z$. Let $t_1, t_2 \in r$. Then there is some tuple $\bar{t} \in r$ such that $\bar{t}(X) = t_1(X)$ and $\bar{t}(Y) = t_2(Y)$, since r satisfies $X \perp Y$. Since r satisfies $XY \perp Z$, for $\bar{t}, t_2 \in r$ there must be some $t \in r$ such that $t(XY) = \bar{t}(XY)$ and $t(Z) = t_2(Z)$. In particular, $t(X) = \bar{t}(X) = t_1(X)$, $t(Y) = \bar{t}(Y) = t_2(Y)$,

and $t(Z) = t_2(Z)$. Hence, there is some $t \in r$ such that $t(X) = t_1(X)$ and $t(YZ) = t_2(YZ)$. That is, r also satisfies the independence atom $X \perp YZ$.

The soundness of the rules in $\mathfrak{I}$ allows us to mechanically infer several implied independence atoms.

Example 4. Recall that $\Sigma = \{Student \perp Prerequisite\}$ in our running example. A single application of the symmetry rule $\mathcal{S}$ to $Student \perp Prerequisite$ gives us the independence atom $Prerequisite \perp Student \in \Sigma^+_{\mathfrak{I}}$. Consequently, $Prerequisite \perp Student \in \Sigma^*$ due to the soundness of the symmetry rule.

The inference rules in $\mathfrak{I}$ are also complete. That is, every implied independence atom can be inferred by applications of the inference rules in $\mathfrak{I}$. The following theorem is like Theorem 3 of [7]:

Theorem 1. *The set $\mathfrak{I}$ of Horn rules forms a finite axiomatization for the class of independence atoms.*

Proof. We proceed as in [7, Theorem 3], but working with relations instead of random variables. Let R be some relation schema and Σ a set of independence atoms over R. Let $\varphi = X \perp Y \notin \Sigma^+_{\mathfrak{I}}$. Without loss of generality we assume that for all non-empty sets $X' \subseteq X$ and $Y' \subseteq Y$ with $X'Y' \neq XY$, $X' \perp Y' \in \Sigma^+_{\mathfrak{I}}$ holds. An independence atom φ with these properties is called *minimal*. Indeed, if $\varphi = X \perp Y$ is not minimal, then we can remove attributes from X or from Y to obtain a minimal atom $\varphi' = X' \perp Y' \notin \Sigma^+_{\mathfrak{I}}$. Note that, if X' and Y' are both singletons, then $X' \perp Y'$ is a minimal atom due to the trivial independence axiom $\mathcal{T}$. For each minimal atom φ' we construct a relation $r_{\varphi'}$ that satisfies Σ and violates φ'. Due to the decomposition rule $\mathcal{D}$, $r_{\varphi'}$ also violates φ and, hence, φ is not implied by Σ.

Let $\varphi = X \perp Y \notin \Sigma^+_{\mathfrak{I}}$ be a minimal atom. For all $A \in R$ assume that $dom(A) = \{0, 1\}$, and let $Z = R - XY$. Let $A_0 \in X$. Define $r_\varphi \subseteq dom(R)$ as follows: for all $t \in dom(R)$ we have,

$$t \in r_\varphi \text{ if and only if } t(A_0) = \textstyle\sum_{A \in (X - A_0)Y} t(A) \mod 2 \,.$$

Clearly, $r = r(XY) \times \prod_{A \in Z} dom(A)$.

We show first that r_φ violates the independence atom $X \perp Y$. Let t be a tuple where $t(A_0) = 1$ and $t(A) = 0$ for all $A \in XY - A_0$. Then $t \in r(X) \times r(Y)$, but $t \notin r(XY)$.

It remains to show that r_φ satisfies every independence atom $V \perp W \in \Sigma$.

Case 1. Assume that $V \subseteq Z$ or $W \subseteq Z$. Say, for example, that $V \subseteq Z$. By construction, for every tuple $t_1 \in r(Z)$ and every tuple in $t_2 \in r(W)$ there is some tuple $t \in r(VW)$ such that $t(V) = t_1(V)$ and $t(W) = t_2(W)$. The case where $W \subseteq Z$ holds is similar. Hence, $r(VW) = r(V) \times r(W)$.

Case 2. Assume that $V \cap XY \neq \emptyset$ and $W \cap XY \neq \emptyset$.

Case 2.1. Suppose $XY \not\subseteq VW$. For $U \subseteq XY$ with $U \neq XY$ we have $r(U) = \prod_{A \in U} r(A)$. Hence, $r(VW) = \prod_{A \in VW} r(A)$. In particular, $r(VW) = r(V) \times r(W)$.

Case 2.2. Suppose $XY \subseteq VW$. Then let $V = X'Y'Z'$, $W = X''Y''Z''$ where $X = X'X''$, $Y = Y'Y''$, and $Z'Z'' \subseteq Z$ holds. Assume that $V \perp W \in \Sigma_{\mathfrak{J}}^+$. We show, under this assumption, the contradiction that $X \perp Y \in \Sigma_{\mathfrak{J}}^+$ holds. Consequently, $V \perp W \notin \Sigma_{\mathfrak{J}}^+$ and this case cannot occur.

Since $X \perp Y$ is a minimal independence atom, $X' \perp Y', X'' \perp Y \in \Sigma_{\mathfrak{J}}^+$. The inference

$$\dfrac{\dfrac{X'' \perp Y}{\mathcal{D}:\ Y \perp X''} \quad \dfrac{X' \perp Y' \quad \dfrac{X'Y'Z' \perp X''Y''Z''}{\mathcal{D}:\ X'Y' \perp X''Y''}}{\mathcal{E}:\quad X' \perp X''Y}}{\dfrac{\mathcal{S}:\quad X''Y \perp X'}{\mathcal{E}:\quad Y \perp X}}{\mathcal{S}:\quad X \perp Y}$$

gives the anticipated contradiction that $X \perp Y \in \Sigma_{\mathfrak{J}}^+$ under the assumption that $V \perp W \in \Sigma_{\mathfrak{J}}^+$ when $XY \subseteq VW$. Note that the inference of $X \perp Y \in \Sigma_{\mathfrak{J}}^+$ remains valid even if some of the X', X'', Y', Y'' are empty, as long as $X = X'X''$ and $Y = Y'Y''$ hold. □

We illustrate the completeness argument on our running example.

Example 5. Let $\Sigma = \{Student \perp Prerequisite\}$ be a set of independence atoms and $\varphi = Prerequiste \perp Year$ be an independence atom over Enrol. The construction from the completeness proof of Theorem 1 may result in the relation on the left, which may result in the relation on the right by suitable substitutions.

Student	*Prerequiste*	*Year*
S1	P1	Y1
S1	P2	Y2
S2	P1	Y1
S2	P2	Y2

Student	*Prerequiste*	*Year*
Hilbert	Phil101	1900
Hilbert	Phys110	1905
Ackermann	Phil101	1900
Ackermann	Phys110	1905

Both relations satisfy Σ and violate φ.

Instead of the exchange rule $\mathcal{E}$, Paredaens used the following rule on the left

$$\dfrac{X' \perp Z \quad X \perp Y}{X \cap X' \perp Y \cup (X \cap Z)}\, X \cap X' \neq \emptyset \qquad \dfrac{UV \perp WW' \quad UWW'' \perp Y}{U \perp YW}$$

for crosses, defined for non-empty attribute sets. This rule produces an independence atom that cannot already be inferred by the decomposition and symmetry rules alone, if $X \cap X' \neq \emptyset$ and $X \cap Z \neq \emptyset$ hold. In that case, the rule can be rewritten as the rule on the right above. This rule can be inferred as follows

$$\dfrac{\dfrac{\dfrac{\dfrac{UV \perp WW'}{\mathcal{D}:\ UV \perp W}}{\mathcal{S}:\ W \perp UV}}{\mathcal{D}:\ W \perp U} \quad \dfrac{UWW'' \perp Y}{\mathcal{D}:\ UW \perp Y}}{\mathcal{S}:\ U \perp W}}{\mathcal{E}:\quad U \perp YW}$$

from the decomposition, symmetry, and exchange rule.

5 Algorithmic Characterization

We establish an algorithmic characterization of the implication problem. In practice, one may simply want to check if a single independence atom φ is implied by a given set Σ of independence atoms. One could compute $\Sigma^* = \Sigma_{\mathfrak{J}}^+$ and check if $\varphi \in \Sigma^*$. However, this algorithm is hardly efficient. Instead, we exploit the extra knowledge about φ to decide more efficiently if φ is implied by Σ.

The divide-and-conquer algorithm is presented as Algorithm 1. On input $(\Sigma, X \perp Y)$ it reduces Σ to $\Sigma' = \Sigma[XY] = \{(V \cap XY) \perp (W \cap XY) \mid V \perp W \in \Sigma\}$ (line 3). If $X \perp Y$ is a trivial independence atom, i.e., if one of its sets is empty, or if the atom or its symmetric atom is included in Σ', then the algorithm returns `true` (line 4-5). If there is no non-trivial atom $U \perp V \in \Sigma'$ where $UV = XY$, then the algorithm returns `false` (line 7-8). Otherwise, there is some non-trivial atom $U \perp V \in \Sigma'$ where $U = QR$, $V = ST$, and $X = QS$, $Y = RT$. In this case, the Algorithm returns `true` if and only if it returns `true` on both inputs $(\Sigma', Q \perp R)$ and $(\Sigma', S \perp T)$ (line 12).

Algorithm 1. Implication

```
 1: procedure IMPLIED(Σ, φ)
 2:     φ ← X ⊥ Y;
 3:     Σ' ← Σ[XY];
 4:     if X = ∅ or Y = ∅ or X ⊥ Y ∈ Σ' or Y ⊥ X ∈ Σ' then
 5:         IMPLIED(Σ, φ) ← true;
 6:     end if;
 7:     if for all U ⊥ V ∈ Σ' with U ≠ ∅ and V ≠ ∅, UV ≠ XY then
 8:         IMPLIED(Σ, φ) ← false;
 9:     else          ▷ ∃U ⊥ V ∈ Σ' with ∅ ≠ U = QR, ∅ ≠ V = ST, X = QS, Y = RT
10:         φ1 ← Q ⊥ R;
11:         φ2 ← S ⊥ T;
12:         IMPLIED(Σ, φ) ← IMPLIED(Σ', φ1) ∧ IMPLIED(Σ', φ2);
13:     end if;
14: end procedure
```

Algorithm 1 works correctly in low-degree polynomial time. This yields the following result, reminiscent of Theorems 8 and 9 of [7]:

Theorem 2. *Algorithm 1 terminates, and* $\textsc{Implied}(\Sigma, \varphi) =$ `true` *if and only if* $\Sigma \models \varphi$. *The time-complexity of Algorithm 1, on input* (Σ, φ), *is in* $\mathcal{O}(|\Sigma| \cdot ||\varphi||^2 + |\Sigma| \cdot ||\Sigma \cup \{\varphi\}||)$.

Proof (Sketch). Let $\varphi = X \perp Y$. Firstly, it follows from an inspection of the inference rules in $\mathfrak{I}$ that $\Sigma \vdash_{\mathfrak{I}} \varphi$ holds if and only if $\Sigma' \vdash_{\mathfrak{I}} \varphi$ holds.

Secondly, for any non-trivial φ, $\Sigma' \vdash_{\mathfrak{I}} \varphi$ holds only if there is some atom $U \perp V \in \Sigma'$ such that $UV = XY$. This follows from the observation that no inference rule in $\mathfrak{I}$ introduces an attribute to its conclusion that does not already occur in one of its premises.

These observations justify lines 2-7 of Algorithm 1. We will now justify lines 9-12. Let $X = QS$, $Y = RT$, $U = QR$, and $V = ST$. Then we show that, if $U \perp V \in \Sigma_{\mathfrak{J}}^+$, then $X \perp Y \in \Sigma_{\mathfrak{J}}^+$ if and only if $Q \perp R \in \Sigma[QR]_{\mathfrak{J}}^+$ and $S \perp T \in \Sigma[ST]_{\mathfrak{J}}^+$.

Assume first that $QR \perp ST, Q \perp R, S \perp T \in \Sigma_{\mathfrak{J}}^+$. Then the following inference shows that $QS \perp RT \in \Sigma_{\mathfrak{J}}^+$, too.

$$\dfrac{\dfrac{\dfrac{S \perp T \qquad \dfrac{QR \perp ST}{\mathcal{S}:\ ST \perp QR}}{\mathcal{E}:\ S \perp QRT}}{\mathcal{D}:\ S \perp RT} \qquad \dfrac{Q \perp R \qquad QR \perp ST}{\mathcal{E}:\ Q \perp RST}}{\mathcal{S}:\ RT \perp S} \qquad \dfrac{}{\mathcal{S}:\ SRT \perp Q}$$

$$\dfrac{\mathcal{E}:\ RT \perp QS}{\mathcal{S}:\ QS \perp RT}$$

If $QS \perp RT \in \Sigma_{\mathfrak{J}}^+$, then there is an inference $U_1 \perp V_1, \ldots, U_k \perp V_k = QS \perp RT$ from Σ. Consequently, $(U_1 \cap QR) \perp (V_1 \cap QR), \ldots, (U_k \cap QR) \perp (V_k \cap QR) = Q \perp R$ is an inference of $Q \perp R$ from $\Sigma[QR]$. Similarly, an inference of $S \perp T$ from $\Sigma[ST]$ can be constructed from an inference of $QS \perp RT$ from Σ.

Note that this shows, in particular, that a selection of $U \perp V$ in line 9 can be made arbitrarily since any selection provides a necessary and sufficient means to check whether $X \perp Y \in \Sigma_{\mathfrak{J}}^+$.

Algorithm 1 terminates since the size of the independence atoms strictly decreases in line 12. If the algorithm did not terminate before, it will terminate when the number of attributes in the two atoms have reached 2 (line 4 or line 7). The first statement of Theorem 2 follows from a simple induction on the number of attributes in φ.

We will now analyze the time complexity of Algorithm 1. The complexity is measured in terms of two types of basic operations: the comparison of two independence atoms and the projection of independence atoms. Both operations are bounded by the number $||\Sigma \cup \{\varphi\}||$ of distinct attributes in $\Sigma \cup \{\varphi\}$. Let $c(\varphi)$ denote the number of basic operations required to solve a problem for an independence atom φ, and assume for now that the distinct attributes in Σ are those in φ. By line 12, $c(\varphi)$ must satisfy the equation $c(\varphi) \leq c(\varphi_1) + c(\varphi_2) + |\Sigma|$, where $|\Sigma|$ denotes the number of atoms in Σ, and where $||\varphi|| = ||\varphi_1|| + ||\varphi_2||$. The solution to this equation is $\mathcal{O}(|\Sigma| \cdot ||\varphi||)$ measured in basic operations. Adding the cost of projecting Σ to the attributes in φ is in $\mathcal{O}(|\Sigma| \cdot ||\Sigma \cup \{\varphi\}||)$. □

Example 6. Let $\Sigma = \{\mathit{Student} \perp \mathit{Prerequisite}\}$ be a set of independence atoms and $\varphi = \mathit{Prerequiste} \perp \mathit{Year}$ be an independence atom over relation schema Enrol. On input (Σ, φ), Algorithm 1 computes $\Sigma' = \emptyset$ in Step 3, and returns Implied$(\Sigma, \varphi) =$ `false` in Step 8, since the condition in Step 7 is trivially satisfied. Hence, by Theorem 2, φ is not implied by Σ.

6 Armstrong Relations

We show that independence atoms enjoy Armstrong relations. That is, for every relation schema R and every set Σ of independence atoms, there is a relation over R that satisfies Σ and violates every independence atom not implied by Σ. The property of enjoying Armstrong relations has been characterized by Fagin in a very general framework [4]. We will exploit this characterization to show how to construct Armstrong relations for independence atoms.

Theorem 3. *[4] Let $\mathcal{S}$ denote a set of sentences. The following properties of $\mathcal{S}$ are equivalent:*

1. Existence of a faithful operator. *There exists an operator $\otimes$ that maps non-empty families of models into models, such that if σ is a sentence in $\mathcal{S}$ and $\langle P_i : i \in I\rangle$ is a non-empty family of models, then σ holds for $\otimes\langle P_i : i \in I\rangle$ if and only if σ holds for each P_i.*
2. Existence of Armstrong models. *Whenever Σ is a consistent subset of $\mathcal{S}$ and Σ^* is the set of sentences in $\mathcal{S}$ that are logical consequences of Σ, then there exists a model (an "Armstrong" model) that obeys Σ^* and no other sentence in $\mathcal{S}$.*
3. Splitting of disjunctions. *Whenever Σ is a subset of $\mathcal{S}$ and $\{\sigma_i : i \in I\}$ is a non-empty subset of $\mathcal{S}$, then $\Sigma \models \bigvee\{\sigma_i : i \in I\}$ if and only if there exists some $i \in I$ such that $\Sigma \models \sigma_i$.* □

Indeed, there is a faithful operator for independence atoms. While Fagin's theorem holds for any cardinality of I [4], we use it only for finite non-empty I. An analog of the below theorem was proved in [7, Theorem 11] for distributions.

Theorem 4. *Let $\{r_i : i = 1, \ldots, n\}$ be a finite set of relations. There exists an operation $\otimes$ that maps finite sets of relations to relations such that for each independence atom σ, the relation $\otimes\{r_i : i = 1, \ldots, n\}$ satisfies σ if and only if for $i = 1, \ldots, n$, r_i satisfies σ.*

Proof. We construct the operation $\otimes$ by using a binary operation $\otimes_b$ such that for every independence atom σ, the relation $r_1 \otimes_b r_2$ satisfies σ if and only if both relations r_1 and r_2 satisfy σ. The operation $\otimes$ is then defined recursively by $\otimes\{r_i : i = 1, \ldots, n\} := (\cdots((r_1 \otimes_b r_2) \otimes_b r_3) \cdots \otimes_b r_n)$. Let r_1, r_2 be relations over relation schema R. Then $r_1 \otimes_b r_2$ is defined by

$$((a_1, a_1'), \ldots, (a_n, a_n')) \in r_1 \otimes_b r_2 \text{ iff } (a_1, \ldots, a_n) \in r_1 \text{ and } (a_1', \ldots, a_n') \in r_2.$$

We now show that for an independence atom $X \perp Y$ over R we have, $r_1 \otimes_b r_2$ satisfies $X \perp Y$ if and only if r_1 satisfies $X \perp Y$ and r_2 satisfies $X \perp Y$.

We show first that if r_1 satisfies $X \perp Y$ and r_2 satisfies $X \perp Y$, then $r_1 \otimes_b r_2$ satisfies $X \perp Y$. Let $t_1 = ((a_1, a_1'), \ldots, (a_n, a_n')), t_2 = ((b_1, b_1'), \ldots, (b_n, b_n')) \in r_1 \otimes_b r_2$. Then, $t_1^1 = (a_1, \ldots, a_n), t_2^1 = (b_1, \ldots, b_n) \in r_1$, and $t_1^2 = (a_1', \ldots, a_n'), t_2^2 = (b_1', \ldots, b_n') \in r_2$. Since r_1 satisfies $X \perp Y$ there is some $\bar{t} = (c_1, \ldots, c_n) \in r_1$ such that $\bar{t} = t_1^1(X)$ and $\bar{t} = t_2^1(Y)$. Since r_2 satisfies $X \perp Y$ there is some

$t' = (c'_1, \ldots, c'_n) \in r_2$ such that $t'(X) = t^2_1(X)$ and $t' = t^2_2(Y)$. Let $t := ((c_1, c'_1), \ldots, (c_n, c'_n)) \in r_1 \otimes_b r_2$. It follows that $t(X) = t_1(X)$ and $t(Y) = t_2(Y)$. Hence, $r_1 \otimes_b r_2$ satisfies $X \perp Y$.

It remains to show that if $r_1 \otimes_b r_2$ satisfies $X \perp Y$, then r_1 satisfies $X \perp Y$ and r_2 satisfies $X \perp Y$. Let $t^1_1 = (a_1, \ldots, a_n), t^1_2 = (b_1, \ldots, b_n) \in r_1$ and $t^2_1 = (a'_1, \ldots, a'_n), t^2_2 = (b'_1, \ldots, b'_n) \in r_2$. Then, $t_1 = ((a_1, a'_1), \ldots, (a_n, a'_n)), t_2 = ((b_1, b'_1), \ldots, (b_n, b'_n)) \in r_1 \otimes_b r_2$. Since $r_1 \otimes_b r_2$ satisfies $X \perp Y$ there is some $t = ((c_1, c'_1), \ldots, (c_n, c'_n)) \in r_1 \otimes r_2$ where $t(X) = t_1(X)$ and $t(Y) = t_2(Y)$. Then $t^{r_1} := (c_1, \ldots, c_n)$ satisfies $t^{r_1}(X) = t^1_1(X)$ and $t^{r_1}(Y) = t^1_2(Y)$. Thus, r_1 satisfies $X \perp Y$. Similarly, $t^{r_2} := (c'_1, \ldots, c'_n)$ satisfies $t^{r_2}(X) = t^2_1(X)$ and $t^{r_2}(Y) = t^2_2(Y)$. Hence, r_2 satisfies $X \perp Y$. □

We illustrate the construction of Armstrong relations on our example.

Example 7. Let $\Sigma = \{Student \perp Prerequisite\}$ be a set of independence atoms over ENROL. From previous examples we have seen the relation r_1 on the left that satisfies Σ and $P \perp Y$, but violates $S \perp Y$, and the relation r_2 on the right that satisfies Σ and $S \perp Y$, but violates $P \perp Y$.

Student	*Prerequiste*	*Year*
S1	P1	Y1
S2	P1	Y2
S1	P2	Y1
S2	P2	Y2

Student	*Prerequiste*	*Year*
S3	P3	Y3
S3	P4	Y4
S4	P3	Y3
S4	P4	Y4

The Armstrong construction results in the relation $r_1 \otimes_b r_2$, defined by $((a_1, a'_1), \ldots, (a_n, a'_n)) \in r_1 \otimes_b r_2$ iff $(a_1, \ldots, a_n) \in r_1$ and $(a'_1, \ldots, a'_n) \in r_2$, on the left, and suitable substitutions yield the relation on the right.

Student	*Prerequiste*	*Year*
(S1,S3)	(P1,P3)	(Y1,Y3)
(S1,S3)	(P1,P4)	(Y1,Y4)
(S1,S4)	(P1,P3)	(Y1,Y3)
(S1,S4)	(P1,P4)	(Y1,Y4)
(S2,S3)	(P1,P3)	(Y2,Y3)
(S2,S3)	(P1,P4)	(Y2,Y4)
(S2,S4)	(P1,P3)	(Y2,Y3)
(S2,S4)	(P1,P4)	(Y2,Y4)
(S1,S3)	(P2,P3)	(Y1,Y3)
(S1,S3)	(P2,P4)	(Y1,Y4)
(S1,S4)	(P2,P3)	(Y1,Y3)
(S1,S4)	(P2,P4)	(Y1,Y4)
(S2,S3)	(P2,P3)	(Y2,Y3)
(S2,S3)	(P2,P4)	(Y2,Y4)
(S2,S4)	(P2,P3)	(Y2,Y3)
(S2,S4)	(P2,P4)	(Y2,Y4)

Student	*Prerequiste*	*Year*
Sheldon	Ethi101	2010
Sheldon	Logi120	2011
Leonard	Ethi101	2010
Leonard	Logi120	2011
Howard	Ethi101	2012
Howard	Logi120	2013
Raj	Ethi101	2012
Raj	Logi120	2013
Sheldon	Chem110	2010
Sheldon	Biol105	2011
Leonard	Chem110	2010
Leonard	Biol105	2011
Howard	Chem110	2012
Howard	Biol105	2013
Raj	Chem110	2012
Raj	Biol105	2013

Indeed the latter two relations are Armstrong relations for Σ. That is, they satisfy Σ and violate $S \perp Y$ and $P \perp Y$, and thereby also $S \perp PY$, $SY \perp P$, $SP \perp Y$, $S \perp YP$, and their symmetric independence atoms.

It can now be shown how an arbitrary set of independence atoms can be visualized as a single Armstrong relation. In practice, Armstrong relations can be used by database designers and business analysts as a communication tool to acquire and discuss the meaningfulness of business rules with domain experts [12]. Just as [7, Theorem 11] obtains for distributions, we obtain:

Theorem 5. *The class of independence atoms enjoys Armstrong relations.*

Proof. Let R be an arbitrary relation schema and Σ a set of independence atoms over R. By Theorem 1, for each $\varphi \notin \Sigma^+_{\mathfrak{I}}$ there is some relation r_φ that satisfies Σ and violates φ. Let $r := \otimes\{r_\varphi \mid \varphi \notin \Sigma^+_{\mathfrak{I}}\}$. The relation is well-defined since the set of all independence atoms over a relation schema is finite. According to Theorem 4, r satisfies all independence atoms in Σ and violates every independence atom not implied by Σ. □

It also follows from our results and Theorem 3 that the set $\mathfrak{I}$ of inference rules is powerful enough to infer all disjunctions of independence atoms that are logically implied by a set of independence atoms, and not merely single independence atoms.

7 Conclusion and Future Work

We investigated independence atoms, introduced in [8], as a new class of relational data dependencies. Our results show that independence atoms form an efficient sub-class of embedded multivalued dependencies whose implication problem is not finitely axiomatizable and undecidable. Our efficient solutions to the implication problem can result in enormous cost savings in data processing, for example when validating the consistency of update operations on relations, or when querying relations. Independence atoms form the database counterpart of probabilistic independence atoms known from probability theory.

In future work we plan to implement our algorithms as a tool, and analyze how the inspection of Armstrong relations can help database designers or business analysts with the task of identifying independence atoms that are semantically meaningful for a given application domain. It is interesting to investigate the minimum number of tuples required in Armstrong relations. It is also a challenging problem to identify means to reduce data redundancy caused by embedded dependencies. For the field of (in)dependence logic, it would be interesting to axiomatize the combined class of independence and dependence atoms.

Acknowledgement. This research is supported by the Marsden Fund Council from Government funding, administered by the Royal Society of New Zealand, and grants 264917 and 251557 of the Academy of Finland.

References

1. Beeri, C., Fagin, R., Howard, J.H.: A complete axiomatization for functional and multivalued dependencies in database relations. In: SIGMOD Conference, pp. 47–61. ACM (1977)

2. Dawid, A.P.: Conditional independence in statistical theory. Journal of the Royal Statistical Society. Series B (Methodological) 41(1), 1–31 (1979)
3. Fagin, R.: Multivalued dependencies and a new normal form for relational databases. ACM Trans. Database Syst. 2(3), 262–278 (1977)
4. Fagin, R.: Horn clauses and database dependencies. J. ACM 29(4), 952–985 (1982)
5. Galil, Z.: An almost linear-time algorithm for computing a dependency basis in a relational database. J. ACM 29(1), 96–102 (1982)
6. Geiger, D., Pearl, J.: Logical and algorithmic properties of conditional independence and graphical models. The Annals of Statistics 21(4), 2001–2021 (1993)
7. Geiger, D., Paz, A., Pearl, J.: Axioms and algorithms for inferences involving probabilistic independence. Inf. Comput. 91(1), 128–141 (1991)
8. Grädel, E., Väänänen, J.A.: Dependence and independence. Studia Logica 101(2), 399–410 (2013)
9. Halpern, J.: Reasoning about uncertainty. MIT Press (2005)
10. Hartmann, S., Link, S.: The implication problem of data dependencies over SQL table definitions. ACM Trans. Datab. Syst. 37(2), 13.1–13.52 (2012)
11. Herrmann, C.: On the undecidability of implications between embedded multivalued database dependencies. Inf. Comput. 204(12), 1847–1851 (2006)
12. Langeveldt, W., Link, S.: Empirical evidence for the usefulness of Armstrong relations on the acquisition of meaningful FDs. Inf. Syst. 35(3), 352–374 (2010)
13. More, S.M., Naumov, P., Sapp, B.: Concurrency Semantics for the Geiger-Paz-Pearl Axioms of Independence. In: CSL, vol. 12, pp. 443–457 (2011)
14. Naumov, P.: Independence in information spaces. Studia Logica 100(5), 953–973 (2012)
15. Paredaens, J.: The interaction of integrity constraints in an information system. J. Comput. Syst. Sci. 20(3), 310–329 (1980)
16. Parker Jr., D., Parsaye-Ghomi, K.: Inferences involving embedded multivalued dependencies and transitive dependencies. In: SIGMOD Conference, pp. 52–57 (1980)
17. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
18. Rissanen, J.: Independent components of relations. ACM Trans. Database Syst. 2(4), 317–325 (1977)
19. Sagiv, Y., Delobel, C., Parker Jr., D., Fagin, R.: An equivalence between relational database dependencies and a fragment of propositional logic. J. ACM 28(3), 435–453 (1981)
20. Sagiv, Y., Walecka, S.F.: Subset dependencies and a completeness result for a subclass of embedded multivalued dependencies. J. ACM 29(1), 103–117 (1982)
21. Studený, M.: Conditional independence relations have no finite complete characterization. In: Transactions of the 11th Prague Conference on Information Theory, pp. 377–396. Kluwer (1992)
22. Thalheim, B.: Dependencies in relational databases. Teubner (1991)
23. van der Waerden, B.L.: Moderne Algebra. J. Springer, Berlin (1940)
24. Whitney, H.: On the Abstract Properties of Linear Dependence. Amer. J. Math. 57(3), 509–533 (1935)