Approximation Logics for Subclasses of Probabilistic Conditional Independence and Hierarchical Dependence on Incomplete Data

Sebastian Link

Abstract Probabilistic conditional independence constitutes a principled approach to handle knowledge and uncertainty in artificial intelligence, and is fundamental in probability theory and multivariate statistics. Similarly, first-order hierarchical dependence provides an expressive framework to capture the semantics of an application domain within a database system, and is essential for the design of databases. For complete data it is well-known that the implication problem associated with probabilistic conditional independence is not axiomatizable by a finite set of Horn rules [56], and the implication problem for first-order hierarchical dependence is undecidable [30]. Moreover, both implication problems do not coincide [56] and neither of them is equivalent to the implication problem of some fragment of Boolean propositional logic [51]. In this article, generalized saturated conditional independence as well as full first-order hierarchical dependence over incomplete data are investigated as expressive subclasses of probabilistic conditional independence and first-order hierarchical dependence, respectively. The associated implication problems are axiomatized by a finite set of Horn rules, and both shown to coincide with that of a propositional fragment under interpretations in the well-known approximation logic $\mathscr{S}$-3. Here, the propositional variables in the set $\mathscr{S}$ are interpreted classically, and correspond to random variables as well as attributes on which incomplete data is not permitted to occur.

1 Introduction

The concept of conditional independence is important for capturing structural aspects of probability distributions, for dealing with knowledge and uncertainty in artificial intelligence, and for learning and reasoning in intelligent systems [24, 49]. A conditional independence (CI) statement $I(Y,Z \mid X)$ represents the independence of

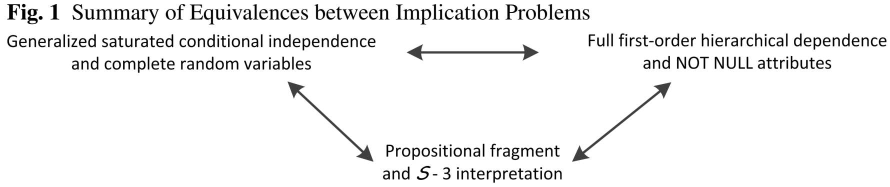
Sebastian Link
The University of Auckland, New Zealand, e-mail: s.link@auckland.ac.nz

two sets of random variables relative to a third: given three mutually disjoint subsets X, Y, and Z of a set S of random variables, if we have knowledge about the state of X, then knowledge about the state of Y does not provide additional evidence for the state of Z and vice versa. A fundamental problem is the implication problem, which is to decide for an arbitrary finite set S, and an arbitrary set $\Sigma \cup \{\varphi\}$ of CI statements over S, whether every probability model that satisfies every CI statement in Σ also satisfies φ. The significance of this problem is due to its relevance for building Bayesian networks [49]. The implication problem for CI statements is not axiomatizable by a finite set of Horn rules [56]. An important subclass of CI statements are saturated conditional independence (SCI) statements. These are CI statements $I(Y,Z \mid X)$ over S that satisfy $XYZ = S$, that is, the set union XYZ of X, Y, and Z is S. Geiger and Pearl have established an axiomatization for the implication problem of SCI statements by a finite set of Horn rules [21].

The notion of saturated conditional independence $I(Y,Z \mid X)$ over S is closely related to that of a multivalued dependency (MVD) $X \twoheadrightarrow Y|Z$ over S, studied in the framework of relational databases [5, 7, 15, 18, 26, 39, 40, 51]. Here, a set X of attributes is used to denote the X-value of a tuple over S, i.e., those tuple components that appear in the columns associated with X. Indeed, $X \twoheadrightarrow Y|Z$ expresses the fact that an X-value uniquely determines the set of associated Y-values independently of joint associations with Z-values where $Z = S - XY$. Thus, given a specific occurrence of an X-value within a tuple, so far not knowing the specific association with a Y-value and Z-value within this tuple, and then learning about the specific associated Y-value does not provide any information about the specific associated Z-value. Previous research has established an equivalence between the implication problem for SCI statements and that for MVDs [61]. In addition it is known that the implication problem of MVDs is equivalent to that of formulae in a Boolean propositional fragment $\mathfrak{F}'$ [51], even in nested databases with finite list, and record constructors [27]. Indeed, Sagiv et al. showed that it suffices to consider two-tuple relations in order to decide the implication problem of MVDs [51]. This enabled them to define truth assignments from two-tuple relations, and vice versa, in such a way that the two-tuple relation satisfies an MVD if and only if the truth assignment is a model for the $\mathfrak{F}'$-formula that corresponds to the MVD. It follows from these results that the implication of SCI statements is equivalent to that of $\mathfrak{F}'$-formulae.

Contribution. The purpose of this article is to summarize recent insight into the relationships between implication problems for fragments of conditional independencies, database dependencies, and propositional logic. The classical equivalences described above are extended in two directions. Firstly, extensions of saturated CI statements, multivalued dependencies, and the propositional fragment $\mathfrak{F}'$ are considered. These extensions include generalized saturated conditional independence (GSCI) statements $I(Y_1,\ldots,Y_k \mid X)$, which declare the independence between any finite number k of sets $Y_1,\ldots,Y_k$ of random variables, given X; as well as Delobel's class of full first-order hierarchical dependencies $X : [Y_1 \mid \ldots \mid Y_k]$ as an extension of MVDs $X \twoheadrightarrow Y_i|R - XY_i$ for $i = 1,\ldots,k$. Secondly, these extensions are handled in the presence of incomplete data. For the probabilistic framework this means that incomplete data can be present in some random variables, and for the database framework this means that null markers can be present in some attribute columns. As a mechanism to control the degree of incomplete data we permit random variables to be specified as complete, that is, incomplete data cannot be assigned to them. Similarly, attributes can be specified as `NOT NULL` to disallow occurrences of null markers in these columns. In fact, the industry standard SQL for defining and querying data permits attributes to be specified as `NOT NULL` [10]. As a main contribution we establish axiomatizations, by a finite set of Horn rules, for the implication problems of i) generalized saturated conditional independencies in the presence of an arbitrary finite set C of complete random variables, and ii) full first-order hierarchical dependencies in the presence of an arbitrary finite set R_s of attributes declared `NOT NULL`. It is shown that both implication problems coincide with the implication problem of a propositional fragment $\mathfrak{F}$ under interpretations by the well-known approximation logic $\mathscr{S}$-3. Indeed, the propositional variables in the set $\mathscr{S}$ correspond to the complete random variables in C as well as the `NOT NULL` attributes in R_s. The main proof arguments are based on special probability models that assign probability one half to two distinct assignments, and on two-tuple relations, since these allow us to define corresponding $\mathscr{S}$-3 truth assignments. The established equivalences are rather special, since any duality between two of these three frameworks fails already for general CI statements, embedded multivalued dependencies and any Boolean propositional fragment over complete data. The equivalences are illustrated in Figure 1. In particular, they should be understood as strong drivers for the advanced treatment of (in)dependence statements as first-class citizens in some uniform framework for reasoning, such as dependence and independence logic [2, 13, 19, 22, 45, 58, 59].

Fig. 1 Summary of Equivalences between Implication Problems



Organization. Generalized conditional independence statements and complete random variables are defined in Section 2. Their combined implication problem is axiomatized in Section 3. In Section 4 we prove the equivalence between the C-implication of GSCI statements and $\mathscr{S}$-3 implication of the propositional fragment $\mathfrak{F}$. In Section 5 this equivalence is extended to include Delobel's class of full first-order hierarchical dependencies and `NOT NULL` attributes. Related work is outlined in Section 6. We conclude in Section 7.

2 Generalized Conditional Independence under Incomplete Data

We use the framework of Geiger and Pearl [21]. We denote by S a finite set of distinct symbols $\{v_1,\dots,v_n\}$, called *random variables*. A *domain mapping* is a mapping that associates a set, $dom(v)$, with each random variable v. The set $dom(v)$ is called the *domain* of v and each of its elements is called a *data value* of v. For $X \subseteq S$ we say that $\mathbf{x}$ is an assignment of X, if $\mathbf{x} \in \prod_{v\in X} dom(v)$. For an assignment $\mathbf{x} = (\mathbf{v}_1,\dots,\mathbf{v}_k)$ of X with $\mathbf{v}_i \in dom(v_i)$, we write $\mathbf{x}(v_i)$ for the data value $\mathbf{v}_i$ of v_i. For some $Y \subseteq X$ we write $\mathbf{x}(Y)$ for the projection of $\mathbf{x}$ onto Y, that is, $\mathbf{x}(Y)$ denotes the restriction of the assignment $\mathbf{x}$ to the random variables in Y.

2.1 Complete Random Variables

In theory one can assume that the data values of assignments always exist and are known. In practice, these assumptions fail frequently. Indeed, it can happen in most samples that some data values do not exist, or that some existing data values are currently unknown. In statistics and machine learning, one speaks commonly of structural zeros in the first case, and of sampling zeros in the second case [17, 54]. In databases, one speaks of inapplicable nulls in the first case, and of unknown nulls in the second case [8, 9, 63]. In practice, it is often difficult to tell whether some data value does not exist, or exists but is currently unknown.

We use the notation $\mathbf{x}(v) = \mu$ to denote that no information is currently available about the data value $\mathbf{x}(v)$ of the random variable v assigned to $\mathbf{x}$. The interpretation of the marker μ as no information means that a data value does either not exist, or a data value exists but is currently unknown.

It is an advantage to gain control over the occurrences of incomplete data values. For this purpose we introduce *complete random variables*. A random variable is defined to be *complete* if and only if $\mu \notin dom(v)$. Although we include μ in domains of random variables that are not complete, we prefer to think of μ as a marker and not as a data value. In what follows we use C to denote the subset of complete random variables. It is a goal of this article to investigate the properties of generalized saturated conditional probabilistic independence in the presence of an arbitrarily chosen set C of complete random variables. Indeed, complete random variables are shown to provide an effective means to control the degree of uncertainty and to soundly approximate classical reasoning.

2.2 Conditional Independence under Complete Random Variables

A *probability model* over $(S = \{v_1,\dots,v_n\},C)$ is a pair (dom,P) where *dom* is a domain mapping that maps each v_i to a finite domain $dom(v_i)$, and $P : dom(v_1) \times \cdots \times dom(v_n) \rightarrow [0,1]$ is a probability distribution having the Cartesian product of these domains as its sample space. Note that $\mu \notin dom(v_i)$ if and only if $v_i \in C$. An assignment $\mathbf{x}$ of $X \subseteq S$ is *complete* if and only if $\mathbf{x}(v) \neq \mu$ holds for all $v \in X$. As usual, for an assignment $\mathbf{x}$ of X, $P(\mathbf{x})$ denotes the marginal probability $P(X = \mathbf{x})$.

Definition 1. The expression $I(Y_1,\ldots,Y_k \mid X)$, where k is a non-negative integer, and $X, Y_1, \ldots, Y_k$ are mutually disjoint subsets of S, is called a *generalized conditional independence* (CI) statement over S. If $XY_1 \cdots Y_k = S$, we call $I(Y_1,\ldots,Y_k \mid X)$ a *generalized saturated conditional independence* (GSCI) statement. Let (dom, P) be a probability model over (S,C). A generalized CI statement $I(Y_1,\ldots,Y_k \mid X)$ is said to *hold for* (dom,P) if for all complete assignments $\mathbf{x}$ of X, and for all assignments $\mathbf{y}_i$ of Y_i for $i = 1,\ldots,k$,

$$P(\mathbf{y}_1,\ldots,\mathbf{y}_k,\mathbf{x}) \cdot P(\mathbf{x})^{k-1} = P(\mathbf{y}_1,\mathbf{x}) \cdot \ldots \cdot P(\mathbf{y}_k,\mathbf{x}) \quad (1)$$

Equivalently, (dom,P) is said to *satisfy* $I(Y_1,\ldots,Y_k \mid X)$.

Remark 1. The expressions $I(Y_1,\ldots,Y_k \mid X)$ are generalized in the sense that they cover CI statements as the special case where $k = 2$. We assume w.l.o.g. that the sets Y_i are non-empty. Indeed, for all positive k we have the property that a probability distribution satisfies $I(\emptyset, Y_2, \ldots, Y_k \mid X)$ if and only if the probability distribution satisfies $I(Y_2,\ldots,Y_k \mid X)$. In particular, for $k = 1$, the CI statement $I(Y \mid X)$ is always satisfied. One may now define an equivalence relation over the set of generalized CI statements over some fixed set S of random variables. Indeed, two such generalized CI statements are equivalent whenever they are satisfied by the same probability distributions over S. However, our inference rules do not need to be applied to such equivalence classes, as Remark 5 shows. For the sake of simplicity, we assume that in GSCI statements $I(Y_1,\ldots,Y_k \mid X)$ the sets Y_i are non-empty.

Remark 2. The satisfaction of generalized CI statements $I(Y_1,\ldots,Y_k \mid X)$ requires equation (1) to hold for *complete* assignments $\mathbf{x}$ of X only. The reason is that the mutual independence between the sets Y_i is conditional on X. That is, assignments that have *no information* about some random variable in X are not taken into account when judging the independence between distinct Y_i.

Remark 3. If every random variable is declared to be complete, that is when $C = S$, and $k = 2$, then Definition 1 reduces to the standard definition of CI statements [21, 49].

We now introduce the running example of this article.

Example 1. Let $\{m(ovie), a(ctor), r(ole), c(rew), f(eature), l(anguage), s(ubtitle)\}$ denote the set S of random variables, that captures properties of blu-rays we want to model. Let $C = \{m,a,r,c,s\}$ denote the set of complete random variables, and let Σ consist of the GSCI statements $I(sar, c, fl \mid m)$ and $I(sc, ar \mid mfl)$, and let φ be $I(s, ar, flc \mid m)$. We may define the following probability model (dom,P) over (S,C):

- $dom(m) = \{\text{Rashomon}, \text{The Seven Samurai}\}$,

- $dom(a) = \{\text{T. Mifune}, \text{M. Kyo}\}$,
- $dom(r) = \{\text{Tajomaru}, \text{Masako}\}$,
- $dom(c) = \{\text{Kurosawa}, \text{Hashimoto}\}$,
- $dom(f) = \{\text{Tailer}, \text{Comments}, \mu\}$,
- $dom(l) = \{\text{Japanese}, \text{Maori}, \mu\}$,

and define P by assigning the probability one half to each of the following two assignments of (S,C):

movie	*actor*	*role*	*crew*	*feature*	*language*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Kurosawa	μ	μ	Suomi
Rashomon	M. Kyo	Masako	Kurosawa	μ	μ	Deutsch

.

It follows that (dom,P) satisfies Σ, but violates φ.

For the remainder of the article we will be interested in GSCI statements. Let $\Sigma \cup \{\varphi\}$ be a set of GSCI statements over S. We say that Σ *C-implies* φ, denoted by $\Sigma \models_C \varphi$, if every probability model over (S,C) that satisfies every GSCI statement in Σ also satisfies the GSCI statement φ. The *implication problem for GSCI statements and complete r.v.* is defined as the following problem.

PROBLEM:	Implication problem of GSCI statements and complete r.v.
INPUT:	Pair (S,C) with set S of random variables and subset $C \subseteq S$ of complete random variables Set $\Sigma \cup \{\varphi\}$ of GSCI statements over S
OUTPUT:	Yes, if $\Sigma \models_C \varphi$; No, otherwise

Example 2. For $S = \{m,a,r,c,f,l,s\}$, $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$ does not C-imply $\varphi = I(s,ar,cfl \mid m)$ for $C = \{m,a,r,c,s\}$, but Σ does C'-imply φ for $C' = \{f,l\}$. A proof of the former is given by the probability model over (S,C) in Example 1, which satisfies Σ, but violates φ. Intuitively, for φ to be implied by Σ one needs to specify f and l to be complete.

For Σ we let $\Sigma_C^* = \{\varphi \mid \Sigma \models_C \varphi\}$ be the *semantic closure* of Σ, i.e., the set of all GSCI statements C-implied by Σ. In order to characterize the implication problem of GSCI statements and complete r.v. we use a syntactic approach by applying inference rules. These inference rules have the form

$$\frac{\text{premise}}{\text{conclusion}}\text{condition}$$

and inference rules without any premises are called axioms. The premise consists of a finite set of GSCI statements, and the conclusion is a singleton GSCI statement. The condition of the rule is simple in the sense that it stipulates a simple syntactic restriction on the application of the rule. An inference rule is called *sound*, if every probability model over (S,C) that satisfies every GSCI statement in the premise of the rule also satisfies the GSCI statement in the conclusion of the rule, given

Table 1 Axiomatization $\mathfrak{G} = \{\mathscr{T}', \mathscr{S}', \mathscr{C}', \mathscr{W}'\}$ of SCI statements when every r.v. is complete

$$\frac{}{I(S,\emptyset \mid \emptyset)}$$
(saturated trivial independence, $\mathscr{T}'$)

$$\frac{I(Y_1,Y_2 \mid X)}{I(Y_2,Y_1 \mid X)}$$
(symmetry, $\mathscr{S}'$)

$$\frac{I(Z,Y \mid X) \qquad I(Z_1,Z_2 \mid XY)}{I(Z_1,Z_2Y \mid X)}$$
(weak contraction, $\mathscr{C}'$)

$$\frac{I(Y_1,Y_2Z \mid X)}{I(Y_1,Y_2 \mid XZ)}$$
(weak union, $\mathscr{W}'$)

that the condition is satisfied. We write $\Sigma \vdash_{\mathfrak{R}} \varphi$ if and only if there is some *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 GSCI statements 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 GSCI statements and complete r.v., if for every S, every $C \subseteq S$ and for every set Σ of GSCI statements over (S,C) we have $\Sigma^+_{\mathfrak{R}} \subseteq \Sigma^*_C$ ($\Sigma^*_C \subseteq \Sigma^+_{\mathfrak{R}}$). The (finite) set $\mathfrak{R}$ is said to be a (finite) *axiomatization* for the implication problem of GSCI statements and complete r.v., if $\mathfrak{R}$ is both sound and complete.

Theorem 1 (Geiger and Pearl 1993). *The set $\mathfrak{G} = \{\mathscr{T}', \mathscr{S}', \mathscr{C}', \mathscr{W}'\}$ from Table 1 forms a finite axiomatization for the implication problem of SCI statements, that is, the special case of the implication problem for GSCI statements and complete r.v. where all GSCI statements are of the form $I(Y_1,Y_2 \mid X)$ and where all random variables are complete.* $\square$

Remark 4. Studený [56] showed that, in the special case where $C = S$ and $k = 2$, the implication problem of CI statements, i.e. to decide for any given set S of random variables and any given set $\Sigma \cup \{\varphi\}$ of CI statements over S of the kind $I(Y_1,Y_2 \mid X)$ whether $\Sigma \models_S \varphi$ holds, cannot be axiomatized by a finite set of Horn rules of the form

$$I(Y_1,Z_1 \mid X_1) \wedge \cdots \wedge I(Y_k,Z_k \mid X_k) \rightarrow I(Y,Z \mid X) \ .$$

3 Axiomatizing GSCI Statements and Complete R.V.

In this section we show that the finite set $\mathfrak{S}$ of Horn rules from Table 2 forms a finite axiomatization for the implication problem of GSCI statements and complete random variables. Our completeness argument applies special probability models which consist of two assignments with probability one half. Special probability models will be further exploited in subsequent sections.

3.1 Sound Inference Rules

Note the following global condition that we enforce on applications of inference rules that infer GSCI statements. It ensures that sets of random variables that occur in GSCI statements are non-empty.

Remark 5. Whenever we apply an inference rule, then we remove all empty sets Y_i from the exact position in which they occur in the sequence of independent sets of random variables. For instance, we can infer $I(\cdot \mid S)$ by an application of the weak union rule $\mathscr{W}$ to the GSCI statement $I(S \mid \emptyset)$.

The rules in $\mathfrak{S}$ are rather intuitive. The saturated trivial independence rule $\mathscr{T}$ is just $\mathscr{T}'$ when we apply the global condition above. The permutation rule $\mathscr{P}$ replaces the symmetry rule $\mathscr{S}'$ to reflect that a GSCI statement holds for a probability distribution, independently of the order in which the sets Y_i of random variables appear. For the case where $k = 2$, the only non-trivial permutation is easily captured by the symmetry rule $\mathscr{S}'$. The weak union rule $\mathscr{W}$ remains unchanged over $\mathscr{W}'$, except for the number of sets of random variables required. The restricted weak contraction rule $\mathscr{C}$ accommodates the arbitrary number of mutually independent sets of random variables. In addition, $\mathscr{C}$ can only be applied when Y-complete assignments are guaranteed. The next example shows that the condition $Y \subseteq C$ is necessary for the soundness of the restricted weak contraction rule $\mathscr{C}$. As a consequence, the implication problem of GSCI statements and complete random variables is different from the implication problem of GSCI statement where all variables are assumed to be complete.

Example 3. Recall Example 2 where $S = \{m,a,r,c,f,l,s\}$, $C = \{m,a,r,c,s\}$, $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$ and $\varphi = I(s,ar,cfl \mid m)$. Indeed, Σ S-implies φ, but Σ does not C-imply φ.

Finally, the merging rule $\mathscr{M}$ is required to state that also the union of independent sets of random variables can be independent of other sets of random variables. In fact, the presence of $\mathscr{M}$ in $\mathfrak{S}$ is necessary since the conclusion of any other rule features at least as many independent sets as the maximum number of independent sets amongst all its premises.

The soundness of the rules in $\mathfrak{S}$ follows from the following proposition and the soundness of the rules in $\mathfrak{G}$. In particular, for the restricted weak contraction rule $\mathscr{C}$ soundness follows under the restriction that assignments must be Y-complete.

Proposition 1. *Let S denote a finite set of random variables and $C \subseteq S$. A probability distribution $\pi = (dom, P)$ over (S,C) satisfies the GSCI statement $I(Y_1, \ldots, Y_k \mid X)$ if and only if for every $i = 1, \ldots, k$, π satisfies the SCI statement $I(Y_i, S - XY_i \mid X)$.*

Proof. Assume that for every $i = 1, \ldots, k$, π satisfies the SCI statement $I(Y_i, S - XY_i \mid X)$. Let $\mathbf{x}$ be a complete assignment over X, and $\mathbf{y}_1, \ldots, \mathbf{y}_k$ be assignments for $Y_1, \ldots, Y_k$, respectively. Then we have

Table 2 Axiomatization $\mathfrak{S} = \{\mathscr{T},\mathscr{P},\mathscr{M},\mathscr{W},\mathscr{C}\}$ of GSCI statements and complete r.v.

$$\frac{}{I(S \mid \emptyset)}$$
(saturated trivial independence, $\mathscr{T}$)

$$\frac{I(Y_1,\ldots,Y_k \mid X)}{I(Y_{\pi(1)},\ldots,Y_{\pi(k)} \mid X)}$$
(permutation, $\mathscr{P}$)

$$\frac{I(Y_1,\ldots,Y_{k-1},Y_k,Z \mid X)}{I(Y_1,\ldots,Y_kZ \mid X)}$$
(merging, $\mathscr{M}$)

$$\frac{I(Y_1,\ldots,Y_{k-1},Y_kZ \mid X)}{I(Y_1,\ldots,Y_k \mid XZ)}$$
(weak union, $\mathscr{W}$)

$$\frac{I(Y_1\cdots Y_k,YZ_1\cdots Z_k \mid X) \quad I(Y_1Z_1,\ldots,Y_kZ_k \mid XY)}{I(Y_1,\ldots,Y_k,YZ_1\cdots Z_k \mid X)} \; Y \subseteq C$$
(restricted weak contraction, $\mathscr{C}$)

$$\begin{aligned} P(\mathbf{x}\mathbf{y}_1\cdots\mathbf{y}_k)\cdot P(\mathbf{x})^{k-1} &= P(\mathbf{x}\mathbf{y}_1)\cdot P(\mathbf{x}\mathbf{y}_2\cdots\mathbf{y}_k)\cdot P(\mathbf{x})^{k-2} \\ &= P(\mathbf{x}\mathbf{y}_1)\cdot P(\mathbf{x}\mathbf{y}_2)\cdot P(\mathbf{x}\mathbf{y}_3\cdots\mathbf{y}_k)\cdot P(\mathbf{x})^{k-3} \\ &= \ldots \\ &= P(\mathbf{x}\mathbf{y}_1)\cdot\ldots\cdot P(\mathbf{x}\mathbf{y}_k)\ , \end{aligned}$$

that is, $\pi = (dom,P)$ satisfies $I(Y_1,\ldots,Y_k \mid X)$.

Vice versa, assume that $\pi = (dom,P)$ over (S,C) satisfies $I(Y_1,\ldots,Y_k \mid X)$. Let $\mathbf{x}$ be a complete assignment over X and $\mathbf{y}_i,\mathbf{y}_1\cdots\mathbf{y}_{i-1}\mathbf{y}_{i+1}\cdots\mathbf{y}_k$ be assignments for Y_i and $S-XY_i$, respectively. Then

$$P(\mathbf{x}\mathbf{y}_i\mathbf{y}_1\cdots\mathbf{y}_{i-1}\mathbf{y}_{i+1}\cdots\mathbf{y}_k)\cdot P(\mathbf{x}) = P(\mathbf{x}\mathbf{y}_i)\cdot P(\mathbf{x}\mathbf{y}_1\cdots\mathbf{y}_{i-1}\mathbf{y}_{i+1}\cdots\mathbf{y}_k)\ ,$$

that is, for every $i = 1,\ldots,k$, π satisfies the SCI statement $I(Y_i,S-XY_i \mid X)$. □

Example 4. For every probability model π over $S = \{m,a,r,c,f,l,s\}$ and every set $C \subseteq S$ of complete random variables, the GSCI statement $I(sar,c,fl \mid m)$ is satisfied by π if and only if all of the SCI statements $I(sar,cfl \mid m)$, $I(c,sarfl \mid m)$ and $I(sarc,fl \mid m)$ are satisfied by π.

Example 5. We can now prove that for $S = \{m,a,r,c,f,l,s\}$, $C = \{f,l\}$, $\Sigma = \{I(sar,c,fl \mid m),I(sc,ar \mid mfl)\}$ does indeed C-imply $\varphi = I(s,ar,cfl \mid m)$, thereby validating our statements from Example 2. In fact, the inference

$$\mathscr{C}: \frac{\mathscr{M}: \dfrac{I(sar,c,fl \mid m)}{I(sar,cfl \mid m)} \qquad I(sc,ar \mid mfl)}{I(s,ar,cfl \mid m)}\{f,l\}\subseteq C$$

shows that $\Sigma \vdash_{\mathfrak{S}} \varphi$ which means that $\Sigma \models_C \varphi$ by soundness of $\mathfrak{S}$.

The following remark shows that Proposition 1 can be used to establish directly that the set $\mathfrak{S}$ of inference rules from Table 2 forms a finite axiomatization for the implication problem of GSCI statements and complete random variables.

Remark 6. The following set $\mathfrak{D}$ of inference rules

$$\begin{array}{cc}
\dfrac{}{I(S,\emptyset \mid \emptyset)} & \dfrac{I(Y_1,Y_2 \mid X)}{I(Y_2,Y_1 \mid X)} \\[2ex]
\dfrac{I(Y_1Y_2,YZ_1Z_2 \mid X) \;\; I(Y_1Z_1,Y_2Z_2 \mid XY)}{I(Y_1,Y_2YZ_1Z_2 \mid X)}\, Y \subseteq C & \dfrac{I(Y_1,Y_2Z \mid X)}{I(Y_1,Y_2 \mid XZ)}
\end{array}$$

forms a finite axiomatization for the implication problem of SCI statements and complete random variables [44]. Suppose that S, $C \subseteq S$ and $\Sigma \cup \{I(Y_1,\ldots,Y_k \mid X)\}$ are given such that $\Sigma \models_C I(Y_1,\ldots,Y_k \mid X)$ holds. For

$$\Sigma_2 = \{I(V_j, S - UV_j \mid U) \mid I(V_1,\ldots,V_m \mid U) \in \Sigma\},$$

and all $i = 1,\ldots,k$ it follows from Proposition 1 that $\Sigma_2 \models_C I(Y_i, S - XY_i \mid X)$ holds, too. The completeness of $\mathfrak{D}$ for the implication of SCI statements and complete random variables means that for all $i = 1,\ldots,k$, $\Sigma_2 \vdash_{\mathfrak{D}} I(Y_i, S - XY_i \mid X)$ holds. Since $\mathfrak{D}$ is subsumed by $\mathfrak{S}$ as the special case where $k = 2$, we also have for all $i = 1,\ldots,k$ that $\Sigma_2 \vdash_{\mathfrak{S}} I(Y_i, S - XY_i \mid X)$ holds. However, the merging rule $\mathscr{M}$ shows that $\Sigma \vdash_{\mathfrak{S}} \sigma$ holds for all $\sigma \in \Sigma_2$. Consequently, for all $i = 1,\ldots,k$, $\Sigma \vdash_{\mathfrak{S}} I(Y_i, S - XY_i \mid X)$ holds. Finally, repeated applications of the restricted weak contraction rule $\mathscr{C}$ and the permutation rule $\mathscr{P}$ show that $\Sigma \vdash_{\mathfrak{S}} I(Y_1,\ldots,Y_k \mid X)$. This establishes the completeness of $\mathfrak{S}$.

Even though the last remark has already established the completeness of $\mathfrak{S}$, we want to illustrate recent techniques for proving completeness without the use of Proposition 1. This will be done in the following subsections.

3.2 The Independence Basis

For some S and $C \subseteq S$, some set Σ of GSCI statements over S, and some $X \subseteq S$ let $IDep_{\Sigma,C}(X) := \{Y \subseteq S - X \mid \Sigma \vdash_{\mathfrak{S}} I(Y, S - XY \mid X)\}$ denote the set of all $Y \subseteq S - X$ such that $I(Y, S - XY \mid X)$ can be inferred from Σ by $\mathfrak{S}$. Note that the empty set $\emptyset$ is an element of $IDep_{\Sigma,C}(X)$.

Lemma 1. *The structure $(IDep_{\Sigma,C}(X), \subseteq, \cup, \cap, (\cdot)^{\mathscr{C}}, \emptyset, S - X)$ forms a finite Boolean algebra, where $(\cdot)^{\mathscr{C}}$ maps a set W to its complement $S - (XW)$.*

Proof. It suffices to show that $IDep_{\Sigma,C}(X)$ is closed under union, intersection, and difference. The soundness of the merging rule $\mathscr{M}$ shows the closure under union.

The soundness of the weak contraction rule $\mathscr{C}$ for the special case where $k = 2$ and $Y = \emptyset$ shows the closure under intersection and difference. $\square$

Recall that an element $a \in P$ of a poset $(P, \sqsubseteq, 0)$ with least element 0 is called an *atom* of $(P, \sqsubseteq, 0)$ precisely when $a \neq 0$ and every element $b \in P$ with $b \sqsubseteq a$ satisfies $b = 0$ or $b = a$ [23]. Further, $(P, \sqsubseteq, 0)$ is said to be *atomic* if for every element $b \in P - \{0\}$ there is an atom $a \in P$ with $a \sqsubseteq b$. In particular, every finite Boolean algebra is atomic [23]. Let $IDepB_{\Sigma,C}(X)$ denote the set of all atoms of $(IDep_{\Sigma,C}(X), \subseteq, \emptyset)$. We call $IDepB_{\Sigma,C}(X)$ the *independence basis* of X with respect to Σ. Its importance is manifested in the following result.

Theorem 2. *Let Σ be a set of GSCI statements over S and $C \subseteq S$. Then $\Sigma \vdash_{\mathfrak{G}} I(Y_1, \ldots, Y_k \mid X)$ if and only if for every $i = 1, \ldots, k$, $Y_i = \bigcup \mathscr{Y}$ for some $\mathscr{Y} \subseteq IDepB_{\Sigma,C}(X)$.*

Proof. Let $\Sigma \vdash_{\mathfrak{G}} I(Y_1, \ldots, Y_k \mid X)$. Then for all $i = 1, \ldots, k$, $\Sigma \vdash_{\mathfrak{G}} I(Y_i, S - XY_i \mid X)$ by the merging rule $\mathscr{M}$. Hence, for all $i = 1, \ldots, k$, $Y_i \in IDep_{\Sigma,C}(X)$. Since every element b of a Boolean algebra is the union over those atoms a with $a \subseteq b$ it follows that for all $i = 1, \ldots, k$, $Y_i = \bigcup \mathscr{Y}$ for $\mathscr{Y} = \{W \in IDepB_{\Sigma,C}(X) \mid W \subseteq Y_i\}$.

Vice versa, let $IDepB_{\Sigma,C}(X) = \{W_1, \ldots, W_n\}$ and for all $i = 1, \ldots, k$, let $Y_i = \bigcup \mathscr{Y}$ for some $\mathscr{Y} \subseteq IDepB_{\Sigma,C}(X)$. Since $I(W_1, \ldots, W_n \mid X) \in \Sigma^+_{\mathfrak{G}}$ holds, successive applications of the permutation rule $\mathscr{P}$ and merging rule $\mathscr{M}$ result in $I(Y_1, \ldots, Y_k \mid X) \in \Sigma^+_{\mathfrak{G}}$. $\square$

Example 6. Recall our example where $S = \{m, a, r, c, f, l, s\}$, $C = \{m, a, r, c, s\}$, $C' = \{f, l\}$, $\Sigma = \{I(sar, c, fl \mid m), I(sc, ar \mid mfl)\}$, and $\varphi = I(s, ar, flc \mid m)$. It follows that $IDepB_{\Sigma,C}(m) = \{sar, c, fl\}$, which we can suitably represent in the form of the single GSCI statement $I(sar, c, fl \mid m)$. According to Theorem 2, $\Sigma \not\models_C \varphi$. Moreover, $\Sigma \models_{C'} \varphi$ since $IDepB_{\Sigma,C'}(m) = \{s, ar, c, fl\}$.

3.3 Completeness

The original completeness proof for multivalued dependencies constructs a counterexample relation with 2^k tuples [5], where k denotes the elements in the (in)dependence basis $Dep_\Sigma(X)$ for the multivalued dependency $X \twoheadrightarrow Y \mid Z \notin \Sigma^+$. The original completeness proof for SCI statements constructs a probability model with $2^{|X|+1}$ values, where $I(Y, Z \mid X) \notin \Sigma^+_{\mathfrak{G}}$ [21]. Here, a recent technique [29] defines special probability models with two assignments of probability one half each. The technique therefore extends the existence of special probability models from the case of marginal SCI statements $I(Y, Z \mid \emptyset)$ [21] to GSCI statements and complete random variables.

Theorem 3. *The set $\mathfrak{S}$ is complete for the implication problem of GSCI statements and complete random variables.*

Proof. Let $\Sigma \cup \{I(Y_1,\ldots,Y_k \mid X)\}$ be a set of GSCI statements over S and $C \subseteq S$, and suppose that $I(Y_1,\ldots,Y_k \mid X)$ cannot be inferred from Σ using $\mathfrak{S}$. We will show that $I(Y_1,\ldots,Y_k \mid X)$ is not C-implied by Σ. For this purpose, we will construct a probability model over (S,C) that satisfies all GSCI statements of Σ, but violates $I(Y_1,\ldots,Y_k \mid X)$.

Let $IDepB_{\Sigma,C}(X) = \{W_1,\ldots,W_n\}$, in particular $S = XW_1\cdots W_n$. Since $I(Y_1,\ldots,Y_k \mid X) \notin \Sigma_{\mathfrak{S}}^+$ we conclude by Theorem 2 that there is some $j \in \{1,\ldots,k\}$ such that Y_j is not the union of some elements of $IDepB_{\Sigma,C}(X)$. Consequently, there is some $i \in \{1,\ldots,n\}$ such that $Y_j \cap W_i \neq \emptyset$ and $W_i - Y_j \neq \emptyset$ hold. Let

$$T := \bigcup_{l\in\{1,\ldots,i-1,i+1,\ldots,k\}} W_l \cap C,$$

and

$$T' := \bigcup_{l\in\{1,\ldots,i-1,i+1,\ldots,k\}} W_l - C.$$

In particular, S is the disjoint union of X, T, T', and W_i. For every $v \in S - C$ we define $dom(v) = \{\mathbf{0},\mathbf{1},\mu\}$; and for every $v \in C$ we define $dom(v) = \{\mathbf{0},\mathbf{1}\}$. We define the following two assignments $\mathbf{a}_1$ and $\mathbf{a}_2$ of S. We define $\mathbf{a}_1(v) = \mathbf{0}$ for all $v \in XW_iT$, $\mathbf{a}_1(v) = \mu$ for all $v \in T'$. We further define $\mathbf{a}_2(v) = \mathbf{a}_1(v)$ for all $v \in XTT'$, and $\mathbf{a}_2(v) = \mathbf{1}$ for all $v \in W_i$. As probability measure we define $P(\mathbf{a}_1) = P(\mathbf{a}_2) = 0.5$. It follows from the construction that (dom,P) does not satisfy $I(Y_1,\ldots,Y_k \mid X)$.

Table 3 Special Probability Model from the Completeness Proof for $\mathfrak{S}$

XT	W_i	T'	P
$\mathbf{0}\cdots\mathbf{0}$	$\mathbf{0}\cdots\mathbf{0}$	$\mu\cdots\mu$	0.5
$\mathbf{0}\cdots\mathbf{0}$	$\mathbf{1}\cdots\mathbf{1}$	$\mu\cdots\mu$	0.5

It remains to show that (dom,P) satisfies every GSCI statement $I(V_1,\ldots,V_m \mid U)$ in Σ. Suppose that for some complete assignment $\mathbf{u}$ of U, $P(\mathbf{u}) = 0$. Then equation (1) will always be satisfied.

If $P(\mathbf{u},\mathbf{v}_o) = 0$ for some complete assignment $\mathbf{u}$ of U, and for some assignment $\mathbf{v}_o$ of V_o, then $P(\mathbf{u},\mathbf{v}_1,\ldots,\mathbf{v}_m) = 0$. Then equation (1) is also satisfied. Suppose that for some complete assignment $\mathbf{u}$ of U, $P(\mathbf{u}) = 0.5$. If for some assignments $\mathbf{v}_l$ of V_l for $l = 1,\ldots,m$, $P(\mathbf{u},\mathbf{v}_1) = \cdots = P(\mathbf{u},\mathbf{v}_m) = 0.5$, then $P(\mathbf{u},\mathbf{v}_1,\ldots,\mathbf{v}_m) = 0.5$, too. Again, equation (1) is satisfied.

It remains to consider the case where $\mathbf{u}$ is some complete assignment of U such that $P(\mathbf{u}) = 1$. In this case, the construction of the probability model tells us that $U \subseteq XT$. Consequently, we can apply the weak union rule $\mathscr{W}$ and permutation rule $\mathscr{P}$ to $I(V_1,\ldots,V_m \mid U) \in \Sigma$ to infer $I(V_1 - XT,\ldots,V_m - XT \mid XT) \in \Sigma_{\mathfrak{S}}^+$. Theorem 2 also shows that $I(W_i, TT' \mid X) \in \Sigma_{\mathfrak{S}}^+$. Now we define $V_l' := V_l - XTT'$ and $Z_l := (V_l - XT) \cap T'$ for $l = 1,\ldots,m$. Consequently, $W_i = V_1'\cdots V_m'$, $T' = Z_1'\cdots Z_m'$, and $V_l - XT = V_l'Z_l'$ for $l = 1,\ldots,m$. An application of the restricted weak contraction rule $\mathscr{C}$ to $I(V_1'Z_1',\ldots,V_m'Z_m' \mid XT)$ and $I(V_1'\cdots V_m', TZ_1'\cdots Z_m' \mid X)$ results in $I(V_1',\ldots,V_m', TZ_1'\cdots Z_m' \mid X) = I(V_1 - XTT',\ldots,V_m - XTT', TT' \mid X)$. It follows from Theorem 2 that $V_l - XTT'$, for every $l = 1,\ldots,m$, is the union of elements from $IDepB_{\Sigma,C}(X)$. Consequently, $V_o - XTT' = W_i$ for some $o \in \{1,\ldots,m\}$ and $V_p - XTT' = \emptyset$ for all $p \in \{1,\ldots,m\} - \{o\}$. Therefore, $W_i \subseteq V_o$ and $W_i \cap V_p = \emptyset$ for all $p \in \{1,\ldots,m\} - \{o\}$. Then, we are either in some previous case where $P(\mathbf{u},\mathbf{v}_l) = 0$ for some $l \in \{1,\ldots,m\}$; or, $P(\mathbf{u},\mathbf{v}_o) = 0.5$, $P(\mathbf{u},\mathbf{v}_p) = 1$ for every $p \in \{1,\ldots,m\} - \{o\}$, and $P(\mathbf{u},\mathbf{v}_1,\ldots,\mathbf{v}_m) = 0.5$. Again, equation (1) is satisfied. This concludes the proof. $\square$

The next example illustrates the construction of the counter-example on our running example.

Example 7. Let $S = \{m,a,r,c,f,l,s\}$ denote the set of random variables from Example 1 and $C = \{m,a,r,c,s\}$, let $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$, and $\varphi = I(s,ar,cfl \mid m)$. The assignments

$$\mathbf{a}_1 = (\text{Rashomon}, \text{T. Mifune}, \text{Tajomaru}, \text{Kurosawa}, \mu, \mu, \text{Suomi})$$

and

$$\mathbf{a}_2 = (\text{Rashomon}, \text{M. Kyo}, \text{Masako}, \text{Kurosawa}, \mu, \mu, \text{Deutsch})$$

taken together with the probability distribution $P(\mathbf{a}_1) = 0.5 = P(\mathbf{a}_2)$ define a probability model that satisfies Σ and violates φ. Indeed, this probability model is an instance of the special probability model used in the completeness proof of Theorem 3, see Table 3. In fact, $W_i = \{a,r,s\}$ and $Y_j = \{s\}$.

3.4 Special Probability Models

We call a probability model (dom,P) over (S,C) *special*, if for every $v \in C$, $dom(v)$ consists of two elements, for every $v \in S - C$, $dom(v)$ consists of two elements and the marker μ, and there are two assignments $\mathbf{a}_1,\mathbf{a}_2$ over (S,C) such that $P(\mathbf{a}_1) = 0.5 = P(\mathbf{a}_2)$. We say that Σ *C-implies* φ in the world of special probability models, denoted by $\Sigma \models_{2,C} \varphi$, if every special probability model over (S,C) that satisfies every GSCI statement in Σ also satisfies the GSCI statement φ. The following variant of the implication problem for GSCI statements and complete random variables emerges.

PROBLEM:	Implication problem for GSCI statements and complete r.v. in the world of special probability models
INPUT:	(S,C) with set S of random variables and subset $C \subseteq S$ of complete random variables Set $\Sigma \cup \{\varphi\}$ of GSCI statements over (S,C)
OUTPUT:	Yes, if $\Sigma \models_{2,C} \varphi$; No, otherwise

The proof of Theorem 3 implies the following result.

Corollary 1. *The implication problem for GSCI statements and complete random variables coincides with the implication problem for GSCI statements and complete random variables in the world of special probability models.*

Proof. Let $\Sigma \cup \{\varphi\}$ be a set of GSCI statements over S and $C \subseteq S$. We need to show that $\Sigma \models_C \varphi$ if and only if $\Sigma \models_{2,C} \varphi$. If it does not hold that $\Sigma \models_{2,C} \varphi$, then it also does not hold that $\Sigma \models_C \varphi$ since every special probability model is a probability model. Vice versa, if it does not hold that $\Sigma \models_C \varphi$, then it does not hold that $\Sigma \vdash_{\mathfrak{S}} \varphi$ since $\mathfrak{S}$ is sound for the implication of GSCI statements and complete random variables. However, the proof of Theorem 3 shows how to construct a special probability model over (S,C) that satisfies every GSCI statement in Σ but does not satisfy φ. Hence, it does not hold that $\Sigma \models_{2,C} \varphi$. $\square$

Corollary 1 shows that to decide the implication problem for GSCI statements and complete random variables it suffices to check special probability models.

Example 8. For the set $S = \{m,a,r,c,f,l,s\}$ of random variables, the subset $C = \{m,a,r,c,s\}$ of complete random variables, and the statements in $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$ and $\varphi = I(s,ar,cfl \mid m)$, the probability model in Example 7 defines a special probability model that satisfies Σ and violates φ. Hence, Σ does not C-imply φ in the world of special probability models.

4 Characterization by an $\mathscr{S}$-3 Fragment

In this section we establish the equivalence between the C-implication of GSCI statements and the implication of formulae in a propositional fragment $\mathfrak{F}$ within Cadoli and Schaerf's well-known approximation logic $\mathscr{S}$-3 [53]. After repeating the syntax and semantics of $\mathscr{S}$-3 logic, we define a mapping of GSCI statements to formulae in $\mathfrak{F}$. The core proof argument establishes an equivalence between special probability models, introduced in the previous section, and special $\mathscr{S}$-3 truth assignments.

4.1 Syntax and Semantics of $\mathscr{S}$-3 logic

Schaerf and Cadoli [53] introduced $\mathscr{S}$-3 logics as "a semantically well-founded logical framework for sound approximate reasoning, which is justifiable from the intuitive point of view, and to provide fast algorithms for dealing with it even when using expressive languages". For a finite set L of propositional variables, let L^* denote the *propositional language* over L, generated from the unary connective $\neg$ (negation), and the binary connectives $\wedge$ (conjunction) and $\vee$ (disjunction). Elements of L^* are also called formulae of L, and usually denoted by φ', ψ' or their subscripted versions. Sets of formulae are denoted by Σ'. We omit parentheses if this does not cause ambiguity.

Let L^ℓ denote the set of all literals over L, i.e., $L^\ell = L \cup \{\neg v' \mid v' \in L\}$. Let $\mathscr{S} \subseteq L$. An $\mathscr{S}$-3 truth assignment of L is a total function $\omega : L^\ell \to \{\mathbb{F}, \mathbb{T}\}$ that maps every propositional variable $v' \in \mathscr{S}$ and its negation $\neg v'$ into opposite truth values ($\omega(v') = \mathbb{T}$ if and only if $\omega(\neg v') = \mathbb{F}$), and that does not map both a propositional variable $v' \in L - \mathscr{S}$ and its negation $\neg v'$ into *false* (we must not have $\omega(v') = \mathbb{F} = \omega(\neg v')$ for any $v' \in L - \mathscr{S}$). Accordingly, for each propositional variable $v' \in L$ and each $\mathscr{S}$-3 truth assignment ω of L there are the following possibilities:

- $\omega(v') = \mathbb{T}$ and $\omega(\neg v') = \mathbb{F}$,
- $\omega(v') = \mathbb{F}$ and $\omega(\neg v') = \mathbb{T}$,
- $\omega(v') = \mathbb{T}$ and $\omega(\neg v') = \mathbb{T}$ (only if $v' \in L - \mathscr{S}$).

$\mathscr{S}$-3 truth assignments generalize both, standard 2-valued truth assignments as well as the 3-valued truth assignments of Levesque [36]. That is, a 2-valued truth assignment is an $\mathscr{S}$-3 truth assignment where $\mathscr{S} = L$, while a 3-valued truth assignment is an $\mathscr{S}$-3 truth assignment with $\mathscr{S} = \emptyset$.

An $\mathscr{S}$-3 truth assignment $\omega : L^\ell \to \{\mathbb{F}, \mathbb{T}\}$ of L can be lifted to a total function $\Omega : L^* \to \{\mathbb{F}, \mathbb{T}\}$. This lifting has been defined as follows [53]. An arbitrary formula φ' in L^* is firstly converted (in linear time in the size of the formula) into its corresponding formula φ'_N in *Negation Normal Form* (NNF) using the following rewriting rules: $\neg(\varphi' \wedge \psi') \mapsto (\neg\varphi' \vee \neg\psi')$, $\neg(\varphi' \vee \psi') \mapsto (\neg\varphi' \wedge \neg\psi')$, and $\neg(\neg\varphi') \mapsto \varphi'$. Therefore, negation in a formula in NNF occurs only at the literal level. The rules for assigning truth values to NNF formulae are as follows:

- $\Omega(\varphi') = \omega(\varphi')$, if $\varphi' \in L^\ell$,
- $\Omega(\varphi' \vee \psi') = \mathbb{T}$ if and only if $\Omega(\varphi') = \mathbb{T}$ or $\Omega(\psi') = \mathbb{T}$,
- $\Omega(\varphi' \wedge \psi') = \mathbb{T}$ if and only if $\Omega(\varphi') = \mathbb{T}$ and $\Omega(\psi') = \mathbb{T}$.

Thus, $\mathscr{S}$-3 logic is non-compositional. An $\mathscr{S}$-3 truth assignment ω is a *model* of a set Σ' of L-formulae if and only if $\Omega(\sigma'_N) = \mathbb{T}$ holds for every $\sigma' \in \Sigma'$. We say that Σ' $\mathscr{S}$-3 *implies* an L-formula φ', denoted by $\Sigma' \models^3_{\mathscr{S}} \varphi'$, if and only if every $\mathscr{S}$-3 truth assignment that is a model of Σ' is also a model of φ'.

4.2 The Propositional Fragment $\mathfrak{F}$

As a first step towards the anticipated duality we define the propositional fragment that corresponds to GSCI statements. Let $\phi : S \to L$ denote a bijection between a set S of random variables and the set $L = \{v' \mid v \in S\}$ of propositional variables. In particular, for $C \subseteq S$ let $\mathscr{S} = \phi(C)$. Thus, complete random variables correspond to propositional variables interpreted classically.

We extend ϕ to a mapping Φ from the set of GSCI statements over S to the fragment $\mathfrak{F}$, that is, $\mathfrak{F}$ is the range of Φ. For a GSCI statement $I(Y_1, \ldots, Y_k \mid X)$ over S, let $\Phi(I(Y_1, \ldots, Y_k \mid X))$ denote the formula

$$\bigvee_{v\in X} \neg v' \vee \bigvee_{i=1}^{k} \left(\bigwedge_{v\in\cup_{j\neq i} Y_j} v' \right).$$

Disjunctions over zero disjuncts are interpreted as *false*, denoted by $\mathbb{F}$, and conjunctions over zero conjuncts are interpreted as *true*, denoted by $\mathbb{T}$. We will simply denote $\Phi(\varphi) = \varphi'$ and $\Phi(\Sigma) = \{\sigma' \mid \sigma \in \Sigma\} = \Sigma'$. Note that for the special case of SCI statements $\varphi = I(Y,Z \mid X)$, that is, GSCI statements where $k = 2$, the formula φ' becomes

$$\bigvee_{v\in X} \neg v' \vee \left(\bigwedge_{v\in Y} v' \right) \vee \left(\bigwedge_{v\in Z} v' \right).$$

Example 9. Let $S = \{m,a,r,c,f,l,s\}$ denote the set of random variables from Example 1 and $C = \{m,a,r,c,s\}$, let $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$, and $\varphi = I(s,ar,cfl \mid m)$. Then $\mathscr{L} = \{m',a',r',c',f',l',s'\}$, $\mathscr{S} = \{m',a',r',c',s'\}$, Σ' consists of

$$\neg m' \vee (c' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge c')$$

and

$$\neg m' \vee \neg f' \vee \neg l' \vee (a' \wedge r') \vee (s' \wedge c'),$$

and $\varphi' = \neg m' \vee (a' \wedge r' \wedge c' \wedge f' \wedge l') \vee (s' \wedge c' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r')$.

4.3 Special Truth Assignments

We will now show that for every set $\Sigma \cup \{\varphi\}$ of GSCI statements over S and every $C \subseteq S$, there is a probability model $\pi = (dom, P)$ over (S,C) that satisfies Σ and violates φ if and only if there is a truth assignment ω'_π that is an $\mathscr{S}$-3 model of Σ' but not an $\mathscr{S}$-3 model of φ'. For arbitrary probability models π it is not obvious how to define the interpretation ω'_π. However, the key to showing the correspondence between counterexample probability models and counterexample truth assignments is Corollary 1. Corollary 1 tells us that for deciding $\Sigma \models_C \varphi$ it suffices to examine special probability models (instead of arbitrary probability models). For a special probability model $\pi = (dom, \{\mathbf{a}_1, \mathbf{a}_2\})$, however, we can define its corresponding special 3-valued truth assignment ω'_π of L as follows:

$$\omega_\pi(v') = \begin{cases} \mathbb{T} \text{ , if } \mathbf{a}_1(v) = \mathbf{a}_2(v) \\ \mathbb{F} \text{ , otherwise} \end{cases} \text{, and}$$

$$\omega_\pi(\neg v') = \begin{cases} \mathbb{T} \text{ , if } \mathbf{a}_1(v) = \mu = \mathbf{a}_2(v) \text{ or } \mathbf{a}_1(v) \neq \mathbf{a}_2(v) \\ \mathbb{F} \text{ , otherwise} \end{cases}.$$

Note that the 3-valued truth assignment is an $\mathscr{S}$-3 truth assignment since it is impossible to have $\mathbf{a}_1(v) = \mu = \mathbf{a}_2(v)$ for any complete random variable $v \in C$. For every $\mathscr{S}$-3 truth assignment ω of L there is some special probability model $\pi = (dom, P)$

over (S,C) such that $\omega_\pi = \omega$. In fact, if $\omega(v') = \mathbb{T} = \omega(\neg v')$ for some $v' \in \mathscr{S}$, then define $dom(v) := \{\mathbf{0},\mathbf{1}\}$ such that the assignments of π are C-complete.

Example 10. Let $S = \{m,a,r,c,f,l,s\}$ denote the set of random variables from Example 1 and $C = \{m,a,r,c,s\}$. The special probability model π defined by

$$\mathbf{a}_1 = (\text{Rashomon}, \text{T. Mifune}, \text{Tajomaru}, \text{Kurosawa}, \mu, \mu, \text{Suomi})$$

and

$$\mathbf{a}_2 = (\text{Rashomon}, \text{M. Kyo}, \text{Masako}, \text{Kurosawa}, \mu, \mu, \text{Deutsch})$$

and the probability distribution $P(\mathbf{a}_1) = 0.5 = P(\mathbf{a}_2)$ translates into the following $\mathscr{S}$-3 interpretation of $\mathscr{L} = \{m',a',r',c',f',l',s'\}$ with $\mathscr{S} = \{m',a',r',c',s'\}$:

- $\omega_\pi(m') = \mathbb{T}$ and $\omega_\pi(\neg m') = \mathbb{F}$
- $\omega_\pi(a') = \mathbb{F}$ and $\omega_\pi(\neg a') = \mathbb{T}$
- $\omega_\pi(r') = \mathbb{F}$ and $\omega_\pi(\neg r') = \mathbb{T}$
- $\omega_\pi(c') = \mathbb{T}$ and $\omega_\pi(\neg c') = \mathbb{F}$
- $\omega_\pi(f') = \mathbb{T}$ and $\omega_\pi(\neg f') = \mathbb{T}$
- $\omega_\pi(l') = \mathbb{T}$ and $\omega_\pi(\neg l') = \mathbb{F}$
- $\omega_\pi(s') = \mathbb{F}$ and $\omega_\pi(\neg s') = \mathbb{T}$

4.4 Semantic Justification of Special Truth Assignments

Next we justify the definition of the special truth assignment and that of the propositional fragment $\mathfrak{F}$ in terms of the special probability models.

Lemma 2. *Let $\pi = (\text{dom}, \{\mathbf{a}_1,\mathbf{a}_2\})$ be a special probability model over (S,C), and let φ denote a GSCI statement over (S,C). Then π satisfies φ if and only if ω'_π is a 3-valued model of φ'.*

Proof. Let $\varphi = I(Y_1,\ldots,Y_k \mid X)$ and

$$\varphi' = \bigvee_{v\in X} \neg v' \vee \bigvee_{i=1}^{k} \left(\bigwedge_{v\in \bigcup_{j\neq i} Y_j} v' \right).$$

Suppose first that π satisfies φ. We need to show that ω'_π is a 3-valued model of φ'. Assume that $\omega'_\pi(\neg v') = \mathbb{F}$ for all $a \in X$. According to the special truth assignment we must have $\mu \neq \mathbf{a}_1(v) = \mathbf{a}_2(v) \neq \mu$ for all $v \in X$. That means $P(\mathbf{a}_1(X)) = 1$. Suppose that for all $i = 2,\ldots,k$ there is some $v \in \bigcup_{j\neq i} Y_j$ such that $\omega'_\pi(v') = \mathbb{F}$. Consequently, there is some $v \in Y_1$ such that $\omega'_\pi(v') = \mathbb{F}$. Hence, $\mathbf{a}_1(v) \neq \mathbf{a}_2(v)$ according to the special truth assignment. Then $P(\mathbf{a}_1(XY_1)) = P(\mathbf{a}_1) = 0.5$. However, since $\mathbf{a}_1(X)$ is complete on X and π satisfies φ we must have $P(\mathbf{a}_1(XY_i)) = 1$ for all $i = 2,\ldots,k$. Hence, for every $v \in Y_2\cdots Y_k$, we have $\mathbf{a}_1(v) = \mathbf{a}_2(v)$. This means that for all $v \in Y_2\cdots Y_k$ we have $\omega'_\pi(v') = \mathbb{T}$. This shows that ω'_π is a 3-valued model of φ'.

Suppose ω'_π is a 3-valued model of φ'. We need to show that π satisfies φ. That is, for every complete assignment $\mathbf{x}$ of X, and every assignment $\mathbf{y}_i$ of Y_i for $i = 1,\ldots,k$, we must show that $P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k)\cdot P(\mathbf{x})^{k-1} = P(\mathbf{x},\mathbf{y}_1)\cdot P(\mathbf{x},\mathbf{y}_k)$ holds. We distinguish between a few cases.

Case 1. If $P(\mathbf{x},\mathbf{y}_i) = 0$ holds for some $i \in \{1,\ldots,k\}$, then $P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k) = 0$ holds, too. For the remaining cases we can therefore assume that for all $i = 1,\ldots,k$, $P(\mathbf{x},\mathbf{y}_i) > 0$. In particular, $P(\mathbf{x}) > 0$.

Case 2. Suppose that $P(\mathbf{x}) = 0.5$. Then $P(\mathbf{x},\mathbf{y}_i) = 0.5$ for all $i = 1,\ldots,k$. Consequently, $(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k)$ equals $\mathbf{a}_1$ or $\mathbf{a}_2$, as $P(\mathbf{x})$ would have to be 1 otherwise. Hence, $P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k) = 0.5$. Therefore, we have

$$P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k)\cdot P(\mathbf{x})^{k-1} = (1/2)^k = P(\mathbf{x},\mathbf{y}_1)\cdot\ldots\cdot P(\mathbf{x},\mathbf{y}_k).$$

Case 3. Suppose $P(\mathbf{x}) = 1$. It follows that $\mathbf{a}_1(X) = \mathbf{x} = \mathbf{a}_2(X)$. Since $\mathbf{x}$ is a complete assignment of X, the special truth assignment entails that $\omega_\pi(\neg v') = \mathbb{F}$ for all $v \in X$. Since ω'_π is a 3-valued model of φ' we conclude that $\omega'_\pi(v') = \mathbb{T}$ for all $v \in S - XY_i$ for some $i \in \{1,\ldots,k\}$. This, however, would mean that $P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_{i-1},\mathbf{y}_{i+1},\ldots,\mathbf{y}_k) = 1$. Since φ is saturated, it follows that $P(\mathbf{x},\mathbf{y}_i) = 0.5$. Consequently, $(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k)$ equals $\mathbf{a}_1$ or $\mathbf{a}_2$. That is, $P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k) = 0.5$. Therefore,

$$P(\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_k)\cdot P(\mathbf{x})^{k-1} = 1/2 = P(\mathbf{x},\mathbf{y}_1)\cdot\ldots\cdot P(\mathbf{x},\mathbf{y}_k).$$

It follows that π satisfies φ. □

4.5 The Equivalence

Corollary 1 and Lemma 2 allow us to establish the anticipated equivalence between the implication problem of GSCI statements and complete random variables and the implication problem of fragment $\mathfrak{F}$ in $\mathscr{S}$-3 logic.

Theorem 4. *Let $\Sigma \cup \{\varphi\}$ be a set of GSCI statements over S and $C \subseteq S$, and let $\Sigma' \cup \{\varphi'\}$ denote the set of its corresponding propositional formulae over L. Then $\Sigma \models_C \varphi$ if and only if $\Sigma' \models^3_{\mathscr{S}} \varphi'$.*

Proof. Based on Corollary 1 it suffices to establish an equivalence between $\Sigma \models_{2,C} \varphi$ and $\Sigma' \models^3_{\mathscr{S}} \varphi'$.

Suppose first that $\Sigma \models_{2,C} \varphi$ does not hold. Then there is some special probability model π over (S,C) that satisfies every GSCI statement σ in Σ but violates φ. Let ω_π denote the special truth assignment associated with π. By Lemma 2 it follows that ω_π is a 3-valued model of every formula σ' in Σ' but not a 3-valued model of φ'. As ω_π is an $\mathscr{S}$-3 truth assignment it follows that $\Sigma' \models^3_{\mathscr{S}} \varphi'$ does not hold.

Suppose now that $\Sigma' \models^3_{\mathscr{S}} \varphi'$ does not hold. Then there is some truth assignment ω over L that is an $\mathscr{S}$-3 model of every formula σ' in Σ', but not an $\mathscr{S}$-3 model of the formula φ'. Define the following special probability model $\pi = (dom, \{\mathbf{a}_1,\mathbf{a}_2\})$ over (S,C). For $v \in C$, let $dom(v) = \{\mathbf{0},\mathbf{1}\}$; and for $v \in S - C$, let $dom(v) = \{\mathbf{0},\mathbf{1},\mu\}$.

We now define $\mathbf{a}_1$ and $\mathbf{a}_2$ as follows. If $\omega(v') = \mathbb{T}$ and $\omega(\neg v') = \mathbb{F}$, then $\mu \neq \mathbf{a}_1(v) = \mathbf{a}_2(v) \neq \mu$. If $\omega(v') = \mathbb{T}$ and $\omega(\neg v') = \mathbb{T}$, then $\mathbf{a}_1(v) = \mu = \mathbf{a}_2(v)$. Finally, if $\omega(v') = \mathbb{F}$ and $\omega(\neg v') = \mathbb{T}$, then $\mu \neq \mathbf{a}_1(v) \neq \mathbf{a}_2(v) \neq \mu$. Since ω is not an $\mathscr{S}$-3 model of φ', it follows that $\mathbf{a}_1 \neq \mathbf{a}_2$. It follows now that $\omega_\pi = \omega$. By Lemma 2 it follows that π satisfies every GSCI statement σ in Σ but violates φ. Hence, $\Sigma \models_{2,C} \varphi$ does not hold. □

Example 11. Let $S = \{m,a,r,c,f,l,s\}$ denote the set of random variables from Example 1 and $C = \{m,a,r,c,s\}$, let $\Sigma = \{I(sar,c,fl \mid m), I(sc,ar \mid mfl)\}$, and $\varphi = I(s,ar,cfl \mid m)$. The special probability model π defined by

$$\mathbf{a}_1 = (\text{Rashomon}, \text{T. Mifune}, \text{Tajomaru}, \text{Kurosawa}, \mu, \mu, \text{Suomi})$$

and

$$\mathbf{a}_2 = (\text{Rashomon}, \text{M. Kyo}, \text{Masako}, \text{Kurosawa}, \mu, \mu, \text{Deutsch})$$

shows that Σ does not C-imply φ. From a logical point of view, the special $\mathscr{S}$-3 interpretation ω_π of $\mathscr{L} = \{m',a',r',c',f',l',s'\}$ with $\mathscr{S} = \{m',a',r',c',s'\}$:

- $\omega_\pi(m') = \mathbb{T}$ and $\omega_\pi(\neg m') = \mathbb{F}$
- $\omega_\pi(a') = \mathbb{F}$ and $\omega_\pi(\neg a') = \mathbb{T}$
- $\omega_\pi(r') = \mathbb{F}$ and $\omega_\pi(\neg r') = \mathbb{T}$
- $\omega_\pi(c') = \mathbb{T}$ and $\omega_\pi(\neg c') = \mathbb{F}$
- $\omega_\pi(f') = \mathbb{T}$ and $\omega_\pi(\neg f') = \mathbb{T}$
- $\omega_\pi(l') = \mathbb{T}$ and $\omega_\pi(\neg l') = \mathbb{F}$
- $\omega_\pi(s') = \mathbb{F}$ and $\omega_\pi(\neg s') = \mathbb{T}$

shows that Σ', consisting of

$$\neg m' \vee (c' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge c')$$

and

$$\neg m' \vee \neg f' \vee \neg l' \vee (a' \wedge r') \vee (s' \wedge c'),$$

does not $\mathscr{S}$-3 imply $\varphi' = \neg m' \vee (a' \wedge r' \wedge c' \wedge f' \wedge l') \vee (s' \wedge c' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r')$.

5 Full Hierarchical Dependencies and `NOT NULL` constraints

In this section we extend the duality between the implication problem of GSCI statements and complete random variables and the implication problem of the fragment $\mathfrak{F}$ under $\mathscr{S}$-3 interpretations to a trinity including the implication problem of Delobel's class of full first-order hierarchical dependencies (FOHDs) [12] and `NOT NULL` constraints. We adapt the technique of special probability models to establish an axiomatization $\mathfrak{H}$ for the implication problem for FOHDs and `NOT NULL` constraints. The completeness proof exploits two-tuple relations. In the database context, two-tuple relations form the counter-part of special probability models, enabling us to establish the anticipated trinity of implication problems. The proof techniques in this section are different from the ones in the previous section in that they explore decomposition arguments rather than probabilities.

5.1 Defining Hierarchical Dependencies under Incomplete Data

Let $\mathfrak{A} = \{\hat{v}_1, \hat{v}_2, \ldots\}$ be a (countably) infinite set of symbols, called *attributes*. A *relation schema* is a finite set $R = \{\hat{v}_1, \ldots, \hat{v}_n\}$ of attributes from $\mathfrak{A}$. Each attribute $\hat{v}$ of a relation schema is associated with a domain $dom(\hat{v})$ which represents the set of possible values that can occur in the column named $\hat{v}$. Note that the validity of our results only depends on having at least two element values in each domain. This is a consequence of our proof techniques. In order to encompass incomplete information the domain of each attribute contains the null marker, denoted by $\texttt{ni} \in dom(\hat{v})$. The intention of $\texttt{ni}$ is to mean "no information". This is the most primitive interpretation, and it can model non-existing as well as unknown information [3, 63]. We stress that the null marker is not a domain value. In fact, it is a purely syntactic convenience that we include the null marker in the domain of each attribute as a distinguished element.

A *tuple* over R is a function $t : R \to \bigcup_{\hat{v}\in R} dom(\hat{v})$ with $t(\hat{v}) \in dom(\hat{v})$ for all $\hat{v} \in R$. The null marker occurrence $t(\hat{v}) = \texttt{ni}$ associated with an attribute $\hat{v}$ in a tuple t means that "no information" is available about the value $t(\hat{v})$ of t on attribute $\hat{v}$. For $X \subseteq R$ let $t(X)$ denote the restriction of the tuple t over R to X, and $dom(X) = \prod_{\hat{v}\in X} dom(\hat{v})$ the Cartesian product of the domains of attributes in X. A (partial) *relation* r over R is a finite set of tuples over R. Let t_1 and t_2 be two tuples over R. It is said that t_1 *subsumes* t_2 if for every attribute $\hat{v} \in R$, $t_1(\hat{v}) = t_2(\hat{v})$ or $t_2(\hat{v}) = \texttt{ni}$ holds. In consistency with previous work [3, 37, 63], the following restriction will be imposed, unless stated otherwise: No relation shall contain two tuples t_1 and t_2 such that t_1 subsumes t_2. With no null markers present this means that no duplicate tuples occur. For a tuple t over R and a set $X \subseteq R$, t is said to be X-total, if for all $\hat{v} \in X$, $t(\hat{v}) \neq \texttt{ni}$. Similar, a relation r over R is said to be X-total, if every tuple t of r is X-total. A relation r over R is said to be a *total relation*, if it is R-total.

We recall the definition of projection and join operations on partial relations [3, 37]. Let r be some relation over R. Let X be some subset of R. The *projection* $r[X]$ of r on X is the set of tuples t for which (i) there is some $t_1 \in r$ such that $t = t_1(X)$ and (ii) there is no $t_2 \in r$ such that $t_2(X)$ subsumes t and $t_2(X) \neq t$. For $Y \subseteq X$, the *Y-total projection* $r_Y[X]$ of r on X is $r_Y[X] = \{t \in r[X] \mid t \text{ is } Y\text{-total}\}$. Given an X-total relation r_1 over R_1 and an X-total relation r_2 over R_2 such that $X = R_1 \cap R_2$ the *natural join* $r_1 \bowtie r_2$ of r_1 and r_2 is the relation over $R_1 \cup R_2$ which contains those tuples t such that there are some $t_1 \in r_1$ and $t_2 \in r_2$ with $t_1 = t(R_1)$ and $t_2 = t(R_2)$ [3, 37]. For example, the relation

movie	*actor*	*role*	*crew*	*feature*	*language*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Kurosawa	ni	ni	Suomi
Rashomon	M. Kyo	Masako	Kurosawa	ni	ni	Deutsch

is the natural join of the following three relations:

movie	*actor*	*role*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Suomi
Rashomon	M. Kyo	Masako	Deutsch

movie	*crew*
Rashomon	Kurosawa

.

movie	*feature*	*subtitle*
Rashomon	ni	ni

Following Atzeni and Morfuni [3], a *null-free subschema* (NFS) over the relation schema R is an expression R_s where $R_s \subseteq R$. The NFS R_s over R is satisfied by a relation r over R, denoted by $\models_r R_s$, if and only if r is R_s-total. SQL, the industry standard for data management, allows attributes to be specified as `NOT NULL` [10].

Definition 2. A *full first-order hierarchical dependency* (FOHD) over the relation schema R is an expression $X : [Y_1 \mid \dots \mid Y_k]$ with a non-negative integer k, $X, Y_1, \dots, Y_k \subseteq R$ such that $Y_1, \dots, Y_k$ form a partition of $R - X$. A relation r over R is said to *satisfy* (or said to be a *model* of) the full first-order hierarchical dependency $X : [Y_1 \mid \cdots \mid Y_k]$ over R, denoted by $\models_r X : [Y_1 \mid \cdots \mid Y_k]$, if and only if $r_X[R] = (\cdots(r_X[XY_k] \bowtie r_X[XY_{k-1}]) \bowtie \cdots) \bowtie r_X[XY_1]$ holds.

The FOHD $\emptyset : [Y_1 \mid \cdots \mid Y_k]$ expresses the fact that any relation over R is the Cartesian product over its projections to attribute sets in $\{Y_i\}_{i=1}^{k}$. For $k = 0$, the FOHD $X : [\,]$ is satisfied trivially, where $[\,]$ denotes the empty list.

Remark 7. In consistency with Remark 1 on GSCI statements, we assume w.l.o.G. that the sets Y_i in FOHDs are non-empty. Indeed, for all positive k we have the property that for all relations r the FOHD $X : [\emptyset, Y_2, \dots, Y_k]$ is satisfied by r if and only if r satisfies the FOHD $X : [Y_2, \dots, Y_k]$. In particular, if $k = 1$, then $X : [\emptyset]$ is equivalent to $X : [\,]$; more specifically, they are both satisfied by all relations.

Example 12. We use now

$$R = \{\hat{m}(\text{ovie}), \hat{v}(\text{ctor}), \hat{r}(\text{ole}), \hat{c}(\text{rew}), \hat{f}(\text{eature}), \hat{l}(\text{anguage}), \hat{s}(\text{ubtitle})\}$$

to denote a relation schema that models information about blu-rays of movies. As the NFS there are at least the two options $R_s = \{\hat{m}, \hat{v}, \hat{r}, \hat{c}, \hat{s}\}$ and $R'_s = \{\hat{f}, \hat{l}\}$. For ease of presentation in this and the following examples we denote attributes by lower-case Latin letters without the $\hat{}$ above them. The following full first-order hierarchical dependencies are specified to enforce consistency in database relations: $\Sigma = \{m : [sar \mid c \mid fl], mfl : [sc \mid ar]\}$. The database design team has identified an additional meaningful FOHD $\varphi = m : [s \mid ar \mid cfl]$, and is wondering whether φ must be enforced in addition to Σ, or whether it is already implicitly enforced by enforcing Σ, i.e., whether φ is R_s-implied or R'_s-implied by Σ, respectively.

5.2 Axiomatization

For the design of a relational database schema semantic constraints are defined on the relations which are intended to be instances of the schema [38]. During the design process one usually needs to determine further constraints which are logically implied by the given ones. As was the case with GSCI statements and propositional formulae before, we can speak of R_s-implication for sets of full first-order hierarchical dependencies. Similarly, we can introduce the notions of soundness and completeness for sets of inference rules. Finite sets of full first-order hierarchical dependencies are denoted by $\hat{\Sigma}$ and single FOHDs by $\hat{\varphi}$.

PROBLEM:	Implication Problem for FOHDs and NFSs
INPUT:	Relation schema R, null-free subschema R_s over R, Set $\hat{\Sigma} \cup \{\hat{\varphi}\}$ of FOHDs over R
OUTPUT:	Yes, if $\hat{\Sigma} \models_{R_s} \hat{\varphi}$; No, otherwise

Table 4 Axiomatization $\mathfrak{F} = \{\hat{\mathscr{U}}, \hat{\mathscr{P}}, \hat{\mathscr{M}}, \hat{\mathscr{A}}, \hat{\mathscr{T}}\}$ of FOHDs and NFS R_s

$$\frac{}{\emptyset : [R]}$$
(universal, $\hat{\mathscr{U}}$)

$$\frac{X : [Y_1 \mid \cdots \mid Y_k]}{X : [Y_{\pi(1)} \mid \cdots \mid Y_{\pi(k)}]}$$
(permutation, $\hat{\mathscr{P}}$)

$$\frac{X : [Y_1 \mid \cdots \mid Y_{k-1} \mid Y_k \mid Z]}{X : [Y_1 \mid \cdots \mid Y_{k-1} \mid Y_k Z]}$$
(merging, $\hat{\mathscr{M}}$)

$$\frac{X : [Y_1 \mid \cdots \mid Y_k Z]}{XZ : [Y_1 \mid \cdots \mid Y_k]}$$
(augmentation, $\hat{\mathscr{A}}$)

$$\frac{X : [Y_1 \cdots Y_k \mid YZ_1 \cdots Z_k] \quad XY : [Y_1 Z_1 \mid \cdots \mid Y_k Z_k]}{X : [Y_1 \mid \cdots \mid Y_k \mid YZ_1 \cdots Z_k]} \; Y \subseteq R_s$$
(restricted transitivity, $\hat{\mathscr{T}}$)

Remark 8. In consistency with Remark 5 on the application of inference rules to GSCI statements, note the following global condition that we enforce on all applications of inference rules that infer FOHDs. Whenever we apply such an inference rule, we remove all empty sets from the exact position in which they occur as elements in the sequence in the conclusion. For instance, we can infer $R : [\,]$ by an application of the *augmentation rule* $\hat{\mathscr{A}}$ to the FOHD $\emptyset : [R]$.

As in the context of GSCI statements, we can define $Dep_{\hat{\Sigma},R_s}(X) := \{Y \subseteq R - X \mid \Sigma \vdash_{\mathfrak{F}} X : [Y \mid R - XY]\}$ as the set of all $Y \subseteq R - X$ such that $X : [Y \mid R - XY]$ can be inferred from $\hat{\Sigma}$ by $\mathfrak{F}$. The special case $Y = \emptyset$ of the restricted transitivity rule $\hat{\mathscr{T}}$ as well the merging rule $\hat{\mathscr{M}}$ show that

$$(Dep_{\hat{\Sigma},R_s}(X),\subseteq,\cup,\cap,(\cdot)^{\mathscr{C}},\emptyset,R-X)$$

forms a finite Boolean algebra where $(\cdot)^{\mathscr{C}}$ maps a set $Y \subseteq R-X$ to its complement $R-(XY)$. Let $DepB_{\hat{\Sigma},R_s}(X)$ denote the set of all atoms of $(Dep_{\hat{\Sigma},R_s}(X),\subseteq,\emptyset)$. We call $DepB_{\hat{\Sigma},R_s}(X)$ the *dependence basis* of X with respect to $\hat{\Sigma}$ and R_s [4]. The proof of the following result follows the proof of Theorem 2.

Theorem 5. *Let $\hat{\Sigma}$ be a set of FOHDs over R. Then $\hat{\Sigma} \vdash_{\mathfrak{F}} X : [Y_1 \mid \cdots \mid Y_k]$ if and only if for every $i=1,\ldots,k$, $Y_i = \bigcup \mathscr{Y}$ for some $\mathscr{Y} \subseteq DepB_{\hat{\Sigma},R_s}(X)$.* □

The completeness proof shows that an FOHD $\hat{\varphi}$ is not R_s-implied by a set of FOHDs $\hat{\Sigma}$ whenever $\hat{\varphi}$ cannot be inferred from $\hat{\Sigma}$ by $\mathfrak{F}$. We will now apply the techniques from the completeness proof for GSCI statements to construct a two-tuple relation that satisfies $\hat{\Sigma}$ but violates $\hat{\varphi}$.

Theorem 6. *The set $\mathfrak{F}$ of inference rules from Table 4 forms an axiomatization for the implication problem of full first-order hierarchical dependencies and null-free subschemata.*

Proof. It remains to show the completeness of $\mathfrak{F}$. Let R be an arbitrary relation schema, let R_s be an NFS over R, and let $\hat{\Sigma}$ be an arbitrary set of FOHDs over R. We need to show that $\hat{\Sigma}^*_{R_s} \subseteq \hat{\Sigma}^+_{\mathfrak{F}}$ holds.

Let $X : [Y_1 \mid \cdots \mid Y_k] \notin \hat{\Sigma}^+_{\mathfrak{F}}$. Let $DepB_{\hat{\Sigma},R_s}(X) = \{W_1,\ldots,W_n\}$, in particular $R = XW_1 \cdots W_n$. Since $X : [Y_1 \mid \cdots \mid Y_k] \notin \hat{\Sigma}^+_{\mathfrak{F}}$ we conclude by Theorem 5 that there is some $j \in \{1,\ldots,k\}$ such that Y_j is not the union of some elements of $DepB_{\hat{\Sigma},R_s}(X)$. Consequently, there is some $i \in \{1,\ldots,n\}$ such that $Y_j \cap W_i \neq \emptyset$ and $W_i - Y_j \neq \emptyset$ hold. Let $T := \bigcup_{l \in \{1,\ldots,i-1,i+1,\ldots,k\}} W_l \cap R_s$, and $T' := \bigcup_{l \in \{1,\ldots,i-1,i+1,\ldots,k\}} W_l - R_s$. In particular, R is the disjoint union of X, T, T', and W_i. We define the following two tuples t_1 and t_2 over R. We define $t_1(\hat{v}) = \mathbf{0}$ for all $\hat{v} \in XW_iT$, $t_1(\hat{v}) = \texttt{ni}$ for all $\hat{v} \in T'$. We further define $t_2(\hat{v}) = t_1(\hat{v})$ for all $\hat{v} \in XTT'$, and $t_2(\hat{v}) = \mathbf{1}$ for all $\hat{v} \in W_i$. The two-tuple relation $r = \{t_1, t_2\}$ is illustrated in Table 5. It is simple to observe that the relation r enjoys the following property: an FOHD $U : [V_1 \mid \cdots \mid V_m]$ is satisfied by r if and only if i) $U \cap T' \neq \emptyset$, or ii) $U \cap W_i \neq \emptyset$, or iii) $W_i \subseteq V_o$ for some $o \in \{V_1,\ldots,V_m\}$. Indeed, if $U \cap T' \neq \emptyset$, then $r_U[Z] = \emptyset$ for all $Z \subseteq R$. If $U \cap T' = \emptyset$ and $U \cap W_i \neq \emptyset$, then the projections $r_U[UV_l]$ contain two tuples for all $l = 1,\ldots,m$ and only the original tuples match on common attributes. If, $U \subseteq XT$ and $W_i \subseteq V_o$, then the projection $r_U[UV_l]$ contains only one tuple for all $V_l \in \{V_1,\ldots,V_n\} - \{V_o\}$, and the projection $r_U[UV_o]$ contains two tuples. The join of those projections is the original relation r. Vice versa, if $U \subseteq XT$ and $W_i \not\subseteq V_o$ for all $V_o \in \{V_1,\ldots,V_m\}$, then the projections $r_U[UV_l]$ contain tuples whose join does not occur in the original relation r (in fact, a projection of some tuple in the joined relation to W_i contains some $\mathbf{0}$s and some $\mathbf{1}$s).

The construction ensures that r violates $X : [Y_1 \mid \cdots \mid Y_k]$ since $X \cap T' = \emptyset$, $X \cap W_i = \emptyset$, and $W_i \not\subseteq Y_s$ for $s = 1,\ldots,k$. Furthermore, r is R_s-total by construction.

It remains to show that r satisfies $\hat{\Sigma}$, that is, every FOHD $U : [V_1 \mid \cdots \mid V_m]$ in $\hat{\Sigma}$. If $U \cap T' \neq \emptyset$ or $U \cap W_i \neq \emptyset$, then r satisfies $U : [V_1 \mid \cdots \mid V_m]$. Otherwise, $U \subseteq$

Table 5 Two-tuple Relation from Completeness Proof of $\mathfrak{F}$

XT	T'	W_i
$\mathbf{0}\cdots\mathbf{0}$	ni$\cdots$ni	$\mathbf{0}\cdots\mathbf{0}$
$\mathbf{0}\cdots\mathbf{0}$	ni$\cdots$ni	$\mathbf{1}\cdots\mathbf{1}$

XT. Consequently, we can apply the augmentation rule $\hat{\mathscr{A}}$ and permutation rule $\hat{\mathscr{P}}$ to $U : [V_1 \mid \cdots \mid V_m] \in \hat{\Sigma}$ to infer $XT : [V_1 - XT \mid \cdots \mid V_l - XT] \in \hat{\Sigma}^+_{\mathfrak{F}}$. Theorem 5 also shows that $X : [W_i \mid TT'] \in \hat{\Sigma}^+_{\mathfrak{F}}$. Now we define $V'_l := V_l - XTT'$ and $Z_l := (V_l - XT) \cap T'$ for $l = 1,\ldots,m$. Consequently, $W_i = V'_1 \cdots V'_m$, $T' = Z'_1 \cdots Z'_m$, and $V_l - XT = V'_l Z'_l$ for $l = 1,\ldots,m$. An application of the restricted transitivity rule $\hat{\mathscr{T}}$ to $XT : [V'_1 Z'_1 | \cdots | V'_m Z'_m]$ and $X : [V'_1 \cdots V'_m | TZ'_1 \cdots Z'_m]$ results in $X : [V'_1 \mid \cdots \mid V'_m \mid TZ'_1 \cdots Z'_m] = X : [V_1 - XTT' \mid \cdots \mid V_m - XTT' \mid TT']$. It follows from Theorem 2 that for every $l = 1,\ldots,m$, $V_l - XTT'$ is the union of elements from $DepB_{\hat{\Sigma},R_s}(X)$. Consequently, $V_o - XTT' = W_i$ for some $o \in \{1,\ldots,m\}$ and, therefore, $W_i \subseteq V_o$. As we have seen above, this means that r indeed satisfies $U : [V_1 \mid \cdots \mid V_m]$. This concludes the proof. □

Example 13. Recall our running example: $R = \{m,a,r,c,f,l,s\}$, $R_s = \{m,a,r,c,s\}$, $\Sigma = \{m : [sar \mid c \mid fl], mfl : [sc \mid ar]\}$, and $\varphi = m : [s \mid ar \mid cfl]$. The construction from Theorem 6 may result in the following relation r

movie	*actor*	*role*	*crew*	*feature*	*language*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Kurosawa	ni	ni	Suomi
Rashomon	M. Kyo	Masako	Kurosawa	ni	ni	Deutsch

that satisfies Σ and R_s, but violates φ. For example, the *movie*-total part of the join of the following projections

movie	*subtitle*
Rashomon	Suomi
Rashomon	Deutsch

movie	*actor*	*role*
Rashomon	T. Mifune	Tajomaru
Rashomon	M. Kyo	Masako

movie	*crew*	*feature*	*language*
Rashomon	Kurosawa	ni	ni

is

movie	*actor*	*role*	*crew*	*feature*	*language*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Kurosawa	ni	ni	Suomi
Rashomon	M. Kyo	Masako	Kurosawa	ni	ni	Deutsch
Rashomon	T. Mifune	Tajomaru	Kurosawa	ni	ni	Deutsch
Rashomon	M. Kyo	Masako	Kurosawa	ni	ni	Suomi

which is different from r.

5.3 Implication of FOHDs and `NOT NULL` Constraints in the World of Two-tuple Relations

A relation r that consists of two tuples is said to be a *two-tuple relation*. We say that Σ R_s*-implies* φ in the world of two-tuple relations, denoted by $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$, if every R_s-total two-tuple relation over R that satisfies every FOHD in $\hat{\Sigma}$ also satisfies the FOHD $\hat{\varphi}$. The following variant of the implication problem for FOHDs and NFSs emerges.

PROBLEM:	Implication problem for FOHDs and NFSs in the world of two-tuple relations
INPUT:	Relation schema R, NFS R_s over R, Set $\hat{\Sigma} \cup \{\hat{\varphi}\}$ of FOHDs over R
OUTPUT:	Yes, if $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$; No, otherwise

The proof of Theorem 6 implies the following result.

Corollary 2. *The implication problem for FOHDs and NFSs coincides with the implication problem for FOHDs and NFSs in the world of two-tuple relations.*

Proof. Let $\hat{\Sigma} \cup \{\hat{\varphi}\}$ be a set of FOHDs over R. We need to show that $\hat{\Sigma} \models_{R_s} \hat{\varphi}$ if and only if $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$. If it does not hold that $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$, then it also does not hold that $\hat{\Sigma} \models_{R_s} \hat{\varphi}$ since every two-tuple relation is a relation. Vice versa, if it does not hold that $\hat{\Sigma} \models_{R_s} \hat{\varphi}$, then it does not hold that $\hat{\Sigma} \vdash_{\mathfrak{F}} \hat{\varphi}$ since $\mathfrak{F}$ is sound for the implication of FOHDs. However, the proof of Theorem 6 shows how to construct an R_s-total two-tuple relation that satisfies every FOHD in $\hat{\Sigma}$ but does not satisfy $\hat{\varphi}$. Hence, it does not hold that $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$. □

Corollary 2 shows that to decide the implication problem for FOHDs and NFSs over R it suffices to check two-tuple relations over R.

Example 14. The two-tuple relation r from Example 13 shows that $\hat{\Sigma}$ does not R_s-imply $\hat{\varphi}_2$ in the world of two-tuple relations.

5.4 Functional and Hierarchical Dependencies

In this subsection we establish a result on the interaction of FOHDs and functional dependencies over two-tuple relations. The finding subsumes a known result on the interaction of multivalued dependencies (MVDs) and functional dependencies over two-tuple relations [1]. Recall that a functional dependency (FD) over relation schema R is an expression $X \to Y$ with $X, Y \subseteq R$. A relation r over R satisfies the FD $X \to Y$ if and only if all tuples $t, t' \in r$ with matching non-null values on all the attributes in X also have matching values on all the attributes in Y, that is, if $t(X) = t'(X)$ and t, t' are X-total, then $t(Y) = t'(Y)$ [3, 37].

Theorem 7. *Let $r = \{t_1, t_2\}$ be a two-tuple relation over relation schema R. Then r satisfies the FOHD $X : [Y_1 \mid \cdots \mid Y_k]$ if and only if there is some $i \in \{1, \ldots, k\}$ such that r satisfies the FD $X \to R - XY_i$.*

Proof. If $t_1(X) \neq t_2(X)$, or t_1 and t_2 are not both X-total, then r satisfies both the FOHD $X : [Y_1 \mid \cdots \mid Y_k]$ and the FDs $X \to R - XY_i$ for all $i = 1, \ldots, k$. For the remainder of the proof we therefore assume that $t_1(X) = t_2(X)$ holds and t_1, t_2 are both X-total, i.e., the tuples in the projections $r_X[XY_i]$ all have matching non-null values on their common attributes, i.e., the attributes in X.

Assume first that r satisfies the FD $X \to R - XY_i$ for some $i \in \{1, \ldots, k\}$. Consequently, the projections $r_X[XY_j]$ contain only one tuple for all $j \in \{1, \ldots, k\} - \{i\}$, and $r_X[XY_i]$ contains at most two tuples. The join $r_X[XY_1] \bowtie \cdots \bowtie r_X[XY_k]$ contains only tuples from r, i.e., $r_X[R] = r_X[XY_1] \bowtie \cdots \bowtie r_X[XY_k]$.

Assume now that r violates the FDs $X \to Y_i$ and $X \to Y_j$ for some $i \neq j$. Then $r_X[XY_i]$ and $r_X[XY_j]$ contain two tuples each. The join $r_X[XY_1] \bowtie \cdots \bowtie r_X[XY_k]$ thus contains tuples that are not originally in r. This concludes the proof. □

Example 15. Recall the following two-tuple relation r from Example 13

movie	*actor*	*role*	*crew*	*feature*	*language*	*subtitle*
Rashomon	T. Mifune	Tajomaru	Kurosawa	ni	ni	Suomi
Rashomon	M. Kyo	Masako	Kurosawa	ni	ni	Deutsch

.

Indeed, r does not satisfy any of the FDs $m \to arcfl$, $m \to rcfls$, nor $m \to as$. According to Theorem 7, r does not satisfy the FOHD $\varphi = m : [s \mid a \mid crfl]$.

5.5 Equivalence to the Propositional Fragment $\mathfrak{F}$

Let $\hat{\phi} : R \to L$ denote a bijection between a relation schema R of attributes $\hat{v}$ and the set $L = \{v' \mid \hat{v} \in R\}$ of propositional variables, where $\hat{\phi}(R_s) = \mathscr{S} \subseteq L$ for an NFS R_s over R. We extend $\hat{\phi}$ to a mapping $\hat{\Phi}$ from the set of FOHDs over R to the fragment $\mathfrak{F}$, that is, $\mathfrak{F}$ is the range of $\hat{\Phi}$. For an FOHD $X : [Y_1 \mid \cdots \mid Y_k]$ over R, let $\hat{\Phi}(X : [Y_1 \mid \cdots \mid Y_k])$ denote the formula

$$\bigvee_{\hat{v} \in X} \neg v' \vee \bigvee_{i=1}^{k} \left(\bigwedge_{\hat{v} \in \cup_{j \neq i} Y_j} v' \right).$$

Recall from before that disjunctions over zero disjuncts are interpreted as $\mathbb{F}$ and conjunctions over zero conjuncts are interpreted as $\mathbb{T}$. We will simply denote $\Phi(\hat{\varphi}) = \varphi'$ and $\hat{\Phi}(\hat{\Sigma}) = \{\sigma' \mid \hat{\sigma} \in \hat{\Sigma}\} = \Sigma'$. Example 9 shows the $\mathfrak{F}$-formulae that correspond to the FOHDs from Example 12.

Note that for the special case of MVDs $\hat{\varphi} = X : [Y \mid Z]$, i.e. FOHDs where $k = 2$, the formula φ' becomes again

$$\bigvee_{\hat{v}\in X}\neg v' \vee \left(\bigwedge_{\hat{v}\in Y} v'\right) \vee \left(\bigwedge_{\hat{v}\in Z} v'\right).$$

We will now show that for any set $\hat{\Sigma}\cup\{\hat{\varphi}\}$ of FOHDs over R there is an R_s-total relation r over R that satisfies $\hat{\Sigma}$ and violates $\hat{\varphi}$ if and only if there is an $\mathscr{S}$-3 truth assignment ω'_r that is an $\mathscr{S}$-3 model of Σ' but not an $\mathscr{S}$-3 model of φ'. For arbitrary relations r it is not obvious how to define the truth assignment ω'_r. However, the key to showing the correspondence between counterexample relations and counterexample truth assignments is Corollary 2. Corollary 2 tells us that for deciding the implication problem of FOHDs and NFSs it suffices to examine two-tuple relations (instead of arbitrary relations). For a two-tuple relation $r = \{t_1,t_2\}$, however, we can define its corresponding special 3-valued truth assignment ω'_r of L as follows:

$$\omega'_r(v') = \begin{cases} \mathbb{T} \text{ , if } t_1(\hat{v}) = t_2(\hat{v}) \\ \mathbb{F} \text{ , otherwise} \end{cases},$$

and

$$\omega'_r(\neg v') = \begin{cases} \mathbb{T} \text{ , if } t_1(\hat{v}) \neq t_2(\hat{v}) \text{ or } t_1(\hat{v}) = \mathtt{ni} = t_2(\hat{v}) \\ \mathbb{F} \text{ , otherwise} \end{cases}.$$

Next we justify the definition of the special truth assignment and that of the propositional fragment $\mathfrak{F}$ in terms of two-tuple relations.

Lemma 3. *Let $r = \{t_1,t_2\}$ be a two-tuple relation over R, and let $\hat{\varphi}$ denote an FOHD over R. Then r satisfies $\hat{\varphi}$ if and only if ω'_r is a 3-valued model of φ'.*

Proof. Let $\hat{\varphi} = X : [Y_1 \mid \cdots \mid Y_k]$ and

$$\varphi' = \bigvee_{\hat{v}\in X}\neg a' \vee \bigvee_{i=1}^{k}\left(\bigwedge_{\hat{v}\in\bigcup_{j\neq i}Y_j} a'\right).$$

Suppose first that r satisfies $\hat{\varphi}$. We need to show that ω'_r is a 3-valued model of φ'. Assume that $\omega'_r(\neg a') = \mathbb{F}$ for all $\hat{v}\in X$. According to the special 3-valued truth assignment we must have $\mathtt{ni} \neq t_1(\hat{v}) = t_2(\hat{v}) \neq \mathtt{ni}$ for all $\hat{v}\in X$. Suppose that for all $i = 2,\dots,k$ there is some $\hat{v}\in\bigcup_{j\neq i}Y_j$ such that $\omega'_r(a') = \mathbb{F}$. Consequently, there is some $\hat{v}\in Y_1$ such that $\omega'_r(a') = \mathbb{F}$. Hence, $t_1(\hat{v}) \neq t_2(\hat{v})$ according to the special 3-valued truth assignment. However, since r satisfies φ, r must satisfy the FD $X \to Y_2\cdots Y_k$ by Theorem 7. Consequently, for every $\hat{v}\in Y_2\cdots Y_k$ we have $t_1(\hat{v}) = t_2(\hat{v})$. This means that for all $\hat{v}\in Y_2\cdots Y_k$ we have $\omega'_r(a') = \mathbb{T}$. This shows that ω'_r is a 3-valued model of φ'.

Suppose ω'_r is a 3-valued model of φ'. We need to show that r satisfies $\hat{\varphi}$. That is, $r = r[XY_1] \bowtie \cdots \bowtie r[XY_k]$ holds. According to Theorem 7 this is equivalent to showing that r satisfies the FD $X \to R - XY_i$ for some $i \in \{1,\dots,k\}$. Suppose that $t_1(X) = t_2(X)$ and t_1,t_2 are both X-total, otherwise there is nothing to show. This implies that $\omega'_r(a') = \mathbb{T}$ for all $\hat{v}\in X$. Assume that for $j = 2,\dots,k$, r violates $X \to R - XY_j$, otherwise there is nothing to show. Consequently, for all $j = 2,\dots,k$ there is some $\hat{v} \in R - XY_j$ such that $\omega'_r(a') = \mathbb{F}$. Since ω'_r satisfies φ' we must have $\omega'_r(a') = \mathbb{T}$ for all $\hat{v} \in Y_2 \cdots Y_k$. Hence, $t_1(Y_2 \cdots Y_k) = t_2(Y_2 \cdots Y_k)$, and r satisfies $X \to R - XY_1$. It follows that r satisfies $\hat{\varphi}$. □

The equivalence between two-tuple relations for FOHDs and special truth assignments extend the existing equivalence between two-tuple relations for multivalued dependencies and special truth assignments [14, 29, 51].

Corollary 2 and Lemma 3 allow us to establish the anticipated equivalence between two-tuple relations and propositional truth assignments.

Theorem 8. *Let $\hat{\Sigma} \cup \{\hat{\varphi}\}$ be a set of FOHDs over relation schema R with NFS R_s, and let $\Sigma' \cup \{\varphi'\}$ denote the set of its corresponding formulae over L with the set $\mathscr{S}$. Then $\hat{\Sigma} \models_{R_s} \hat{\varphi}$ if and only if $\Sigma' \models^3_{\mathscr{S}} \varphi'$.*

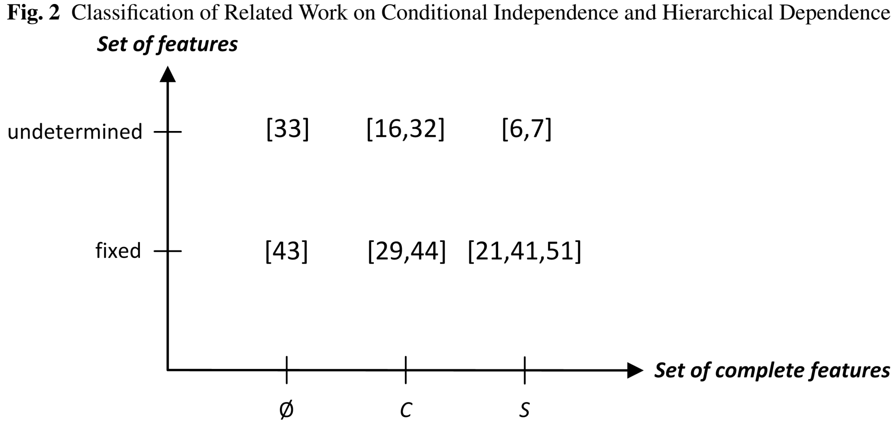
Proof. Based on Corollary 2 it remains to establish the equivalence between $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$ and $\Sigma' \models^3_{\mathscr{S}} \varphi'$. Suppose first that $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$ does not hold. Then there is some R_s-total relation r over R that satisfies every FOHD $\hat{\sigma}$ in $\hat{\Sigma}$ but violates $\hat{\varphi}$. Let ω'_r denote the special 3-valued truth assignment associated with r. By definition ω'_r is an $\mathscr{S}$-3 interpretation. By Lemma 3 it follows that ω'_r is an $\mathscr{S}$-3 model of every formula σ' in Σ' but not an $\mathscr{S}$-3 model of φ'. Consequently, $\Sigma' \models^3_{\mathscr{S}} \varphi'$ does not hold. Suppose now that $\Sigma' \models^3_{\mathscr{S}} \varphi'$ does not hold. Then there is some $\mathscr{S}$-3 truth assignment ω' over L that is an $\mathscr{S}$-3 model for every formula σ' in Σ', but not an $\mathscr{S}$-3 model for the formula φ'. Define the following two-tuple relation $r = \{t_1, t_2\}$ over R: for all $\hat{v} \in R$, let $\mathtt{ni} \neq t_1(\hat{v}) = t_2(\hat{v}) \neq \mathtt{ni}$, if $\omega(a') = \mathbb{T}$ and $\omega(\neg a') = \mathbb{F}$; let $t_1(\hat{v}) = \mathtt{ni} = t_2(\hat{v})$, if $\omega(a') = \mathbb{T} = \omega(\neg a')$; and let $\mathtt{ni} \neq t_1(\hat{v}) \neq t_2(\hat{v}) \neq \mathtt{ni}$, if $\omega(a') = \mathbb{F}$ and $\omega(\neg a') = \mathbb{T}$. In particular, it follows that $\omega'_r = \omega'$. By Lemma 3 it follows that r satisfies every FOHD $\hat{\sigma}$ in $\hat{\Sigma}$ but violates $\hat{\varphi}$. In addition, r is R_s-total since the construction ensures that the null marker `ni` can only occur on attributes outside of R_s. Hence, $\hat{\Sigma} \models_{2,R_s} \hat{\varphi}$ does not hold. □

Examples 9 and 13 illustrate the equivalences between the implication problem of FOHDs and NFSs and the implication problem of formulae in $\mathfrak{F}$ under $\mathscr{S}$-3 interpretations.

6 Related Work

Dawid [11] has started to investigate fundamental properties of conditional independence, leading to a claim that "rather than just being another useful tool in the statistician's kitbag, conditional independence offers a new language for the expression of statistical concepts and a framework for their study". Geiger and Pearl [20, 21, 49] have systematically investigated the implication problem for fragments of conditional independence statements over different probability models. In particular, they have established an axiomatization of saturated conditional independence (SCI) statements by a finite set of Horn rules [20]. Studený [56] showed that no axiomatization by a finite set of Horn rules exists for general conditional independence statements. Recently, Naumov and Nicholls [46] established a complete infinite recursively enumerable axiomatization of the propositional theory for conditional independence statements. Niepert et al. [47, 48] established an axiomatization for stable conditional independence statements, which subsume saturated statements, and showed that their associated implication problem is coNP-complete. Recently, this line of work has been extended to incomplete data, in which the implication problem changes [6, 32, 33, 34, 41, 42, 43, 44]. Figure 2 shows a classification of this work by distinguishing between implication problems in fixed and undetermined sets of attributes, random variables, or propositional variables, respectively (referred to collectively as features), and by distinguishing between the sets of features that can be declared complete (either only the empty set $\emptyset$, or the entire set S, or an arbitrary subset C of S). The present article is a summary of the results and techniques applied to fixed sets of features. Similar results hold when the set of features remains undetermined [6, 32, 33]. In particular, the results establish strong bonds with database semantics and approximation logics.

Fig. 2 Classification of Related Work on Conditional Independence and Hierarchical Dependence



In fact, database theory has studied more than 100 different classes of database dependencies [57] over strictly relational data, where incomplete data must not occur. These dependencies enforce the semantics of application domains within a database system [38]. Here, multivalued dependencies [15] are an expressive class whose implication problem can be decided in almost linear time [4, 18, 50]. In particular, they form the basis for the Fourth Normal Form in database design which characterizes database schemata whose instances are free from data redundancy [15, 60, 62]. The implication problem of multivalued dependencies is equivalent to that of a Boolean propositional fragment [52], and to that of SCI statements [61]. Furthermore, it is known that the equivalence between MVD implication and that of their corresponding propositional counterpart cannot be extended to an equivalence between the implication problem of embedded MVDs and that of any Boolean propositional fragment [52]. We also note that the implication problem of embedded multivalued dependencies is undecidable [30, 31] and not axiomatizable by a finite set of Horn rules [55]. Studený also showed that the implication problem of embedded MVDs and that of CI statements do not coincide [56]. Again, this line of work has been extended to incomplete data [16, 25, 28, 29, 35] and the present article can be understood as a summary of these findings.

It is important to point out that the results in this article can be proven more directly in different ways. Firstly, for a set $\Sigma \cup \{I(Y_1,\ldots,Y_k \mid X)\}$ of GSCI statements over S with $C \subseteq S$ it holds that $\Sigma \models_C I(Y_1,\ldots,Y_k \mid X)$ if and only if $\Sigma[XC] \models_S I(Y_1,\ldots,Y_k \mid X)$, where $\Sigma[U] = \{I(W_1,\ldots,W_m \mid V) \in \Sigma \mid V \subseteq U\}$. This embedding translates every instance of the implication problem for GSCI statements and complete random variables into an instance of the implication problem for GSCI statements. This illustrates the significance of the special case where $C = S$. Secondly, every instance of the implication problem for GSCI statements and complete random variables can be translated into an instance of an implication problem for SCI statements and complete random variables, see Remark 6. Finally, the results for full first-order hierarchical dependencies from Section 5 can be obtained by exploiting a strong correspondence between relations that satisfy an FOHD $X : [Y_1|\cdots|Y_k]$ and probability models that satisfy $I(Y_1,\ldots,Y_k \mid X)$. For instance, a two tuple relation $r = \{t_1,t_2\}$ satisfies $X : [Y_1|\cdots|Y_k]$ if and only if the special probability model $\tau(r)$ satisfies the GSCI statement $I(Y_1,\ldots,Y_k \mid X)$, where $\tau(r)$ is obtained by stipulating $P(t_1) = 0.5 = P(t_2)$. Vice versa, the special probability model π satisfies the GSCI statement $I(Y_1,\ldots,Y_k \mid X)$ if and only if the relation $\tau'(\pi)$ satisfies the FOHD $X : [Y_1|\cdots|Y_k]$, where the two tuples in $\tau'(\pi)$ are simply the two assignments of π that have probability one half. Nevertheless, the main focus of this article is not on the results for GSCI statements, but on the techniques used to obtain them.

7 Conclusion

Conditional independence is a core concept in disciplines as diverse as artificial intelligence, databases, probability theory, and statistics. The implication problem for conditional independence statements is paramount for many applications including Bayesian networks and database design. It is known that the implication problem for general conditional independence statements cannot be axiomatized by a finite set of Horn rules, and is coNP-complete to decide for their stable fragment, already in the idealized case where all data is complete. This article showcases the equivalences between three different implication problems: i) generalized saturated conditional independence statements in the presence of a set of complete random variables, ii) a fragment of propositional logic under $\mathscr{S}$-3 interpretations, and iii) Delobel's class of full first-order hierarchical database dependencies in the presence of a set of attributes declared `NOT NULL`. Axiomatizations in the form of finite sets of Horn rules were established, and algorithms to decide the associated implication problems in almost linear time are also available [44]. The key to these equivalences are special probability models and two-tuple relations. It is further known that none of these equivalences holds between the frameworks of conditional independence statements, any fragment of Boolean propositional logic, and general first-order hierarchical dependencies, already in the case of complete data [44].

This body of work is a strong advocate for investigating notions of dependence and independence as first-class citizens within standard frameworks for reasoning, as successfully started in dependence and independence logics.

Acknowledgements I would like to thank the anonymous reviewer for feedback that largely improved the presentation of the paper. This research is supported by the Marsden Fund Council from Government funding, administered by the Royal Society of New Zealand.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley, Boston, MA, USA (1995)
2. Abramsky, S., Kontinen, J., Väänänen, J.A., Vollmer, H.: Dependence logic: Theory and applications (Dagstuhl seminar 13071). Dagstuhl Reports **3**(2), 45–54 (2013)
3. Atzeni, P., Morfuni, N.: Functional dependencies and constraints on null values in database relations. Information and Control **70**(1), 1–31 (1986)
4. Beeri, C.: On the membership problem for functional and multivalued dependencies in relational databases. ACM Trans. Database Syst. **5**(3), 241–259 (1980)
5. Beeri, C., Fagin, R., Howard, J.H.: A complete axiomatization for functional and multivalued dependencies in database relations. In: Proceedings of the SIGMOD International Conference on Management of Data, pp. 47–61. ACM, Toronto, Canada (1977)
6. Biskup, J., Hartmann, S., Link, S.: Probabilistic conditional independence under schema certainty and uncertainty. In: Proceedings of the 6th International Conference on Scalable Uncertainty Management (SUM), *Lecture Notes in Computer Science*, vol. 7520, pp. 365–378. Springer (2012)
7. Biskup, J., Link, S.: Appropriate inferences of data dependencies in relational databases. Ann. Math. Artif. Intell. **63**(3-4), 213–255 (2012)
8. Codd, E.F.: Extending the database relational model to capture more meaning. ACM Trans. Database Syst. **4**(4), 397–434 (1979)
9. Codd, E.F.: Missing information (applicable and inapplicable) in relational databases. SIGMOD Record **15**(4), 53–78 (1986)
10. Date, C., Darwen, H.: A guide to the SQL standard. Addison-Wesley Professional, Reading, MA, USA (1997)
11. Dawid, A.P.: Conditional independence in statistical theory. Journal of the Royal Statistical Society. Series B (Methodological) **41**(1), 1–31 (1979)
12. Delobel, C.: Normalization and hierarchical dependencies in the relational data model. ACM Trans. Database Syst. **3**(3), 201–222 (1978)
13. Durand, A., Kontinen, J.: Hierarchies in dependence logic. ACM Trans. Comput. Log. **13**(4), 31 (2012)
14. Fagin, R.: Functional dependencies in a relational data base and propositional logic. IBM Journal of Research and Development **21**(6), 543–544 (1977)
15. Fagin, R.: Multivalued dependencies and a new normal form for relational databases. ACM Trans. Database Syst. **2**(3), 262–278 (1977)
16. Ferrarotti, F., Hartmann, S., Link, S.: Reasoning about functional and full hierarchical dependencies over partial relations. Inf. Sci. **235**, 150–173 (2013)
17. Fienberg, S.: The analysis of incomplete multi-way contingency tables. Biometrics **28**(1), 177–202 (1972)

18. Galil, Z.: An almost linear-time algorithm for computing a dependency basis in a relational database. J. ACM **29**(1), 96–102 (1982)
19. Galliani, P.: Inclusion and exclusion dependencies in team semantics - on some logics of imperfect information. Ann. Pure Appl. Logic **163**(1), 68–84 (2012)
20. Geiger, D., Pearl, J.: Logical and algorithmic properties of independence and their application to bayesian networks. Ann. Math. Artif. Intell. **2**, 165–178 (1990)
21. Geiger, D., Pearl, J.: Logical and algorithmic properties of conditional independence and graphical models. The Annals of Statistics **21**(4), 2001–2021 (1993)
22. Grädel, E., Väänänen, J.A.: Dependence and independence. Studia Logica **101**(2), 399–410 (2013)
23. Graetzer, G.: General lattice theory. Birkhäuser, Boston, USA (1998)
24. Halpern, J.: Reasoning about uncertainty. MIT Press (2005)
25. Hartmann, S., Kirchberg, M., Link, S.: Design by example for SQL table definitions with functional dependencies. The VLDB Journal **21**(1), 121–144 (2012)
26. Hartmann, S., Link, S.: On a problem of Fagin concerning multivalued dependencies in relational databases. Theor. Comput. Sci. **353**(1-3), 53–62 (2006)
27. Hartmann, S., Link, S.: Characterising nested database dependencies by fragments of propositional logic. Ann. Pure Appl. Logic **152**(1-3), 84–106 (2008)
28. Hartmann, S., Link, S.: When data dependencies over SQL tables meet the Logics of Paradox and S-3. In: Proceedings of the 29th ACM SIGMOD-SIGART-SIGACT Symposium on Principles of Database Systems (PoDS), pp. 317–326. ACM, Indianapolis, USA (2010)
29. Hartmann, S., Link, S.: The implication problem of data dependencies over SQL table definitions: axiomatic, algorithmic and logical characterizations. ACM Trans. Database Syst. **37**(2), Article 13 (2012)
30. Herrmann, C.: On the undecidability of implications between embedded multivalued database dependencies. Inf. Comput. **122**(2), 221–235 (1995)
31. Herrmann, C.: Corrigendum to "on the undecidability of implications between embedded multivalued database dependencies". Inf. Comput. **204**(12), 1847–1851 (2006)
32. Köhler, H., Link, S.: Logics for approximating implication problems of saturated conditional independence. In: E. Fermé, J. Leite (eds.) Logics in Artificial Intelligence - 14th European Conference, JELIA 2014, Funchal, Madeira, Portugal, September 24-26, 2014. Proceedings, *Lecture Notes in Computer Science*, vol. 8761, pp. 224–238. Springer (2014)
33. Köhler, H., Link, S.: Saturated conditional independence with fixed and undetermined sets of incomplete random variables. In: N. Zhang, J. Tian (eds.) Uncertainty in Artificial Intelligence - 30th International Conference, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014. Proceedings, pp. 224–238. AUAI Press (2014)
34. Kontinen, J., Link, S., Väänänen, J.A.: Independence in database relations. In: Proceedings of the 20th International Workshop on Logic, Language, Information, and Computation (WoLLIC), *Lecture Notes in Computer Science*, vol. 8071, pp. 179–193. Springer (2013)
35. Le, V.B.T., Link, S., Ferrarotti, F.: Effective recognition and visualization of semantic requirements by perfect sql samples. In: Proceedings of the 32th International Conference on Conceptual Modeling (ER), *Lecture Notes in Computer Science*, vol. 8217, pp. 227–240. Springer (2013)
36. Levesque, H.: A knowledge-level account of abduction. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1061–1067. Morgan Kaufmann, Detroit, USA (1989)
37. Lien, E.: On the equivalence of database models. J. ACM **29**(2), 333–362 (1982)
38. Link, S.: Consistency enforcement in databases. In: L.E. Bertossi, G.O.H. Katona, K. Schewe, B. Thalheim (eds.) Semantics in Databases, Second International Workshop, Dagstuhl Castle, Germany, January 7-12, 2001, Revised Papers, *Lecture Notes in Computer Science*, vol. 2582, pp. 139–159. Springer (2003)
39. Link, S.: Charting the completeness frontier of inference systems for multivalued dependencies. Acta Inf. **45**(7-8), 565–591 (2008)
40. Link, S.: Characterizations of multivalued dependency implication over undetermined universes. J. Comput. Syst. Sci. **78**(4), 1026–1044 (2012)

41. Link, S.: Propositional reasoning about saturated conditional probabilistic independence. In: Proceedings of the 19th International Workshop on Logic, Language, Information and Computation (WoLLIC), *Lecture Notes in Computer Science*, vol. 7456, pp. 257–267. Springer, Buenos Aires, Argentina (2012)
42. Link, S.: Approximate reasoning about generalized conditional independence with complete random variables. In: Proceedings of the 7th International Conference on Scalable Uncertainty Management (SUM), *Lecture Notes in Computer Science*, vol. 8078, pp. 269–282. Springer (2013)
43. Link, S.: Reasoning about saturated conditional independence under uncertainty: Axioms, algorithms, and levesque's situations to the rescue. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI). AAAI Press (2013)
44. Link, S.: Sound approximate reasoning about saturated conditional probabilistic independence under controlled uncertainty. J. Applied Logic **11**(3), 309–327 (2013)
45. Lohmann, P., Vollmer, H.: Complexity results for modal dependence logic. Studia Logica **101**(2), 343–366 (2013)
46. Naumov, P., Nicholls, B.: R.E. axiomatization of conditional independence. In: Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge (TARK), pp. 148–155 (2013)
47. Niepert, M., Gyssens, M., Sayrafi, B., Gucht, D.V.: On the conditional independence implication problem: A lattice-theoretic approach. Artif. Intell. **202**, 29–51 (2013)
48. Niepert, M., Van Gucht, D., Gyssens, M.: Logical and algorithmic properties of stable conditional independence. Int. J. Approx. Reasoning **51**(5), 531–543 (2010)
49. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Franciso, U.S.A. (1988)
50. Sagiv, Y.: An algorithm for inferring multivalued dependencies with an application to propositional logic. J. ACM **27**(2), 250–262 (1980)
51. Sagiv, Y., Delobel, C., Parker Jr., D.S., Fagin, R.: An equivalence between relational database dependencies and a fragment of propositional logic. J. ACM **28**(3), 435–453 (1981)
52. Sagiv, Y., Delobel, C., Parker Jr., D.S., Fagin, R.: Correction to "An equivalence between relational database dependencies and a fragment of propositional logic". J. ACM **34**(4), 1016–1018 (1987)
53. Schaerf, M., Cadoli, M.: Tractable reasoning via approximation. Artif. Intell. **74**, 249–310 (1995)
54. Stefanini, F.: Graphical models for eliciting structural information. In: Classification and Data Mining, pp. 139–146. Springer (2013)
55. Stott Parker Jr., D., Parsaye-Ghomi, K.: Inferences involving embedded multivalued dependencies and transitive dependencies. In: Proceedings of the International Conference on Management of Data (SIGMOD), pp. 52–57. ACM (1980)
56. Studený, M.: Conditional independence relations have no finite complete characterization. In: S. Kubik, J. Visek (eds.) Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, pp. 377–396. Kluwer (1992)
57. Thalheim, B.: Dependencies in relational databases. Teubner (1991)
58. Väänänen, J.A.: Dependence Logic - A New Approach to Independence Friendly Logic, *London Mathematical Society student texts*, vol. 70. Cambridge University Press (2007)
59. Väänänen, J.A., Hodges, W.: Dependence of variables construed as an atomic formula. Ann. Pure Appl. Logic **161**(6), 817–828 (2010)
60. Vincent, M.: Semantic foundations of 4NF in relational database design. Acta Inf. **36**(3), 173–213 (1999)
61. Wong, S., Butz, C., Wu, D.: On the implication problem for probabilistic conditional independency. Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans **30**(6), 785–805 (2000)
62. Wu, M.: The practical need for fourth normal form. In: Proceedings of the Twenty-third Technical Symposium on Computer Science Education (SIGCSE), pp. 19–23. ACM (1992)
63. Zaniolo, C.: Database relations with null values. J. Comput. Syst. Sci. **28**(1), 142–166 (1984)