Saturated Conditional Independence with Fixed and Undetermined Sets of Incomplete Random Variables

Henning Koehler
School of Engineering & Advanced Technology
Massey University, New Zealand
h.koehler@massey.ac.nz

Sebastian Link
Department of Computer Science
The University of Auckland, New Zealand
s.link@auckland.ac.nz

Abstract

The implication problem for saturated conditional independence statements is studied in the presence of fixed and undetermined sets of incomplete random variables. Here, random variables are termed incomplete since they admit missing data. Two different notions of implication arise. In the classic notion of V-implication, a statement is implied jointly by a set of statements and a fixed set V of random variables. In the alternative notion of pure implication, a statement is implied by a given set of statements alone, leaving the set of random variables undetermined. A first axiomatization for V-implication is established that distinguishes purely implied from V-implied statements. Axiomatic, algorithmic and logical characterizations of pure implication are established. Pure implication appeals to applications in which the existence of random variables is uncertain, for example, when independence statements are integrated from different sources, when random variables are unknown or shall remain hidden.

1 INTRODUCTION

The concept of conditional independence (CI) 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 [Darwiche (2009); Dawid (1979); Pearl (1988)]. Application areas include natural language processing, speech processing, computer vision, robotics, computational biology, and error-control coding [Darwiche (2009); Halpern (2005); Niepert et al. (2013)]. Central to these applications is the implication problem, which is to decide for an arbitrary set V of random variables, and an arbitrary set $\Sigma \cup \{\varphi\}$ of CI statements over V, whether every probability model that satisfies every element in Σ also satisfies φ. Indeed, non-implied CI statements represent new opportunities to construct complex probability models with polynomially many parameters and to efficiently organize distributed probability computations [Geiger and Pearl (1993)]. An algorithm for deciding the implication problem can also test the consistency of independence and dependence statements collected from different sources; which is particularly important as these statements often introduce non-linear constraints resulting in unfeasible CSP instances [Geiger and Pearl (1993); Niepert et al. (2013)]. While the decidability of the implication problem for CI statements relative to discrete probability measures remains open, it is not axiomatizable by a finite set of Horn rules [Studený (1992)] and already coNP-complete for stable CI statements [Niepert, Van Gucht, and Gyssens (2010)]. An important subclass are therefore saturated CI (SCI) statements, in which all given random variables occur. In fact, graph separation and SCI statements enjoy the same axioms [Geiger and Pearl (1993)], and the implication problem of SCI statements is decidable in almost linear time [Galil (1982)]. These results contribute to the success story of Bayesian networks in AI and machine learning [Darwiche (2009); Geiger and Pearl (1993)], and have recently been carried over to the presence of missing data [Link (2013a)]. Here, independence is not judged on conditions that carry missing data. The findings complement a long line of AI research on the recognized need to reveal missing data and to explain where they come from, e.g. [Chickering and Heckerman (1997); Dempster, Laird, and Rubin (1977); Friedman (1997); Lauritzen (1995); Marlin et al. (2011); Saar-Tsechansky and Provost (2007); Singh (1997); Zhu et al. (2007)]. It is important to realize that implication problems of SCI statements in the presence of missing data differ from implication problems in the absence of data. For an illustration, consider a simplified burglary example. A *r(obbery)* sets off an *a(larm)* causing *s(heldon)* or *b(atman)* to call security. The independence between *sb* and *r*, given *a*, can be stated as the SCI statement $I(sb, r|a)$ over $V = \{b, a, r, s\}$. In the absence of missing data, $I(s, b|ar)$ and $I(sb, r|a)$ together do V-imply $I(s, br|a)$. With missing data present, however, $I(s, b|ar)$

and $I(sb, r|a)$ together do not V-imply $I(s, br|a)$:

r	a	b	s	P
$-$	true	true	true	0.5
$-$	true	false	false	0.5

Here, $I(s, b|ar)$ is satisfied as the assignments on the condition ar involve missing data, represented by $-$.

Most of the literature on the implication problem for SCI statements have focused on the notion of implication in which the underlying set V of random variables is assumed to be fixed. However, the assumption that V is fixed may not be practical: for example, the fact that not all random variables are known yet should not prevent us from declaring some independence statements; or even if we know all random variables, we may not want to disclose all of them; or when independence statements are integrated from different sources. Instead, we may want to state that given a, sb is independent from the set of remaining random variables, no matter what they are. This statement could be written as $I(sb|a)$. The intriguing point here is the difference between declaring $I(sb|a)$ and declaring $I(r|a)$ when V is left undetermined. In fact, the probability model

r	a	b	s	e	P
true	true	$-$	$-$	true	0.5
false	true	$-$	$-$	false	0.5

satisfies $I(sb|a)$, but does not satisfy $I(r|a)$. We conclude that $I(sb|a)$ implies $I(r|a)$ for the fixed set V, but $I(sb|a)$ does not imply $I(r|a)$ when the set of random variables is left undetermined.

The example illustrates the need to distinguish between different notions of semantic implication. The first notion is that of V-implication. For example, Link (2013a) established an axiomatization $\mathfrak{U}_V$ for the V-implication problem of SCI statements in the presence of missing data. The alternative, stronger notion of pure implication leaves the set of random variables undetermined: the pure implication problem is to decide for every given set $\Sigma \cup \{\varphi\}$ of SCI statements, whether for every probability model π that involves at least all the random variables in $\Sigma \cup \{\varphi\}$ and that satisfies Σ, π also satisfies φ. Pure implication allows us to use independence statements without knowing all the random variables. This lowers barriers for their use and makes them applicable in demanding frameworks where some variables shall remain unknown for some users and where we still want to know how complex probability distributions can be organized efficiently. That is, pure implication enables us to reason under uncertainty about the random variables, while V-implication does not. For illustration, suppose we want to keep the random variable r hidden. Then it is impossible to reason about SCI statements under the notion of V-implication. With pure implication we can still state $I(sb|a)$ and $I(b|a)$, and our results show that we can even conclude $I(s|a)$ from that.

Contribution. In Section 2 we show that the only existing finite axiomatization $\mathfrak{U}_V$ for the V-implication of SCI statements cannot distinguish between purely implied and V-implied SCI statements. That is, there are purely implied SCI statements for which every inference by $\mathfrak{U}_V$ applies the V-symmetry rule; giving incorrectly the impression that the pure implication of an SCI statement depends on V. In Section 3 we establish a finite axiomatization $\mathfrak{C}_V$ such that every purely implied SCI statement can be inferred without any application of the V-symmetry rule; every V-implied SCI statement can be inferred with only a single application of the V-symmetry rule, and this application is done in the last step of the inference. In Section 4 we establish a finite axiomatization $\mathfrak{C}$ for the pure implication of SCI statements. As $\mathfrak{C}$ results from $\mathfrak{C}_V$ by removal of the symmetry rule, the results show that the symmetry rule is only necessary to infer those SCI statements that are V-implied but not implied. In Section 5, pure implication is characterized by V-implication where V involves random variables that do not occur in any of the given SCI statements. In Sections 6, 7 and 8 this result is exploited to characterize the pure implication problem i) logically by a propositional fragment under interpretations by Levesque's situations, ii) by multivalued database dependencies involving missing data, and iii) by an algorithm that decides pure implication in almost linear time. Related work is discussed in Section 9. We conclude in Section 10.

2 IMPLICATION UNDER FIXED SETS OF RANDOM VARIABLES

We summarize the semantics of CI statements in the presence of missing data from Link (2013a). A definition is given that embodies the ability of an axiomatization to separate V-implied from purely implied SCI statements. It is shown that the existing axiomatization $\mathfrak{U}_V$ for V-implication from Link (2013a) does not have this ability.

We denote by $\mathfrak{V}$ a countably infinite set of distinct symbols $\{v_1, v_2, \dots\}$ of *random variables*. A *domain mapping* is a mapping that associates a set, *dom*(v_i), with each random variable v_i of a finite set $V \subseteq \mathfrak{V}$. This set is called the *domain* of v_i and each of its elements is a *data value* of v_i. We assume that each domain *dom*(v_i) contains the element $-$, which we call the *marker*. Although we use the element $-$ like any other data value, we prefer to think of $-$ as a marker, denoting that no information is currently available about the data value of v_i. The interpretation of this marker as no information means that a data value does either not exist (known as a structural zero in statistics, and the null marker inapplicable in databases), or a data value exists but is currently unknown (known as a sampling zero in statistics, and the null marker applicable in databases). The disadvantage of using this interpretation is a loss in knowledge when representing data values known to not exist, or known to exist but currently unknown. This interpretation overcomes the computational difficulties when more expressive interpretations of missing data are used. As another key advantage one can represent missing data values, even if it is unknown whether they do not exist, or exist but are currently unknown. Strictly speaking, we shall call such random variables *incomplete* as their data values may be missing. For simplicity, we continue to speak off random variables for the remainder of this paper, although we really do mean incomplete random variables. For $X = \{v_1, \ldots, v_k\} \subseteq V$ we say that $\mathbf{a}$ is an assignment of X, if $\mathbf{a} \in \mathit{dom}(v_1) \times \cdots \times \mathit{dom}(v_k)$. For an assignment $\mathbf{a}$ of X we write $\mathbf{a}(y)$ for the projection of $\mathbf{a}$ onto $Y \subseteq X$. We say that $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$ is X*-complete*, if $\mathbf{a}_i \neq -$ for all $i = 1, \ldots, k$.

A *probability model* over a finite set $V = \{v_1, \ldots, v_n\}$ of random variables is a pair (dom, P) where dom is a domain mapping that maps each v_i to a finite domain $\mathit{dom}(v_i)$, and $P : \mathit{dom}(v_1) \times \cdots \times \mathit{dom}(v_n) \rightarrow [0, 1]$ is a probability distribution having the Cartesian product of these domains as its sample space.

The expression $I(Y, Z|X)$ where X, Y and Z are disjoint subsets of V is called a *conditional independence* (CI) *statement* over V. The set X is called the *condition* of $I(Y, Z|X)$. If $XYZ = V$, we call $I(Y, Z|X)$ a *saturated* CI (SCI) statement. Let (dom, P) be a probability model over V. Following Link (2013a), a CI statement $I(Y, Z|X)$ is said to *hold for* (dom, P) if for every complete assignment $\mathbf{x}$ of X, and for every assignment $\mathbf{y}, \mathbf{z}$ of Y and Z, respectively,

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \cdot P(\mathbf{x}) = P(\mathbf{x}, \mathbf{y}) \cdot P(\mathbf{x}, \mathbf{z}). \qquad (1)$$

Equivalently, (dom, P) is said to *satisfy* $I(Y, Z|X)$.

The satisfaction of $I(Y, Z|X)$ requires Equation 1 to hold for *complete* assignments $\mathbf{x}$ of X only. The reason is that the independence between an assignment $\mathbf{y}$ and an assignment $\mathbf{z}$ is conditional on the assignment $\mathbf{x}$. Indeed, in case there is *no information* about the assignment $\mathbf{x}$, then there should not be any requirement on the independence between $\mathbf{y}$ and $\mathbf{z}$.

SCI statements interact with one another, and these interactions have been formalized by the following notion of semantic implication. Let $\Sigma \cup \{\varphi\}$ be a set of SCI statements over V. We say that Σ V*-implies* φ, denoted by $\Sigma \models_V \varphi$, if every probability model over V that satisfies every SCI statement $\sigma \in \Sigma$ also satisfies φ. The V*-implication problem* is the following problem.

PROBLEM:	V-implication problem
INPUT:	Set V of random variables Set $\Sigma \cup \{\varphi\}$ of SCI statements over V
OUTPUT:	Yes, if $\Sigma \models_V \varphi$; No, otherwise

Table 1: Axiomatization $\mathfrak{U}$ under Incomplete RVs

$\dfrac{}{I(V - X, \emptyset \| X)}$ (triviality, $\mathcal{T}'$)	$\dfrac{I(Y, Z \| X)}{I(Z, Y \| X)}$ (symmetry, $\mathcal{S}$)
$\dfrac{I(YZ, UW \| X) \; I(YU, ZW \| X)}{I(YZU, W \| X)}$ (algebra, $\mathcal{A}'$)	$\dfrac{I(Y, ZW \| X)}{I(Y, Z \| XW)}$ (weak union, $\mathcal{W}'$)

For Σ we let $\Sigma^*_V = \{\varphi \mid \Sigma \models_V \varphi\}$ be the *semantic closure* of Σ, i.e., the set of all SCI statements V-implied by Σ. In order to determine the V-implied SCI statements we use a syntactic approach by applying inference rules. These inference rules have the form

$$\frac{\text{premises}}{\text{conclusion}}$$

and inference rules without any premises are called axioms. An inference rule is called V*-sound*, if the premises of the rule V-imply the conclusion of the rule. We let $\Sigma \vdash_{\mathfrak{R}} \varphi$ denote the *inference* of φ from Σ by the set $\mathfrak{R}$ of inference rules. That is, there is some sequence $\gamma = [\sigma_1, \ldots, \sigma_n]$ of SCI 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 V*-sound* (V*-complete*) for the V-implication of SCI statements, if for every V and for every set Σ of SCI statements over V, we have $\Sigma^+_{\mathfrak{R}} \subseteq \Sigma^*_V$ ($\Sigma^*_V \subseteq \Sigma^+_{\mathfrak{R}}$). The (finite) set $\mathfrak{R}$ is said to be a (finite) *axiomatization* for the V-implication of SCI statements if $\mathfrak{R}$ is both V-sound and V-complete.

Table 1 contains the set $\mathfrak{U} = \{\mathcal{T}', \mathcal{S}, \mathcal{A}', \mathcal{W}'\}$ of inference rules that form a finite axiomatization for the V-implication of SCI statements under incomplete random variables, as established in Link (2013a).

Motivated by the introductory remarks we now write $I(Y|X)$ instead of writing $I(V - XY, Y|X)$ for an SCI statement over V. It is first shown that the system $\mathfrak{U}_V = \{\mathcal{T}, \mathcal{S}, \mathcal{A}, \mathcal{W}\}$ from Table 2 forms a finite axiomatization for the V-implication of such SCI statements under incomplete random variables.

Proposition 1 *$\mathfrak{U}_V$ is a finite axiomatization for the V-implication of SCI statements under incomplete random variables.*

Proof Let $V \subseteq \mathfrak{V}$ be a finite set of random variables. Let $\Sigma = \{I(Y_1|X_1), \ldots, I(Y_n|X_n)\}$ and $\varphi = I(Y|X)$ be a (set of) SCI statement(s) over V. We can show by an induction over the inference length that $\Sigma \vdash_{\mathfrak{U}_V} \varphi$ if

Table 2: Axiomatization $\mathfrak{U}_V$ under Incomplete RVs

$\dfrac{}{I(\emptyset \mid X)}$ (triviality, $\mathcal{T}$)	$\dfrac{I(Y \mid X)}{I(V-XY \mid X)}$ (V-symmetry, $\mathcal{S}_V$)
$\dfrac{I(Y \mid X) \quad I(Z \mid X)}{I(YZ \mid X)}$ (union, $\mathcal{U}$)	$\dfrac{I(Y \mid X)}{I(Y-Z \mid XZ)}$ (weak union, $\mathcal{W}$)

and only if $\Sigma' = \{I(Y_1, V - X_1Y_1|X_1), \ldots, I(Y_n, V - X_nY_n|X_n)\} \vdash_{\mathfrak{U}} I(V - XY, Y|X)$. Hence, the V-soundness (V-completeness) of $\mathfrak{U}_V$ follows from the the V-soundness (V-completeness) of $\mathfrak{U}$. ▮

Example 2 *Consider $\Sigma = \{I(sb|a), I(b|a)\}$ and $\varphi = I(s|a)$ as a (set of) SCI statement(s) over $V = \{b, a, r, s\}$. Then $\Sigma \models_V \varphi$ as we can show, for example, by the following inference:*

$$\cfrac{\cfrac{\cfrac{I(sb|a)}{\mathcal{S}_V:\ I(r|a)} \qquad I(b|a)}{\mathcal{U}:\ I(rb|a)}}{\mathcal{S}_V:\ I(s|a)}\ .$$

However, since the inference applies the V-symmetry rule it is not clear whether φ is implied by Σ alone, that is, whether it is true that for all V' that include at least a, s, b it holds that $\Sigma \models_{V'} \varphi$. In fact, if we were to find an inference of φ from Σ by $\mathfrak{U}_V$ that never applies the V-symmetry rule $\mathcal{S}_V$, then we would know that φ is not only V-implied by Σ but even implied by Σ alone.

The last example motivates the following definition. It addresses the property of an inference system to first infer all those SCI statements implied by a set of SCI statements alone, without any application of the symmetry rule, and, subsequently, apply the V-symmetry rule once to some of these SCI statements to infer all V-implied SCI statements that do depend on the underlying set V of random variables.

Definition 3 *Let $\mathfrak{S}_V$ denote a set of inference rules that is V-sound for the V-implication of SCI statements, and in which the V-symmetry rule $\mathcal{S}_V$ is the only inference rule that is dependent on V. We say that $\mathfrak{S}_V$ is* conscious of pure implication*, if for every V, and every set $\Sigma \cup \{\varphi\}$ of SCI statements over V such that φ is V-implied by Σ there is some inference of φ from Σ by $\mathfrak{S}_V$ such that the V-symmetry rule $\mathcal{S}_V$ is applied at most once, and, if it is applied, then it is applied in the last step of the inference only.*

Example 2 and Definition 3 motivate the question if $\mathfrak{U}_V$ is conscious of pure implication.

Theorem 4 *$\mathfrak{U}_V$ is not conscious of pure implication.*

Proof Let $V = \{b, a, r, s\}$ and $\Sigma = \{I(b|a), I(bs|a)\}$. One can show that $I(s|a) \notin \Sigma^+_{\{\mathcal{T},\mathcal{W},\mathcal{U}\}}$. Moreover, for all Y such that $r \in Y$, $I(Y|a) \notin \Sigma^+_{\{\mathcal{T},\mathcal{W},\mathcal{U}\}}$, see Lemma 10 from Section 4. However, $I(s|a) \in \Sigma^+_{\mathfrak{U}_V}$ as shown in Example 2. Consequently, in any inference of $I(s|a)$ from Σ by $\mathfrak{U}_V$ the V-symmetry rule $\mathcal{S}_V$ must be applied at least once, but is not just applied in the last step as $r \in V - \{b, a, s\}$. ▮

In view of Theorem 4 it is natural to ask whether there is any axiomatization that is conscious of pure implication.

3 GAINING CONSCIOUSNESS

Theorem 4 has shown that axiomatizations are, in general, not conscious of pure implication. We will now establish a finite conscious axiomatization for the V-implication of SCI statements under incomplete random variables. For this purpose, we consider the *difference rule $\mathcal{D}$* as a new V-sound inference rule:

$$\frac{I(Y|X) \quad I(Z|X)}{I(Y - Z|X)}\ .$$

The V-soundness of the difference rule $\mathcal{D}$ follows easily from the algebra rule $\mathcal{A}'$.

Theorem 5 *Let Σ be a set of SCI statements over V. For every inference γ from Σ by the system $\mathfrak{U}_V = \{\mathcal{T}, \mathcal{S}_V, \mathcal{U}, \mathcal{W}\}$ there is an inference ξ from Σ by the system $\mathfrak{C}_V = \{\mathcal{T}, \mathcal{S}_V, \mathcal{U}, \mathcal{W}, \mathcal{D}\}$ such that*
1. γ and ξ infer the same SCI statement,
2. $\mathcal{S}_V$ is applied at most once in ξ,
3. if $\mathcal{S}_V$ is applied in ξ, then as the last rule.

Proof The proof is done by induction on the length l of γ. For $l = 1$, the statement $\xi := \gamma$ has the desired properties. Suppose for the remainder of the proof that $l > 1$, and let $\gamma = [\sigma_1, \ldots, \sigma_l]$ be an inference of σ_l from Σ by $\mathfrak{U}_V$. We distinguish between four different cases according to how σ_l is obtained from $[\sigma_1, \ldots, \sigma_{l-1}]$.

Case 1. σ_1 is obtained from the triviality axiom $\mathcal{T}$, or is an element of Σ. In this case, $\xi := [\sigma_l]$ has the desired properties.

Case 2. We obtain σ_l by an application of the weak union rule $\mathcal{W}$ to a premise σ_i with $i < l$. Let ξ_i be obtained by applying the induction hypothesis to $\gamma_i = [\sigma_1, \ldots, \sigma_i]$. Consider the inference $\xi := [\xi_i, \sigma_l]$. If in ξ_i the V-symmetry rule $\mathcal{S}_V$ is not applied, then ξ has the desired properties. If in ξ_i the $\mathcal{S}_V$ is applied as the last rule, then the last two steps in ξ are of the following form:

$$\cfrac{\cfrac{I(Y|X)}{\mathcal{S}_V:\ I(V - XY|X)}}{\mathcal{W}:\ I(V - XYZ|XZ)}\ .$$

However, these steps can be replaced as follows:

$$\cfrac{\cfrac{I(Y|X)}{\mathcal{W}:\ I(Y-Z|XZ)}}{\mathcal{S}_V:\ I(V-XYZ|XZ)}\ .$$

The resulting inference has the desired properties.

Case 3. We obtain σ_l by an application of the union rule $\mathcal{U}$ to premises σ_i and σ_j with $i,j < l$. Let ξ_i and ξ_j be obtained by applying the induction hypothesis to $\gamma_i = [\sigma_1,\ldots,\sigma_i]$ and $\gamma_j = [\sigma_1,\ldots,\sigma_j]$, respectively. Consider the inference $\xi := [\xi_i,\xi_j,\sigma_l]$. We distinguish between four cases according to the occurrence of the V-symmetry rule $\mathcal{S}_V$ in ξ_i and ξ_j.

Case 3.1. If $\mathcal{S}_V$ does not occur in ξ_i nor in ξ_j, then ξ has the desired properties.

Case 3.2. If $\mathcal{S}_V$ occurs in ξ_i as the last rule but does not occur in ξ_j, then the last step of ξ_i and the last step of ξ are of the following form:

$$\cfrac{\cfrac{I(Y|X)}{\mathcal{S}_V:\ I(V-XY|X)}\qquad I(Z|X)}{\mathcal{U}:\ I((V-XY)Z|X)}\ .$$

However, these steps can be replaced as follows:

$$\cfrac{\cfrac{I(Y|X)\qquad I(Z|X)}{\mathcal{D}:\ I(Y-Z|X)}}{\mathcal{S}_V:\ I(\underbrace{V-((Y-Z)X)}_{=(V-XY)Z}|X)}\ .$$

The resulting inference has the desired properties.

Case 3.3. If $\mathcal{S}_V$ occurs in ξ_j as the last rule but does not occur in ξ_i, then the last step of ξ_j and the last step of ξ are of the following form:

$$\cfrac{I(Y|X)\qquad \cfrac{I(Z|X)}{\mathcal{S}_V:\ I(V-XZ|X)}}{\mathcal{U}:\ I((V-XZ)Y|X)}\ .$$

However, these steps can be replaced as follows:

$$\cfrac{\cfrac{I(Z|X)\qquad I(Y|X)}{\mathcal{D}:\ I(Z-Y|X)}}{\mathcal{S}_V:\ I(\underbrace{V-((Z-Y)X)}_{=(V-XZ)Y}|X)}\ .$$

The resulting inference has the desired properties.

Case 3.4. If $\mathcal{S}_V$ occurs in ξ_i as the last rule and occurs in ξ_j as the last rule, then the last steps of ξ_i and ξ_j and the last step of ξ are of the following form:

$$\cfrac{\cfrac{I(Y|X)}{\mathcal{S}_V:\ I(V-XY|X)}\qquad \cfrac{I(Z|X)}{\mathcal{S}_V:\ I(V-XZ|X)}}{\mathcal{U}:\ I((V-XY)(V-XZ)|X)}\ .$$

However, these steps can be replaced as follows:

$$\cfrac{I(Y|X)\qquad \cfrac{I(Y|X)\qquad I(Z|X)}{\mathcal{D}:\ I(Y-Z|X)}}{\cfrac{\mathcal{D}:\ I(\underbrace{Y-(Y-Z)}_{=Y\cap Z}|X)}{\mathcal{S}_V:\ I(\underbrace{V-((Y\cap Z)X)}_{=(V-XY)(V-XZ)}|X)}}\ .$$

The resulting inference has the desired properties.

Case 4. We obtain σ_l by an application of the V-symmetry rule $\mathcal{S}_V$ to a premise σ_i with $i < l$. Let ξ_i be obtained by applying the induction hypothesis to $\gamma_i = [\sigma_1,\ldots,\sigma_i]$. Consider the inference $\xi := [\xi_i,\sigma_l]$. If in ξ_i the V-symmetry rule $\mathcal{S}_V$ is not applied, then ξ has the desired properties. If in ξ_i the V-symmetry rule $\mathcal{S}_V$ is applied as the last rule, then the last two steps in ξ are of the following form.

$$\cfrac{\cfrac{I(Y|X)}{\mathcal{S}_V:\ I(V-XY|X)}}{\mathcal{S}_V:\ I(\underbrace{V-(V-XY)X}_{=Y}|X)}$$

The inference obtained from deleting these steps has the desired properties. ■

Example 6 *Recall Example 2 where $V = \{b,a,r,s\}$, $\Sigma = \{I(sb|a), I(b|a)\}$ and $\varphi = I(s|a)$. While the inference of φ from Σ using $\mathfrak{A}_V$ in Example 2 showed that $\Sigma \models_V \varphi$ holds, it did leave open the question whether Σ purely implies φ. Indeed, no inference of φ from Σ by $\mathfrak{A}_V$ can provide this insight by Theorem 4. However, using $\mathfrak{C}_V$ we can obtain the following inference of φ from Σ:*

$$\cfrac{I(sb|a)\qquad I(b|a)}{\mathcal{D}:\ I(s|a)}\ .$$

Indeed, the V-symmetry rule $\mathcal{S}_V$ is unnecessary to infer φ from Σ.

Examples 2 and 6 indicate that the implication of $I(s|a)$ by $\Sigma = \{I(sb|a), I(b|a)\}$ does not depend on the fixed set V of random variables. In what follows we will formalize the stronger notion of pure implication as motivated in the introduction. Theorem 5 shows that the set $\mathfrak{C} := \mathfrak{C}_V - \{\mathcal{S}_V\}$ of inference rules is nearly V-complete for the V-implication of SCI statements under incomplete random variables.

Theorem 7 *Let $\Sigma \cup \{I(Y|X)\}$ be a set of SCI statements over the set V of incomplete random variables. Then $I(Y|X) \in \Sigma^+_{\mathfrak{C}_V}$ if and only if $I(Y|X) \in \Sigma^+_{\mathfrak{C}}$ or $I(V-XY|X) \in \Sigma^+_{\mathfrak{C}}$.* ■

Theorem 7 indicates that $\mathfrak{C}$ can infer every implied SCI statement that is independent from the set V of incomplete random variables. Another interpretation of Theorem 7 is the following. In using $\mathfrak{C}$ to infer V-implied statements, the fixation of V can be deferred until the last step of an inference.

4 PURE IMPLICATION

In this section we formalize the notion of pure implication as motivated in the introduction. It is shown that the set $\mathfrak{C}$ of inference rules forms a finite axiomatization for pure implication. On the one hand, this allows us to distinguish between V-implied and purely implied statements. On the other hand, the notion of pure implication can be applied whenever this notion of implication is more convenient to use, for examples, when there is uncertainty about additional random variables that may be required in the future, when some variables are unknown, or when some variables are meant to remain hidden.

A *probability model* is a triple (V, dom, P) where $V = \{v_1, \ldots, v_n\} \subseteq \mathfrak{V}$ is a finite set of incomplete random variables, *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. The expression $I(Y|X)$ where X and Y are finite, disjoint subsets of $\mathfrak{V}$ is called a *saturated conditional independence* (SCI) *statement*. We say that the SCI statement $I(Y|X)$ *holds for* (V, dom, P) if $XY \subseteq V$ and for every complete assignment $\mathbf{x}$ of X, every assignment $\mathbf{y}$ of Y, and every assignment $\mathbf{z}$ of $V - XY$, respectively,

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}) \cdot P(\mathbf{x}) = P(\mathbf{x}, \mathbf{y}) \cdot P(\mathbf{x}, \mathbf{z}).$$

Equivalently, (V, dom, P) is said to *satisfy* $I(Y|X)$. For an SCI statement $\sigma = I(Y|X)$ let $V_\sigma := XY$, and for a finite set Σ of SCI statements let $V_\Sigma := \bigcup_{\sigma \in \Sigma} V_\sigma$ denote the random variables that occur in it.

Definition 8 *Let $\Sigma \cup \{\varphi\}$ be a finite set of SCI statements. We say that Σ* purely implies *φ, denoted by $\Sigma \models \varphi$, if and only if every probability model (V, dom, P) with $V_{\Sigma \cup \{\varphi\}} \subseteq V$ that satisfies every SCI statement $\sigma \in \Sigma$ also satisfies φ.*

In the definition of pure implication the set of incomplete random variables is left undetermined. The only requirement is that the SCI statements must apply to the probability model. The pure implication problem for SCI statements can be stated as follows.

PROBLEM:	Pure Implication Problem
INPUT:	Set $\Sigma \cup \{\varphi\}$ of SCI statements
OUTPUT:	Yes, if $\Sigma \models \varphi$; No, otherwise

Pure implication is stronger than V-implication.

Table 3: Axiomatization $\mathfrak{C}$ for Pure Implication

$\dfrac{}{I(\emptyset \vert X)}$ (triviality, $\mathcal{T}$)	$\dfrac{I(Y \vert X)}{I(Y - Z \vert XZ)}$ (weak union, $\mathcal{W}$)
$\dfrac{I(Y \vert X) \quad I(Z \vert X)}{I(YZ \vert X)}$ (union, $\mathcal{U}$)	$\dfrac{I(Y \vert X) \quad I(Z \vert X)}{I(Y - Z \vert X)}$ (difference, $\mathcal{D}$)

Proposition 9 *Let $\Sigma \cup \{\varphi\}$ be a finite set of SCI statements, such that $V_{\Sigma \cup \{\varphi\}} \subseteq V$. If $\Sigma \models \varphi$, then $\Sigma \models_V \varphi$, but the other direction may fail.*

Proof The first statement follows directly from the definitions of pure and V-implication. For the other direction, let $V = \{b, a, r, s\}$, $\Sigma = \{I(r|a)\}$ and let φ be $I(sb|a)$. Clearly, Σ V-implies φ. However, Σ does not purely imply φ as the example from the introduction shows. ▮

Soundness and completeness for pure implication are defined as their corresponding notions in the context of some fixed set V by dropping the reference to V. While triviality axiom $\mathcal{T}$, weak union rule $\mathcal{W}$, and union rule $\mathcal{U}$ are all sound, the V-symmetry rule $\mathcal{S}_V$ is V-sound but not sound.

We shall now prove that $\mathfrak{C}$ forms a finite axiomatization for the pure implication of SCI statements. For this purpose, we prove two lemmata in preparation. The correctness of the first lemma can easily be observed by inspecting the inference rules in $\mathfrak{C}$. For each of the rules, every random variable that occurs on the left-hand side of the bar in the conclusion of the rule, already appears on the left-hand side of the bar in at least one premise of the rule.

Lemma 10 *Let $\Sigma = \{I(Y_1|X_1), \ldots, I(Y_n|X_n)\}$ be a finite set of SCI statements. If $I(Y|X) \in \Sigma^+_{\mathfrak{C}}$, then $Y \subseteq Y_1 \cup \ldots \cup Y_n$.* ▮

For the next lemma one may notice that the random variables that do not occur in V_Σ can always be introduced in the last step of an inference, by applying the weak union rule $\mathcal{W}$.

Lemma 11 *Let Σ be a finite set of SCI statements. If $I(Y|X) \in \Sigma^+_{\mathfrak{C}}$, then there is an inference $\gamma = [\sigma_1, \ldots, \sigma_l]$ of $I(Y|X)$ from Σ by $\mathfrak{C}$ such that every attribute occurring in $\sigma_1, \ldots, \sigma_{l-1}$ is an element of V_Σ.*

Proof Define $W := V_\Sigma$ and let $\bar{\xi} := [I(V_1|U_1), \ldots, I(V_{l-1}|U_{l-1})]$ be an inference of $I(Y|X)$ from Σ by $\mathfrak{C}$. Consider the sequence

$$\xi := [I(V_1 \cap W|U_1 \cap W), \ldots, I(V_{l-1} \cap W|U_{l-1} \cap W)] \, .$$

We claim that ξ is an inference of $I(Y \cap W|X \cap W)$ from Σ by $\mathfrak{C}$. For if $I(V_i|U_i)$ is an element of Σ or was obtained by an application of the triviality axiom $\mathcal{T}$, then $I(Y \cap W|X \cap W) = I(Y|X)$. One can verify that if $I(V_i|U_i)$ is the result of applying one of the rules $\mathcal{U}, \mathcal{W}, \mathcal{D}$, then $I(V_i \cap W|U_i \cap W)$ is the result of the same rule applied to the corresponding premises in ξ.

Now by Lemma 10 we know that $Y \subseteq W$, hence $Y \cap W = Y$. However, this means that we can infer $I(Y|X)$ from $I(Y \cap W|X \cap W)$ by a single application of the weak union rule $\mathcal{W}$:

$$\frac{I(Y \cap W|X \cap W)}{I(\underbrace{(Y \cap W) - X}_{=Y}|\underbrace{(X \cap W) \cup X}_{=X})}.$$

Hence, the inference $[\xi, I(Y|X)]$ has the desired properties. ▮

We are now prepared to prove the following result.

Theorem 12 *The set $\mathfrak{C} = \{\mathcal{T}, \mathcal{W}, \mathcal{U}, \mathcal{D}\}$ forms a finite axiomatization for the pure implication of SCI statements under incomplete random variables.*

Proof Let $\Sigma = \{I(Y_1|X_1), \ldots, I(Y_n|X_n)\}$ be a finite set of SCI statements and $I(Y|X)$ an SCI statement. We have to show that

$$I(Y|X) \in \Sigma^* \quad \text{if and only if} \quad I(Y|X) \in \Sigma^+_{\mathfrak{C}}.$$

Let $T := X \cup Y \cup V_\Sigma$. In order to prove the soundness of $\mathfrak{C}$ we assume that $I(Y|X) \in \Sigma^+_{\mathfrak{C}}$ holds. Let (V, dom, P) be a probability model that satisfies every element of Σ, and where $T \subseteq V$ holds. We must show that (V, dom, P) also satisfies $I(Y|X)$. According to Lemma 11 there is an inference γ of $I(Y|X)$ from Σ by $\mathfrak{C}$ such that $U \cup W \subseteq T \subseteq V$ holds for each SCI statement $I(W|U)$ that occurs in γ. Since each rule in $\mathfrak{C}$ is sound we can conclude (by induction) that each SCI statement occurring in γ is satisfied by (V, dom, P). In particular, (V, dom, P) satisfies $I(Y|X)$.

In order to prove the completeness of $\mathfrak{C}$ we assume that $I(Y|X) \notin \Sigma^+_{\mathfrak{C}}$. Let $V \subseteq \mathfrak{V}$ be a finite set of random variables such that T is a proper subset of V, i.e., $T \subset V$. Consequently, $V - XY$ is not a subset of T. Hence, by Lemma 10, $I(V - XY|X) \notin \Sigma^+_{\mathfrak{C}}$. Now from $I(Y|X) \notin \Sigma^+_{\mathfrak{C}}$ and from $I(V - XY|X) \notin \Sigma^+_{\mathfrak{C}}$ we conclude that $I(Y|X) \notin \Sigma^+_{\mathfrak{C}_V}$ by Theorem 7. Since $\mathfrak{C}_V$ is V-complete for the V-implication of SCI statements it follows that Σ does not V-imply $I(Y|X)$. Hence, Σ does not purely imply $I(Y|X)$ by Proposition 9. ▮

Example 13 *Recall Example 6 where $V = \{b, a, r, s\}$, and Σ consists of the two SCI statements $I(bs|a)$ and $I(b|a)$. The inference of $I(s|a)$ from Σ by $\mathfrak{C}_V$ in Example 6 is actually an inference by $\mathfrak{C}$. Hence, $I(s|a)$ is purely implied by Σ, as one would expect intuitively.*

5 PURE AND V-IMPLICATION

Instances $\Sigma \models \varphi$ of the pure implication problem can be characterized by the instance $\Sigma \models_V \varphi$ of the V-implication problem for any set V of incomplete random variables that *properly* contains $V_{\Sigma \cup \{\varphi\}}$.

Theorem 14 *Let $\Sigma \cup \{\varphi\}$ be a set of SCI statements. Then the following are equivalent:*
1. $\Sigma \models \varphi$
2. for some V such that $V_{\Sigma \cup \{\varphi\}} \subset V$, $\Sigma \models_V \varphi$
3. for all V such that $V_{\Sigma \cup \{\varphi\}} \subset V$, $\Sigma \models_V \varphi$

Proof It is clear that *3.* entails *2.* Let $\varphi = I(Y|X)$, and let V be any finite set of random variables such that $V_{\Sigma \cup \{\varphi\}} \subset V$. If *2.* holds, then Theorem 7 and Theorem 12 show that *1.* holds or $\Sigma \vdash_{\mathfrak{C}} I(V - XY|X)$ holds. However, Lemma 10 shows that the latter condition cannot hold as $V - XY$ contains some random variable that does not occur in V_Σ. Hence, *2.* entails *1.* If *1.* holds, then Theorem 7 and Theorem 12 show that *3.* holds as well. ▮

Example 15 $\Sigma = \{I(bs|a), I(b|a)\}$ *purely implies* $I(s|a)$ *as, for instance,* $\Sigma \models_V I(s|a)$ *for* $V = \{b, a, r, s\}$. $\Sigma' = \{I(bs|a)\}$ *does not purely imply* $I(r|a)$ *as for* $V = \{b, e, a, r, s\}$, Σ' *does not V-imply* $I(r|a)$ *as witnessed in the introduction.*

In the following we apply Theorem 14 to establish characterizations of pure implication in terms of logical formulae under Levesque's situations, database dependencies, and algorithmic solutions. For a set $\Sigma \cup \{\varphi\}$ of SCI statements we write $V_c = V_{\Sigma \cup \{\varphi\}} \cup \{v_0\}$ for some $v_0 \notin V_{\Sigma \cup \{\varphi\}}$, $\sigma_c = I(V_c - XY, Y|X)$ for $\sigma = I(Y|X) \in \Sigma \cup \{\varphi\}$ and $\Sigma_c = \{\sigma_c \mid \sigma \in \Sigma\}$. In particular, $\Sigma \models \varphi$ if and only if $\Sigma_c \models_{V_c} \varphi_c$.

6 LEVESQUE'S SITUATIONS

We recall the framework for situations from Levesque (1989), and exploit them to establish a logical characterization of the pure implication problem.

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., $L^\ell = L \cup \{\neg v' \mid v' \in L\}$. A *situation* of L is a total function $\omega : L^\ell \to \{\mathbb{F}, \mathbb{T}\}$ that does not map both a propositional variable $v' \in L$ and its negation $\neg v'$ to $\mathbb{F}$. That is, we must not have $\omega(v') = \mathbb{F} = \omega(\neg v')$ for any $v' \in L$.

A situation $\omega : L^{\ell} \to \{\mathbb{F}, \mathbb{T}\}$ of L can be lifted to a total function $\Omega : L^{*} \to \{\mathbb{F}, \mathbb{T}\}$. Assuming φ' is in Negation Normal Form, this lifting is defined by:

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

A situation ω is a *model* of a set Σ' of L-formulae if and only if $\Omega(\sigma') = \mathbb{T}$ holds for every $\sigma' \in \Sigma'$. We say that Σ' *implies* an L-formula φ', denoted by $\Sigma' \models_{L} \varphi'$, if and only if every situation that is a model of Σ' is also a model of φ'.

Equivalences. Let $\phi : V_c \to L_c$ denote a bijection between a set V_c of random variables and the set $L_c = \{v' \mid v \in V\}$ of propositional variables. We extend ϕ to a mapping Φ from the set of SCI statements over V_c to the set L_c^{*}. For an SCI statement $I(Y, Z \mid X)$ over V_c, let $\Phi(I(Y, Z \mid X))$ denote

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

Disjunctions over zero disjuncts are $\mathbb{F}$ and conjunctions over zero conjuncts are $\mathbb{T}$. We will denote $\Phi(\varphi_c) = \varphi'_c$ and $\Phi(\Sigma_c) = \{\Phi(\sigma_c) \mid \sigma \in \Sigma_c\} = \Sigma'_c$.

In our example, for $\varphi_c = I(bse, r \mid a)$ we have $\varphi'_c = \neg a' \vee (b' \wedge s' \wedge e') \vee r'$, and for $\Sigma_c = \{I(re, bs \mid a)\}$ we have $\Sigma'_c = \{\neg a' \vee (b' \wedge s') \vee (r' \wedge e')\}$.

It was shown in Link (2013a) that for any set $\Sigma_c \cup \{\varphi_c\}$ of SCI statements over V_c there is a probability model $\pi = (dom, P)$ over V_c that satisfies Σ_c and violates φ_c if and only if there is a situation ω_π over L_c that is a model of Σ'_c but not a model of φ'_c. For arbitrary probability models π it is not obvious how to define the situation ω_π. However, if Σ_c does not V_c-imply φ_c, then there is a special probability model $\pi = (dom, \{\mathbf{a}_1, \mathbf{a}_2\})$ over V_c that i) has two assignments $\mathbf{a}_1, \mathbf{a}_2$ of probability one half each, ii) satisfies all SCI statements in Σ_c and iii) violates φ_c. Given such π, let ω_π denote the following special situation of L_c, taken from Link (2013a):

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

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

From the results in Link (2013a) and Theorem 14 we obtain the following logical characterization of pure implication.

Theorem 16 *Let $\Sigma \cup \{\varphi\}$ be a finite set of SCI statements and $L_c = \{v' \mid v \in V_{\Sigma \cup \{\varphi\}} \cup \{v_0\}\}$. Then $\Sigma \models \varphi$ if and only if $\Sigma'_c \models_{L_c} \varphi'_c$.*

Proof Theorem 14 shows that $\Sigma \models \varphi$ if and only if $\Sigma_c \models_{V_c} \varphi_c$ for $V_c = V_{\Sigma \cup \{\varphi\}} \cup \{v_0\}$. By (Link, 2013a, Thm.6), $\Sigma_c \models_{V_c} \varphi_c$ if and only if $\Sigma'_c \models_{L_c} \varphi'_c$. ▮

Recall that $\Sigma = \{I(sb \mid a)\}$ does not purely imply $\varphi = I(s, br \mid a)$ as the special probability model π defined by

r	a	b	s	e	P
true	true	–	–	true	0.5
false	true	–	–	false	0.5

satisfies Σ_c, but violates φ_c. Any special situation where $\omega_\pi(b') = \mathbb{T} = \omega_\pi(s')$, $\omega_\pi(\neg a') = \omega_\pi(r') = \omega_\pi(e') = \mathbb{F}$ is a model of $\Sigma'_c = \{\neg a' \vee (b' \wedge s') \vee (r' \wedge e')\}$, but not a model of $\varphi'_c = \neg a' \vee (b' \wedge s' \wedge e') \vee r'$.

7 DATABASE DEPENDENCIES

Database dependencies enforce the semantics of application domains in database systems [Link (2001)]. Let $\mathfrak{A} = \{\hat{v}_1, \hat{v}_2, \ldots\}$ be an infinite set of distinct symbols, called attributes. A *relation schema* is a finite non-empty subset R of $\mathfrak{A}$. Each attribute $\hat{v} \in R$ has an infinite domain $dom(\hat{v})$. In order to encompass missing data values the domain of each attribute contains the null marker $-$. The intention of $-$ is to mean "no information" [Lien (1982)]. 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 $\hat{v} \in R$. For $X \subseteq R$ let $t(X)$ denote the restriction of t to X. A *relation* r over R is a finite set of tuples over R. For a tuple t over R and a set $X \subseteq R$, t is said to be X*-total*, if for all $\hat{v} \in X$, $t(\hat{v}) \neq -$. A relation over R is a *total relation*, if it is R-total. A *multivalued dependency* (MVD) over R is a statement $X \twoheadrightarrow Y$ where X and Y are disjoint subsets of R [Lien (1982)]. The MVD $X \twoheadrightarrow Y$ over R is satisfied by a relation r over R if and only if for all $t_1, t_2 \in r$ the following holds: if t_1 and t_2 are X-total and $t_1(X) = t_2(X)$, then there is some $t \in r$ such that $t(XY) = t_1(XY)$ and $t(X(R - XY)) = t_2(X(R - XY))$. Thus, the relation r satisfies $X \twoheadrightarrow Y$ when every X-total value determines the set of values on Y independently of the set of values on $R - Y$. For a set $\hat{\Sigma} \cup \{\hat{\varphi}\}$ of MVDs over R, $\hat{\Sigma}$ R-implies $\hat{\varphi}$, denoted by $\hat{\Sigma} \models_R \hat{\varphi}$, if and only if every relation over R that satisfies all elements in $\hat{\Sigma}$ also satisfies $\hat{\varphi}$.

For a set $\Sigma_c \cup \{\varphi_c\}$ of SCI statements over V_c one may associate the set $\hat{\Sigma_c} \cup \{\hat{\varphi_c}\}$ of MVDs over $R_c := \{\hat{v} \mid v \in V_c\}$, where $\hat{\sigma_c} = X \twoheadrightarrow Y$ for $\sigma_c = I(Y, Z|X)$ and $\hat{\Sigma_c} = \{\hat{\sigma_c} \mid \sigma \in \Sigma_c\}$.

Theorem 17 *Let $\Sigma \cup \{\varphi\}$ be a finite set of SCI statement. Then $\Sigma \models \varphi$ if and only if $\hat{\Sigma_c} \models_{R_c} \hat{\varphi_c}$.*

Proof Theorem 14 shows that $\Sigma \models \varphi$ if and only if $\Sigma_c \models_{V_c} \varphi_c$ for $V_c = V_{\Sigma \cup \{\varphi\}} \cup \{v_0\}$. By (Link, 2013a, Thm.8), $\Sigma_c \models_{V_c} \varphi_c$ if and only if $\hat{\Sigma_c} \models_{R_c} \hat{\varphi_c}$. ▮

8 ALGORITHM & COMPLEXITY

(Link, 2013a, Thm. 7) shows that $\Sigma_c \models_{V_c} \varphi_c$ for $\varphi_c = I(Z,Y|X)$ holds if and only if $\Sigma_c[X] \models_{V_c} \varphi_c$ holds classically, that is, when no domain contains the marker. Here, $\Sigma_c[X] := \{I(V,W|U) \mid I(V,W|U) \in \Sigma_c \wedge U \subseteq X\}$. The *independence basis IDepB*$_{\Sigma_c[X]}(X)$ consists of the minimal $Y \subseteq V_c - X$ such that $\Sigma_c[X] \models_{V_c} I(Z,Y|X)$. By (Link, 2013a, Thm. 8), $\Sigma \models \varphi$ if and only if $\widehat{\Sigma_c[X]} \models_{R_c} \hat{\varphi}_c$, that is, every total relation over R_c that satisfies $\widehat{\Sigma_c[X]}$ also satisfies $\hat{\varphi}_c$. Galil (1982) gave an efficient algorithmic solution to the latter problem.

Theorem 18 *Using the algorithm in Galil (1982), the pure implication problem $\Sigma \models I(Y|X)$ can be decided in time $\mathcal{O}(|\Sigma_c| + \min\{k_{\Sigma_c[X]}, \log \bar{p}_{\Sigma_c[X]}\} \times |\Sigma_c[X]|)$. Herein, $|\Sigma_c|$ denotes the total number of random variables in Σ_c, $k_{\Sigma_c[X]}$ denotes the cardinality of $\Sigma_c[X]$, and $\bar{p}_{\Sigma_c[X]}$ denotes the number of sets in IDepB$_{\Sigma_c[X]}(X)$ that have non-empty intersection with Y.* $\blacksquare$

9 RELATED WORK

Dawid (1979) first investigated 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 (1993) have systematically investigated the implication problem for fragments of CI statements over different probability models. In particular, they have established an axiomatization of SCI statements by a finite set of Horn rules. Studený (1992) showed that no axiomatization by a finite set of Horn rules exists for general CI statements. Niepert, Van Gucht, and Gyssens (2010) established an axiomatization for stable CI statements, which subsume SCI statements, and showed that their associated implication problem is coNP-complete. Independently, database theory has investigated the concept of embedded multivalued dependencies (MVDs) whose implication problem is undecidable [Herrmann (1995)] and not axiomatizable by a finite set of Horn rules [Stott Parker Jr. and Parsaye-Ghomi (1980)]. Studený (1992) also showed that the implication problem of embedded MVDs and that of CI statements do not coincide. In contrast, the implication problems of MVDs, SCI statements and some fragement of Boolean propositional logic all coincide [Geiger and Pearl (1993); Sagiv et al. (1981); Wong, Butz, and Wu (2000)]. These findings have been established for the notion of implication over fixed sets of variables and the idealized case where all data values are known. Biskup, Hartmann, and Link (2012) differentiated between V-implication and pure implication for SCI statements with complete random variables only, applying ideas from database theory in Biskup (1980) and Link (2012). In the case of missing data, equivalences between implication problems for MVDs with null markers, SCI statements with incomplete random variables, and a fragment of propositional logic under Levesque's situations were established recently in Link (2013a) and Hartmann and Link (2012). However, the notion of pure implication for conditional independence statements has not been studied yet in the context of missing data.

10 CONCLUSION

Recently, probabilistic conditional independence statements were studied in the presence of incomplete random variables, which admit missing data values. The associated implication problem for saturated CI statements was characterized axiomatically by a finite set $\mathfrak{U}_V$ of Horn rules, logically by a propositional fragment under interpretations by Levesque's situations, and algorithmically by an equivalence to database dependencies. In this paper it was shown that there is a difference between SCI statements V-implied jointly by a given set of SCI statements and a fixed set V of incomplete random variables, and those purely implied by a given set of SCI statements alone. It was shown that $\mathfrak{U}_V$ cannot separate V-implied from purely implied SCI statements. An axiomatization $\mathfrak{C}_V$ was then established that can infer any purely implied SCI statement without applications of the V-symmetry rule $\mathcal{S}_V$, and infer any V-implied SCI statement with a single application of $\mathcal{S}_V$ in the very last step of the inference only. The system $\mathfrak{C}$ that results from $\mathfrak{C}_V$ by removing $\mathcal{S}_V$ was proven to from a finite axiomatization for the stronger notion of pure implication. The pure implication problem $\Sigma \models \varphi$ was characterized by the V-implication problem $\Sigma \models_V \varphi$ for sets V that properly contain the random variables that occur $\Sigma \cup \{\varphi\}$. This result enabled us to characterize pure implication logically and algorithmically as well. Our results clarify the role of the V-symmetry rule $\mathcal{S}_V$ as a pure means to infer V-implied SCI statements. The notion of pure implication is appealing when the existence of random variables is uncertain, for example, when independence statements are integrated from different sources, when random variables are unknown or when they shall remain hidden. It is future work to extend the findings of this paper to the general case where an arbitrary finite set S of complete random variables can be specified, thereby, covering the current setting by the case where $S = \emptyset$ and the classical setting by the case where every random variable is complete. This would extend the work in Link (2013b) where the notion of pure implication was not considered.

Acknowledgements

This research is supported by the Marsden fund council from Government funding, administered by the Royal Society of New Zealand.

References

Biskup, J.; Hartmann, S.; and Link, S. 2012. Probabilistic conditional independence under schema certainty and uncertainty. In *SUM*, 365–378.

Biskup, J. 1980. Inferences of multivalued dependencies in fixed and undetermined universes. *Theor. Comput. Sci.* 10(1):93–106.

Chickering, D. M., and Heckerman, D. 1997. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning* 29(2-3):181–212.

Darwiche, A. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press.

Dawid, A. P. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1):1–31.

Dempster, A.; Laird, N. M.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:139.

Friedman, N. 1997. Learning belief networks in the presence of missing values and hidden variables. In *ICML*, 125–133.

Galil, Z. 1982. An almost linear-time algorithm for computing a dependency basis in a relational database. *J. ACM* 29(1):96–102.

Geiger, D., and Pearl, J. 1993. Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics* 21(4):2001–2021.

Halpern, J. 2005. *Reasoning about uncertainty*. MIT Press.

Hartmann, S., and Link, S. 2012. The implication problem of data dependencies over SQL table definitions: axiomatic, algorithmic and logical characterizations. *ACM Trans. Database Syst.* 37(2):Article 13.

Herrmann, C. 1995. On the undecidability of implications between embedded multivalued database dependencies. *Inf. Comput.* 122(2):221–235.

Lauritzen, S. 1995. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* 19:191–201.

Levesque, H. 1989. A knowledge-level account of abduction. In *IJCAI*, 1061–1067.

Lien, E. 1982. On the equivalence of database models. *J. ACM* 29(2):333–362.

Link, S. 2001. Consistency enforcement in databases. In *Semantics in Databases*, 139–159.

Link, S. 2012. Characterizations of multivalued dependency implication over undetermined universes. *J. Comput. Syst. Sci.* 78(4):1026–1044.

Link, S. 2013a. Reasoning about saturated conditional independence under uncertainty. In *AAAI*.

Link, S. 2013b. Sound approximate reasoning about saturated conditional probabilistic independence under controlled uncertainty. *J. Applied Logic* 11(3):309–327.

Marlin, B. M.; Zemel, R. S.; Roweis, S. T.; and Slaney, M. 2011. Recommender systems, missing data and statistical model estimation. In *IJCAI*, 2686–2691.

Niepert, M.; Gyssens, M.; Sayrafi, B.; and Gucht, D. V. 2013. On the conditional independence implication problem: A lattice-theoretic approach. *Artif. Intell.* 202:29–51.

Niepert, M.; Van Gucht, D.; and Gyssens, M. 2010. Logical and algorithmic properties of stable conditional independence. *Int. J. Approx. Reasoning* 51(5):531–543.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Franciso, U.S.A.: Morgan Kaufmann.

Saar-Tsechansky, M., and Provost, F. J. 2007. Handling missing values when applying classification models. *Journal of Machine Learning Research* 8:1623–1657.

Sagiv, Y.; Delobel, C.; Parker Jr., D. S.; and Fagin, R. 1981. An equivalence between relational database dependencies and a fragment of propositional logic. *J. ACM* 28(3):435–453.

Singh, M. 1997. Learning bayesian networks from incomplete data. In *AAAI/IAAI*, 534–539.

Stott Parker Jr., D., and Parsaye-Ghomi, K. 1980. Inferences involving embedded multivalued dependencies and transitive dependencies. In *SIGMOD*, 52–57.

Studený, M. 1992. Conditional independence relations have no finite complete characterization. In *11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 377–396.

Wong, S.; Butz, C.; and Wu, D. 2000. On the implication problem for probabilistic conditional independency. *Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans* 30(6):785–805.

Zhu, X.; Zhang, S.; Zhang, J.; and Zhang, C. 2007. Cost-sensitive imputing missing values with ordering. In *AAAI*, 1922–1923.