Artificial Intelligence 202 (2013) 29–51

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

On the conditional independence implication problem: A lattice-theoretic approach

Mathias Niepert [a,*], Marc Gyssens [b], Bassem Sayrafi [c], Dirk Van Gucht [d]

[a] Computer Science & Engineering, University of Washington, 185 Stevens Way, Seattle, WA 98195-2350, USA
[b] School for Information Technology, Hasselt University and Transnational University of Limburg, Martelarenlaan 42, B-3500 Hasselt, Belgium
[c] Computer Science Department, Birzeit University, PO Box 14, Birzeit, West Bank, Palestine
[d] Computer Science Department, Indiana University, Lindley Hall 215, 150 S. Woodlawn Ave., Bloomington, IN 47405-7104, USA

ARTICLE INFO

Article history:
Received 4 July 2011
Received in revised form 4 June 2013
Accepted 19 June 2013
Available online 28 June 2013

Keywords:
Conditional independence
Probability and lattice theory

ABSTRACT

Conditional independence is a crucial notion in the development of probabilistic systems which are successfully employed in areas such as computer vision, computational biology, and natural language processing. We introduce a lattice-theoretic framework that permits the study of the conditional independence (CI) implication problem relative to the class of discrete probability measures. Semi-lattices are associated with CI statements and a finite, sound and complete inference system relative to semi-lattice inclusions is presented. This system is shown to be (1) sound and complete for inferring general from saturated CI statements and (2) complete for inferring general from general CI statements. We also show that the general probabilistic CI implication problem can be reduced to that for elementary CI statements. The completeness of the inference system together with its lattice-theoretic characterization yields a criterion we can use to falsify instances of the probabilistic CI implication problem as well as several heuristics that approximate this falsification criterion in polynomial time. We also propose a validation criterion based on representing constraints and sets of constraints as sparse 0–1 vectors which encode their semi-lattices. The validation algorithm works by finding solutions to a linear programming problem involving these vectors and matrices. We provide experimental results for this algorithm and show that it is more efficient than related approaches.

1. Introduction

Conditional independence is an important concept in artificial intelligence and machine learning. It plays a fundamental role in working with probabilistic systems successfully employed in areas such as computer vision, speech recognition, computational biology, and robotics. Numerous real-world systems can be modeled by a probability distribution over a set of random variables. Unfortunately, reasoning over the full joint probability distribution is intractable for all but the smallest number of cases. It is the very notion of conditional independence that facilitates the *decomposition* of joint probability distributions into smaller parts which are then processed in sophisticated ways to compute a-posteriori probabilities. Bayesian and Markov networks are among the most commonly used probabilistic graphical models leveraging conditional independencies to answer probabilistic queries and to learn probabilistic parameters more efficiently [1]. A deeper theoretical

* Corresponding author.
E-mail addresses: mniepert@cs.washington.edu (M. Niepert), marc.gyssens@uhasselt.be (M. Gyssens), bassem.sayrafi@gmail.com (B. Sayrafi), vgucht@cs.indiana.edu (D. Van Gucht).

0004-3702/$ – see front matter
http://dx.doi.org/10.1016/j.artint.2013.06.005

investigation of the mathematical and algorithmic properties of conditional independence is therefore central to the understanding of probabilistic models [2,3].

A deep theoretical understanding of conditional independence and the way it is leveraged in probabilistic graphical models, however, also allows us to understand the shortcomings of said models. Indeed, not every joint probability distribution can be decomposed according to a graphical structure without loss of information [4,5]. Finding ways of assessing the suitability of graphical models for the representation of a given distribution is therefore crucial. In particular, Studený [6] has brought this issue to the forefront, leading to an impressive body of work on algebraic representations of conditional independence structures, providing deep links to algebraic geometry [6,7], supermodular functions on sets, and novel algorithms for reasoning with conditional independencies [8,9].

More motivation for this research is provided by the problem of knowledge elicitation in the field of reasoning under uncertainty [10,11]. Consider the problem of eliciting knowledge from several domain experts to model a probabilistic system. The resulting incomplete expert feedback might be a combination of some specific subjective probabilities, (conditional) independency and dependency information for the random variables under consideration, and conditional probabilities. Statistical tests on different heterogeneous data sets may provide additional sources of evidence. Each piece of information can be interpreted as a *constraint* on the joint probability distribution to be modeled, and finding a suitable model as a *constraint satisfaction problem* (CSP), and the approach to harness CSP solvers for instances of this and related problems is well-known [12,13]. However, (conditional) independence and dependence statements pose a special problem, because they often introduce *non-linear* constraints resulting in unfeasible CSP instances. Therefore, a remaining challenge is to test for *consistency* of the (conditional) independence and dependence information collected from different sources, which requires an algorithm deciding the implication problem for CI statements [14].

A central notion in the realm of reasoning about conditional independence is, therefore, the *probabilistic conditional independence implication problem*, that is, to decide whether a set of CI statements implies a single CI statement relative to some class of discrete probability measures. While it remains open whether the implication problem for the class of all discrete probability measures is decidable, it is known that there exists no finite, sound and complete inference system [15]. However, there exist finite *sound* inference systems that have attracted special interest. The most prominent one is the *semi-graphoid* axiom system (Pearl [3]), which we refer to as System $\mathscr{G}$ in the present paper. One of the main contributions of this work is to extend the semi-graphoids to a finite inference system which we refer to as System $\mathscr{A}$ which, although not sound, is *complete* for the general probabilistic implication problem. In the way that the semi-graphoid inference rules provide a *lower bound* on what can be inferred, System $\mathscr{A}$ provides an *upper bound*. We demonstrate that, in the general case where the number of variables is not fixed and where no finite axiomatization exists, considering both lower and upper bounds provides deep insights into the implication problem and allows us to develop a novel algorithm for both validating and rejecting implication problem instances.

The techniques we use to obtain these results are made possible through the introduction of a *lattice-theoretic framework*. We associate semi-lattices of sets of variables with CI statements. Derivability of a single CI statement from a set of CI statements in System $\mathscr{A}$ is then characterized by the inclusion of the semi-lattice of the former in the union of the semi-lattices of the latter. We also use this framework to show that derivability in System $\mathscr{A}$ in the context of arbitrary CI statements can be reduced to derivability in the context of *elementary* CI statements, that is, CI statements that express independence between two single variables given a third set of variables. This result has important ramifications from a practical point of view because the use of elementary CI statements allows for a canonical representation of CI statements. We then introduce the *additive implication problem* for CI statements relative to certain classes of real-valued functions and specify properties of these classes that guarantee either soundness or completeness of $\mathscr{A}$. Through the concept of *multi-information functions* induced by probability measures [6], we finally link the additive implication problem for this class of functions to the *multiplication-based* probabilistic CI implication problem. This allows us to show, for instance, that System $\mathscr{A}$ is *sound and complete* for the inference of *arbitrary* CI statements from sets of *saturated* CI statements relative to both the class of all binary probability measures and the class of all discrete probability measures. Saturated CI statements by definition involve all variables under consideration.

The combination of the lattice-inclusion techniques and the completeness of System $\mathscr{A}$ for the general probabilistic conditional independence problem allows us to derive criteria that can be used to *falsify* or *validate* instances of this implication problem. We introduce an *approximate logical implication* algorithm which combines these falsification and validation criteria. The validation algorithm is based on our results regarding the reduction of derivability in System $\mathscr{A}$ for general CI statements to derivability for elementary CI statements. It represents a set of such elementary CI statements as a sparse 0–1 matrix, and validates instances of the implication problem by solving linear programs with this matrix as constraint matrix. Thus, by only requiring the algorithm to decide the majority of the probabilistic conditional independence implication problems, we can leverage linear constraint solvers for our purposes. We present an experimental evaluation in which we investigate the fraction of instances of the implication problem that can be decided with the novel approach.

We report the results of extensive experiments designed to assess the viability of our approach. We relate the experimental results for our falsification criteria to those obtained earlier from a *racing algorithm* introduced by Bouckaert and Studený [8]. The linear programming techniques used in the validation criteria were subsequently also used by Bouckaert et al. [9]. We also compare the experimental results for our validation criteria with their results. The results of our experiments show that our approximation algorithm works very effectively and compares favorably to the related work, often outperforming it by several orders of magnitude.

2. Related work

Probabilistic conditional independence is an important notion in several disciplines [2,3], forming the theoretical basis for probabilistic graphical models in particular and efficient probabilistic inference in general. Numerous lines of research have been devoted to the study of its mathematical, logical, and algorithmic properties. The first axiomatic approach to conditional independence was introduced by Dawid [2]. A closely related inference system for probabilistic CI was developed by Pearl and Paz [16]. In both lines of work, a three-place relation was characterized using similar axiom systems. While Dawid termed the relation a separoid (cf. Dawid [17]), Pearl and Paz termed it a graphoid. These relations were shown to surface in the context of probabilistic conditional independence and other notions in statistics and artificial intelligence. Shafer [18] provides a comprehensive survey of some of the advances in the use and investigation of probabilistic conditional independence up until the year 1996, published in a special issue of *Annals of Mathematics and Artificial Intelligence* [19–23]. A line of work that has stimulated considerable interest in probabilistic conditional independence is Pearl's book about probabilistic expert systems [3]. A more recent overview of the principles and practice of graphical models can be found in, e.g., Koller and Friedman [1].

A considerable amount of related work exists in the area of uncertainty in artificial intelligence. Studený's work on structural representations of conditional independence [6], for instance, is highly related to the present work. In fact, there is a close connection between the lattice-theoretic representation and Studený's theory of imsets [6], an algebraic framework for the study of supermodular functions on sets. Building upon Studený's work, Hemmecke et al. [7] used, among other tools, integer programming to solve some open problems concerning imsets. Studený discussed the use of the maximization problem over the class of l-standardized supermodular functions which can be posed as a linear program [24]. A different but related linear programming formulation for a validation algorithm for the CI implication problem was introduced by Niepert et al. [25]. Independently, Bouckaert et al. [9] leveraged the theory of imsets [6] to assemble linear programs to verify various instances of the CI implication problem. Their approach is similar to ours, except that they do not use the several orders of magnitude more compact and, therefore, more efficient representation. We will compare the two linear programming formulations in the experimental section (Section 9).

Similarly influential to the present work is the work of Geiger and Pearl [14] and Malvestuto [26] about the algorithmic and logical properties of conditional independence. In their work, probabilistic conditional independence is approached from a purely logical point of view and syntactic inference systems and their properties with respect to various classes of probability measures are investigated. The authors also discuss possible algorithms for the probabilistic conditional independence implication problem and the computational complexity of some of the problems and also pose several open problems.

This article is a substantial extension, in terms of both theoretical and experimental results, of previous conference papers by the same authors [25,27]. With this article we extend the theory to fully cover the case of elementary CI statements (Section 5) which serves as the basis for most of the results in Section 9. In addition to additional examples and a refinement of the theoretical results with respect to the additive implication problem (Section 6) and the significant generalization of the validation algorithm (Section 7), the experimental section (Section 9) was expanded with several additional experiments not included in previous publications. The construction of the constraint matrix of the validation algorithm is also a novel contribution increasing the efficiency of the validation algorithm. The lattice-theoretic framework that is developed is a continuation of previous work by three of the present authors and Purdom [28–32]. The theory is applicable in several different areas of computer science. Sayrafi's work, for instance, focused primarily on the class of frequency functions over finite relational databases [31]. The derived inference rules were used in frequent itemset mining algorithms as heuristics to prune the search space [33].

Matúš [34] and de Waal and van der Gaag [35,36] introduce stable independence as a notion to more compactly represent information about conditional independence.[1] To this end, conditional independence statements are partitioned in a stable and a non-stable part allowing for more efficient storage and retrieval of probabilistic conditional independencies. Stable independence was also studied in earlier work of some of the present authors [5]. Though not elaborated upon in this paper, the theory we develop is also applicable to stable sets of CI statements.

3. Preliminaries

We define conditional independence (CI) statements as a syntactic notion. Finite inference systems are frequently used to decide logical entailment of such statements at the syntactic level. In the context of CI statements, we consider System $\mathscr{G}$, the semi-graphoid axioms of Pearl [3] and System $\mathscr{S}$ of Malvestuto [26] and Geiger and Pearl [14]. We add to this a novel inference system, referred to as System $\mathscr{A}$, with which we compare the existing ones on a purely syntactic level.

With regard to set notation, we often write AB for the union $A \cup B$, ab for the set $\{a, b\}$, and a for the singleton set $\{a\}$ if no confusion is possible. Throughout the paper, S denotes a finite universe containing all sets under consideration. For a set A, we write $\overline{A}$ for $S - A$, the complement of A with respect to S.

We begin by defining *conditional independence* (CI) *statements*.

[1] Two sets of variables are stably independent if they are independent relative to a set of variables C and remain independent relative to each superset of C.

$I(A,\emptyset\|C)$	**Triviality**
$I(A,B\|C) \rightarrow I(B,A\|C)$	**Symmetry**
$I(A,BD\|C) \rightarrow I(A,B\|C)$	**Decomposition**
$I(A,B\|CD)$ & $I(A,D\|C) \rightarrow I(A,BD\|C)$	**Contraction**
$I(A,BD\|C) \rightarrow I(A,B\|CD)$	**Weak union**

Fig. 1. The inference rules of the semi-graphoid System $\mathscr{G}$.

$I(A,\emptyset\|C)$	**Triviality**
$I(A,B\|C) \rightarrow I(B,A\|C)$	**Symmetry**
$I(A,B\|CD)$ & $I(AB,D\|C) \rightarrow I(A,BD\|C)$	**Weak Contraction**
$I(A,BD\|C) \rightarrow I(A,B\|CD)$	**Weak union**

Fig. 2. The inference rules of System $\mathscr{S}$ for *saturated* CI statements.

$I(A,\emptyset\|C)$	**Triviality**
$I(A,B\|C) \rightarrow I(B,A\|C)$	**Symmetry**
$I(A,BD\|C) \rightarrow I(A,B\|C)$	**Decomposition**
$I(A,B\|CD)$ & $I(A,D\|C) \rightarrow I(A,BD\|C)$	**Contraction**
$I(A,B\|C) \rightarrow I(A,B\|CD)$	**Strong union**
$I(A,B\|C)$ & $I(D,E\|AC)$ & $I(D,E\|BC) \rightarrow I(D,E\|C)$	**Strong contraction**

Fig. 3. The inference rules of System $\mathscr{A}$.

Definition 1. A *conditional independence (CI) statement* is an expression $I(A, B|C)$ where A, B, and C are pairwise disjoint subsets of S. If $ABC = S$, $I(A, B|C)$ is *saturated*. If $A = \emptyset$, or $B = \emptyset$, or both, $I(A, B|C)$ is *trivial*.

In this paper, we consider three inference systems for CI statements. If $\mathscr{I}$ is such an inference system, $\mathscr{C}$ a set of CI statements, and c a single CI statement, we denote by $\mathscr{C} \vdash_{\mathscr{I}} c$ that c is derivable from $\mathscr{C}$ under the inference rules of System $\mathscr{I}$. The *closure* $\mathscr{C}^+_{\mathscr{I}}$ of $\mathscr{C}$ under $\mathscr{I}$ is the set $\{c \mid \mathscr{C} \vdash_{\mathscr{I}} c\}$.

We first consider the well-known semi-graphoid axiom system [3], denoted here as System $\mathscr{G}$ and shown in Fig. 1, and System $\mathscr{S}$ [14,26], shown in Fig. 2. In contrast to System $\mathscr{G}$, the rules of System $\mathscr{S}$ are only applicable to saturated CI statements and yield saturated CI statements only. In all rules, the sets are assumed pairwise disjoint. Clearly, we have the following.

Proposition 2. *System $\mathscr{S}$ can be inferred from System $\mathscr{G}$; the* weak contraction *rule can be inferred from the* symmetry, decomposition, *and* contraction *inference rules.*

Finally, we consider the set of inference rules in Fig. 3, denoted as System $\mathscr{A}$. Besides four rules of the semi-graphoid axiom system, it contains the *strong union* and *strong contraction* rules. Clearly, $\mathscr{G}$ is subsumed by System $\mathscr{A}$.

Proposition 3. *System $\mathscr{G}$ can be inferred from System $\mathscr{A}$; the* weak union *rule can be inferred from the* decomposition *and* strong union *rules.*

From System $\mathscr{A}$, many other rules can be inferred, such as the *composition* rule shown in Fig. 4. As System $\mathscr{A}$ plays a pivotal role in this work, we use the following abbreviations. If $\mathscr{C}$ is a set of CI statements and c is a single CI statement, then $\mathscr{C} \vdash c$ is short for $\mathscr{C} \vdash_{\mathscr{A}} c$; similarly, $\mathscr{C}^+$ is short for $\mathscr{C}^+_{\mathscr{A}}$.

We emphasize that in this and the next section we take a purely syntactic viewpoint towards CI statements, inference rules, and inference systems. In particular, we do not imply that the rules exhibited in Figs. 1, 2, 3, and 4 are necessarily

$I(A,B\|C)$ & $I(A,D\|C) \to I(A,BD\|C)$	**Composition**

Fig. 4. The *composition* rule.

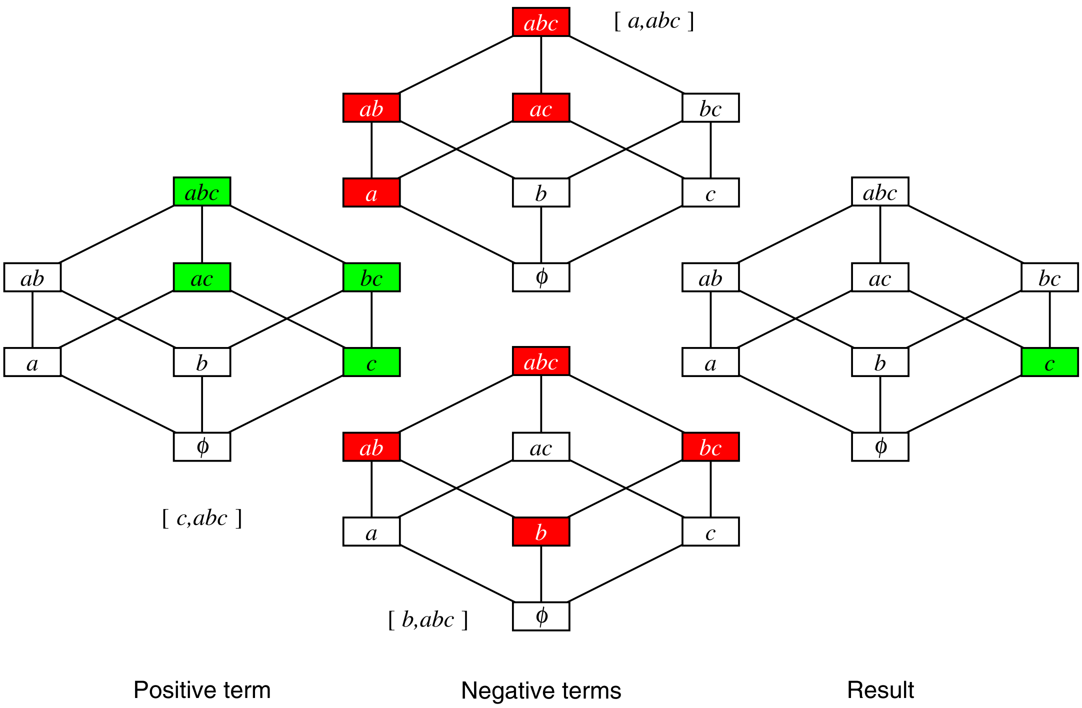

Fig. 5. For $S = \{a, b, c\}$, $\mathscr{L}(a, b|c) = [c, abc] - ([a, abc] \cup [b, abc]) = \{c\}$.

sound and/or complete. Notice, for example, that *strong union* is *not* a sound rule for the inference of CI statements relative to the class of discrete probability measures, as is illustrated by Examples 35 and 38.

4. Lattice-theoretic framework

We first introduce the lattice-theoretic framework which is at the basis of the theory developed in this work. We associate CI statements with so-called *meet semi-lattices*, i.e., partially ordered sets in which each pair of elements has an infimum (a meet) [37]. If the semi-lattice is a class of sets, the infimum of a pair of its elements is their set intersection. In the following, the term "semi-lattice" is used for a meet semi-lattice on a class of sets. Subsequently, we characterize derivability of CI statements in terms of their associated semi-lattices.

Given subsets A and B of S, we write $[A, B]$ for $\{U \mid A \subseteq U \subseteq B\}$. We now associate semi-lattices with conditional independence statements as follows.

Definition 4. Let $I(A, B|C)$ be a CI statement. The *semi-lattice* of $I(A, B|C)$, denoted $\mathscr{L}(A, B|C)$, equals $[C, S] - ([A, S] \cup [B, S]) = \{U \subseteq S \mid C \subseteq U, A - U \neq \emptyset, B - U \neq \emptyset\}$.

Notice that $\mathscr{L}(A, B|C)$ is indeed closed under intersection.

Example 5. First, let $S = \{a, b, c\}$. Then, $\mathscr{L}(a, b|c) = \{c\}$, as illustrated in Fig. 5. Next, let $S = \{a, b, c, d\}$. Then,

$$\mathscr{L}(bc, d|a) = [a, abcd] - \big([bc, abcd] \cup [d, abcd]\big) = \{a, ab, ac\};$$

$$\mathscr{L}(bc, \emptyset|a) = [a, abcd] - \big([bc, abcd] \cup [\emptyset, abcd]\big) = \emptyset; \quad \text{and}$$

$$\mathscr{L}(ab, cd|\emptyset) = [\emptyset, abcd] - \big([ab, abcd] \cup [cd, abcd]\big) = \{\emptyset, a, b, c, d, ac, ad, bc, bd\}.$$

For a CI statement c, $\mathscr{L}(c)$ denotes the semi-lattice of c, and, for a set of CI statements $\mathscr{C}$, $\mathscr{L}(\mathscr{C}) = \bigcup_{c' \in \mathscr{C}} \mathscr{L}(c')$.[2]

Definition 6. Let $I(A, B|C)$ be a CI statement. The set of witness sets of $I(A, B|C)$, denoted $\mathscr{W}(A, B|C)$, equals $\{\{a, b\} \mid a \in A \;\&\; b \in B\}$.

Note that, if $I(A, B|C)$ is trivial, then $\mathscr{W}(A, B|C) = \emptyset$. Using the previously defined notion of *witness sets* of a CI statement, we can rewrite the associated semi-lattice as a union of lattices. This is crucial as it opens the door to proofs by induction on the lattice-elements. The following result is a straightforward consequence of Definition 4.

Proposition 7. *For a CI statement* $c = I(A, B|C)$, $\mathscr{L}(c) = \bigcup_{W \in \mathscr{W}(c)} [C, \overline{W}]$.

Example 8. Let $S = \{a, b, c, d\}$ and consider the CI statement $I(bc, d|a)$. Then, $\mathscr{L}(bc, d|a) = [a, S] - ([bc, S] \cup [d, S]) = \{a, ab, ac\}$. Alternatively, we may apply Proposition 7 and deduce from $\mathscr{W}(bc, d|a) = \{bd, cd\}$ that $\mathscr{L}(bc, d|a) = [a, \overline{bd}] \cup [a, \overline{cd}] = [a, ac] \cup [a, ab] = \{a, ab, ac\}$.

We now show that, given a set of CI statements $\mathscr{C}$ and a single CI statement c, $\mathscr{C} \vdash c$ if and only if $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$. This characterization of derivability of CI statements under System $\mathscr{A}$ in terms of their associated semi-lattices is at the basis of our work on the CI implication problem. The implication from left to right can be shown by a straightforward structural induction. Therefore, we focus on the implication from right to left, for which we need the notion of *witness decomposition*.

Definition 9. The *witness decomposition* of the CI statement $I(A, B|C)$, denoted $\mathrm{wdec}(A, B|C)$, equals $\{I(a, b|C) \mid a \in A \;\&\; b \in B\}$. For a set of CI statements $\mathscr{C}$, $\mathrm{wdec}(\mathscr{C}) = \bigcup_{c' \in \mathscr{C}} \mathrm{wdec}(c')$.

We now justify the name "decomposition" by showing that, for any CI statement c, we have that (a) the closure of the witness decomposition under $\mathscr{A}$ is equal to the closure of c under $\mathscr{A}$, and (b) the union of semi-lattices associated with the elements of the witness decomposition is equal to the semi-lattice associated with c.

Proposition 10. *Let c be a CI statement. Then,* (1) $\{c\}^+ = \mathrm{wdec}(c)^+$; *and* (2) $\mathscr{L}(c) = \bigcup_{c' \in \mathrm{wdec}(c)} \mathscr{L}(c')$.

Proof. Let $c = I(A, B|C)$. For the first statement, the inclusion from right to left can be derived straightforwardly using the *decomposition* and *symmetry* rules. To see the reverse inclusion, let $a \in A$. For all $b \in B$, $I(a, b|C) \in \mathrm{wdec}(c)$. By repeatedly applying the *composition* rule, we infer $I(a, B|C)$, and, by *symmetry*, $I(B, a|C)$. By another repeated application of the *composition* rule, we infer $I(B, A|C)$, and, by *symmetry*, $I(A, B|C)$. Hence, $\{c\}^+ \subseteq \mathrm{wdec}(c)^+$. For the second statement, let $I(a, b|C) \in \mathrm{wdec}(c)$, and let $W = \{a, b\}$. Clearly, $\mathscr{W}(a, b|C) = \{W\}$ and, by Proposition 7, $\mathscr{L}(a, b|C) = [C, \overline{W}]$. The statement now follows from Proposition 7. $\square$

We are now in the position to prove the fundamental equivalence between derivability under System $\mathscr{A}$ and the inclusion relationship between the lattice decompositions of the consequent and the antecedents.

Theorem 11. *Let $\mathscr{C}$ be a set of CI statements, and let c be a single CI statement. Then $\mathscr{C} \vdash c$ if and only if $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$.*

Proof. As said before, we focus on the "if". Let $I(a, b|C) \in \mathrm{wdec}(c)$, and let $W = \{a, b\}$. Since $\mathscr{L}(c) = \bigcup_{c' \in \mathrm{wdec}(c)} \mathscr{L}(c')$, $\mathscr{L}(a, b|C) \subseteq \mathscr{L}(\mathscr{C})$ (1). By Proposition 10, it suffices to show that $I(a, b|C) \in \mathrm{wdec}(\mathscr{C})^+$. Thereto, we prove the stronger statement $\forall V \in [C, \overline{W}]$: $I(a, b|V) \in \mathrm{wdec}(\mathscr{C})^+$ by downward induction on $[C, \overline{W}]$. For the base case, we must show that $I(a, b|\overline{W}) \in \mathrm{wdec}(\mathscr{C})^+$. By statement (1), $\overline{W} \in \mathscr{L}(\mathscr{C})$. By Proposition 10, there exist a', b' such that $I(a', b'|C') \in \mathrm{wdec}(\mathscr{C})$ and such that $\overline{W} \in \mathscr{L}(a', b'|C')$. Hence, $\{a', b'\} = \{a, b\}$ and $C' \subseteq \overline{W}$, and we can derive $I(a, b|\overline{W})$ through *strong union* and *symmetry*. For the inductive step, let $C \subseteq V \subset \overline{W}$. The inductive hypothesis states that, for all V' with $V \subset V' \subseteq \overline{W}$, $I(a, b|V') \in \mathrm{wdec}(\mathscr{C})^+$ (2). As in the base case, we can show that there exist a' and b' such that $I(a', b'|V) \in \mathrm{wdec}(\mathscr{C})^+$ using *strong union*. Let $W' = \{a', b'\}$. If $W' = W$ then $\{a', b'\} = \{a, b\}$ and we can use *symmetry* to infer $I(a, b|V)$. Otherwise, we distinguish two cases:

(i) *Exactly one of a' and b' is in W.* Assume $a' = a \in W$. Using *contraction* on $I(a, b'|V)$ and on $I(a, b|Vb')$ (in $\mathrm{wdec}(\mathscr{C})^+$ by (2)), we derive $I(a, b'b|V)$, and, using *decomposition*, we derive $I(a, b|V)$.

(ii) *Neither a' nor b' is in W.* Using *strong contraction* on $I(a', b'|V)$ and on $I(a, b|Va')$ and $I(a, b|Vb')$ (both in $\mathrm{wdec}(\mathscr{C})^+$ by (2)), we derive $I(a, b|V)$. $\square$

We conclude this section with an illustrative example.

[2] Observe that the union of meet semi-lattices is not always a meet semi-lattice.

$I(a,b\|C) \rightarrow I(b,a\|C)$	**Symmetry**
$I(a,b\|Cd)$ & $I(a,d\|C) \rightarrow I(a,b\|C)$	**Elem. contraction**
$I(a,b\|C) \rightarrow I(a,b\|CD)$	**Strong union**
$I(a,b\|C)$ & $I(d,e\|aC)$ & $I(d,e\|bC) \rightarrow I(d,e\|C)$	**Strong contraction**

Fig. 6. The inference rules of System $\mathscr{E}$.

Example 12. Let $S = \{a, b, d, e\}$, let $\mathscr{C} = \{I(a, b|\emptyset), I(d, e|a), I(d, e|b)\}$, and let $c = I(d, e|\emptyset)$. Each of these four CI statements has only one witness set, which is $\{a, b\}$ for the first and $\{d, e\}$ for the other three. On the one hand, by Proposition 7, we have that $\mathscr{L}(\mathscr{C}) = [\emptyset, de] \cup [a, ab] \cup [b, ab] = \{\emptyset, a, b, d, e, ab, de\}$ and $\mathscr{L}(c) = [\emptyset, ab] = \{\emptyset, a, b, ab\}$. Therefore, $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$. On the other hand, we can derive $I(d, e|\emptyset)$ from $I(a, b|\emptyset)$, $I(d, e|a)$, and $I(d, e|b)$ using the *strong contraction* rule. Therefore, $\mathscr{C} \vdash c$, consistent with Theorem 11.

5. Elementary CI statements

A *witness decomposition* of a CI statement (Definition 9) is equivalent to the original statement for derivations under System $\mathscr{A}$ (Proposition 10). Therefore, the CI statements involved in witness decompositions deserve further study.

Definition 13. A CI statement is *elementary* if it is of the form $I(a, b|C)$, with a and b distinct elements of S and C a subset of $S - \{a, b\}$.

Elementary CI statements play an important role in Section 8 and following, where we apply the developed theory to obtain falsification and validation algorithms for the probabilistic implication problem. Here, we study some of their properties.

Definition 14. Let $c = I(A, B|C)$ be an arbitrary CI statement. We define $\mathscr{D}_{\mathrm{el}}(c) = \{I(a, b|V) \mid C \subseteq V \subseteq ABC - \{a, b\}, a \in A, b \in B\}$. Moreover, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) = \bigcup_{c' \in \mathscr{C}} \mathscr{D}_{\mathrm{el}}(c')$.

Given a set of arbitrary CI statements $\mathscr{C}$ and an arbitrary CI statement c, we consider the sets of elementary CI statements $\mathscr{D}_{\mathrm{el}}(\mathscr{C})$ and $\mathscr{D}_{\mathrm{el}}(c)$. In Theorem 15, we state that $\mathscr{C}^+_{\mathscr{G}} = \mathscr{D}_{\mathrm{el}}(\mathscr{C})^+_{\mathscr{G}}$ and $\{c\}^+_{\mathscr{G}} = \mathscr{D}_{\mathrm{el}}(c)^+_{\mathscr{G}}$ (and hence also $\mathscr{C}^+ = \mathscr{D}_{\mathrm{el}}(\mathscr{C})^+$ and $\{c\}^+ = \mathscr{D}_{\mathrm{el}}(c)^+$).[3] We also present System $\mathscr{E}$, shown in Fig. 6, which involves elementary CI statements only. In Theorem 16, we show that $\mathscr{C} \vdash c$ if and only if, for all $c_{\mathrm{el}} \in \mathscr{D}_{\mathrm{el}}(c)$, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \vdash_{\mathscr{E}} c_{\mathrm{el}}$.

These results yield a *normal form* for derivations under System $\mathscr{A}$: first, from the given set of CI statements, derive a set of elementary CI statements from the given set of CI statements under System $\mathscr{G}$; then, from this set, derive a new set of elementary CI statements under System $\mathscr{E}$; and, finally, from this last set, derive the targeted CI statement under inference System $\mathscr{G}$. (Notice that System $\mathscr{E}$ can be inferred from System $\mathscr{A}$.)

We start with the first and last stage of this normalization, which is straightforward.

Theorem 15. *Let c be an arbitrary CI statement and let $\mathscr{C}$ be a set of arbitrary CI statements. Then, $\{c\}^+_{\mathscr{G}} = \mathscr{D}_{\mathrm{el}}(c)^+_{\mathscr{G}}$ and $\mathscr{C}^+_{\mathscr{G}} = \mathscr{D}_{\mathrm{el}}(\mathscr{C})^+_{\mathscr{G}}$, and hence also $\{c\}^+ = \mathscr{D}_{\mathrm{el}}(c)^+$ and $\mathscr{C}^+ = \mathscr{D}_{\mathrm{el}}(\mathscr{C})^+$.*

We now turn to the second stage of the normalization, which involves System $\mathscr{E}$, shown in Fig. 6. System $\mathscr{E}$ is constructed in such a way that *only elementary CI statements* can be derived. In addition, it has the following property.

Theorem 16. *Let $\mathscr{C}$ be a set of* elementary *CI statements, and let c be a single* elementary *CI statement. Then $\mathscr{C} \vdash_{\mathscr{E}} c$ if and only if $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$.*

Proof. System $\mathscr{E}$ can be inferred from System $\mathscr{A}$. The "only if" direction then follows from Theorems 11 and 15. The proof of the "if" direction is a straightforward adaptation of the proof of the "if" direction of Theorem 11. $\square$

Theorems 11, 15, and 16 yield the following corollary.

[3] We emphasize that the sets of elementary CI statements considered here are *not* witness decompositions of the corresponding arbitrary CI statements. While a witness decomposition can be obtained using only some rules of System $\mathscr{G}$ (namely, the *decomposition* and *symmetry* rules), reconstruction of the original CI statement requires the *strong union* rule.

Corollary 17. *Let $\mathscr{C}$ be an arbitrary set of CI statements, and let c be a single arbitrary CI statement. Then the following statements are equivalent*:

- $\mathscr{C} \vdash c$;
- *for all* $c_{\mathrm{el}} \in \mathscr{D}_{\mathrm{el}}(c)$, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \vdash c_{\mathrm{el}}$; *and*
- *for all* $c_{\mathrm{el}} \in \mathscr{D}_{\mathrm{el}}(c)$, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \vdash_{\mathscr{E}} c_{\mathrm{el}}$.

6. The additive implication problem

An important insight into the study of the probabilistic implication problem for CI statements was gained by Studený's seminal work on conditional independence structures [6]. He showed that, for every CI statement $I(A, B|C)$, a discrete probability measure P *satisfies* $I(A, B|C)$, meaning that $P(C)P(ABC) = P(AC)P(BC)$ for *all possible* assignments to the variables in the sets A, B, and C, if and only if the *multi-information function* M_P induced by P, a function defined in Section 7, satisfies $M_P(C) + M_P(ABC) = M_P(AC) + M_P(BC)$. This relationship is non-trivial and remarkable because it relates a condition that is defined by numerous constraints (one per variable assignment) to a single constraint over an additive set function. Hence, he reduced the *multiplication-based* probabilistic CI implication problem to an *addition-based* implication problem. Therefore, we first develop the mathematical machinery for the *additive implication problem* in this section. This theory is then directly applicable to analyze the probabilistic implication problem via Studený's transformation.

6.1. Preliminaries and problem statement

We first introduce some terminology. In this work, a *real-valued function* is a function $F : 2^S \to \mathbf{R}$, associating a real number to each subset of S.

Definition 18. Let $I(A, B|C)$ be a CI statement, and F be a real-valued function. We say that F *a-satisfies* $I(A, B|C)$, denoted $\models^a_F I(A, B|C)$, if $F(C) + F(ABC) = F(AC) + F(BC)$.

For a real-valued function F and a class of CI statements $\mathscr{C}$, we say that F a-satisfies $\mathscr{C}$, denoted $\models^a_F \mathscr{C}$, if F a-satisfies each CI statement in $\mathscr{C}$.

Definition 19. Let $\mathscr{C}$ be a set of CI statements, c be a single CI statement, and $\mathscr{F}$ be a class of real-valued functions. We say that $\mathscr{C}$ *a-implies* c relative to $\mathscr{F}$, denoted $\mathscr{C} \models^a_{\mathscr{F}} c$, if each function in $\mathscr{F}$ that a-satisfies all the CI statements in $\mathscr{C}$ also a-satisfies c.

Now, given a class of real-valued functions $\mathscr{F}$ and classes of CI statements $\mathfrak{C}_1$ and $\mathfrak{C}_2$, the *additive implication problem* $(\mathfrak{C}_1, \mathfrak{C}_2)_{\mathscr{F}}$ is to decide whether $\mathscr{C} \models^a_{\mathscr{F}} c$, for $\mathscr{C} \subseteq \mathfrak{C}_1$ and $c \in \mathfrak{C}_2$.

We proceed by defining the notion of density of a real-valued function.

Definition 20. The *density* of the real-valued function F is the real-valued function ΔF defined by $\Delta F(X) = \sum_{U \in [X,S]} (-1)^{|U|-|X|} F(U)$, for each $X \subseteq S$.

The following well-known relationship between a real-valued function and its density, also known as its *Möbius inversion*, justifies the name (see, e.g., [38]).

Proposition 21. *Let F be a real-valued function. Then, for each $X \subseteq S$, $F(X) = \sum_{U \in [X,S]} \Delta F(U)$.*

The *a-satisfaction* of a real-valued function for a CI statement can be characterized in terms of an equation involving its density function. This characterization is central in developing our results and follows from a more general result in [31].

Proposition 22. *Let $I(A, B|C)$ be a CI statement and F be a real-valued function. Then, $\models^a_F I(A, B|C)$ if and only if $\sum_{U \in \mathscr{L}(A,B|C)} \Delta F(U) = 0$.*

In the remainder of this section, we study properties of classes of real-valued functions that guarantee soundness and/or completeness of inference systems subsumed by System $\mathscr{A}$ for additive implication. These results are related to probabilistic conditional independence implication in Section 7.

6.2. General soundness properties

We first define *soundness*, and then exhibit several characterizations.

Definition 23. Let $\mathscr{I}$ be an inference system for CI statements, $\mathscr{F}$ be a class of real-valued functions, and $\mathfrak{C}_1$ and $\mathfrak{C}_2$ be classes of CI statements. Then, $\mathscr{I}$ is *sound* for the additive implication problem $(\mathfrak{C}_1, \mathfrak{C}_2)_{\mathscr{F}}$ if, for each set of CI statements $\mathscr{C} \subseteq \mathfrak{C}_1$, and for each single CI statement $c \in \mathfrak{C}_2$, $\mathscr{C} \vdash_{\mathscr{I}} c$ implies $\mathscr{C} \models^a_{\mathscr{F}} c$.

In the remainder of Section 6, we shall consider $\mathfrak{C}_{\text{all}}$, the class of all CI statements over the universe S, and $\mathfrak{C}_{\text{sat}}$, the class of all saturated CI statements over S.

In order to characterize soundness for additive implication for several inference systems, we introduce the *zero-density property*.

Definition 24. A class of real-valued functions $\mathscr{F}$ has the *zero-density property* on a set of CI statements $\mathfrak{C}$ if, for each real-valued function $F \in \mathscr{F}$ and for each CI statement $c \in \mathfrak{C}$, $\models^a_F c$ implies $\Delta F(U) = 0$ for each $U \in \mathscr{L}(c)$.

Example 30 lists several real-valued functions illustrating the zero-density property.

We now provide characterizations of soundness for System $\mathscr{A}$:

Theorem 25. *Let $\mathscr{I}$ consist of the* symmetry, weak union, *and* strong union *rules. Let $\mathscr{F}$ be a class of real-valued functions. The following statements are equivalent*:

(1) *System $\mathscr{I}$ is sound for the additive implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$;*
(2) *$\mathscr{F}$ has the zero-density property on $\mathfrak{C}_{\text{all}}$; and*
(3) *System $\mathscr{A}$ is sound for the additive implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$.*

Proof. (1) $\Rightarrow$ (2). Let $I(A, B|C)$ be a CI statement, and let $F \in \mathscr{F}$ be a real-valued function such that $\models^a_F I(A, B|C)$. We show that $\Delta F(V) = 0$, for each $V \in \mathscr{L}(A, B|C)$. The proof goes by downward induction on $\mathscr{L}(A, B|C)$. By Proposition 7, we have $\mathscr{L}(A, B|C) = \bigcup_{W \in \mathscr{W}(I(A,B|C))} [C, \overline{W}]$. Hence, for the base case, we must prove that $\Delta F(\overline{W}) = 0$, for each $W \in \mathscr{W}(A, B|C)$. Let $W = \{a, b\} \in \mathscr{W}(I(A, B|C))$, with $a \in A$ and $b \in B$. Now, $I(a, b|C \cup (A - \{a\}) \cup (B - \{b\}))$ is derivable from $I(A, B|C)$ using *weak union* and *symmetry*. Furthermore, $I(a, b|\overline{W})$ is derivable from $I(a, b|C \cup (A - \{a\}) \cup (B - \{b\}))$ using *strong union*, since $C \cup (A - \{a\}) \cup (B - \{b\}) \subseteq \overline{W}$. Thus, $\{I(A, B|C)\} \vdash_{\mathscr{I}} I(a, b|\overline{W})$, and, hence, by the soundness of $\mathscr{I}$, $\{I(A, B|C)\} \models^a_{\mathscr{F}} I(a, b|\overline{W})$. Since $\models^a_F I(A, B|C)$, it follows that $\models^a_F I(a, b|\overline{W})$. Because $\mathscr{L}(a, b|\overline{W}) = \{\overline{W}\}$, we can invoke Proposition 22 to conclude that $\Delta F(\overline{W}) = 0$. For the inductive step, let $V \in \mathscr{L}(A, B|C)$. The inductive hypothesis states that $\Delta F(U) = 0$ for all $U \in \mathscr{L}(A, B|C)$ that are strict supersets of V. Similar to the base case, we can infer that $\models^a_F I(A', B'|V)$ with $A' = A - V$, $B' = B - V$, and V pairwise disjoint. Hence, by Proposition 22, $\sum_{U \in \mathscr{L}(A',B'|V)} \Delta F(U) = 0$ (1). Clearly, $\mathscr{L}(A', B'|V) \subseteq \mathscr{L}(A, B|C)$. Also, all sets in $\mathscr{L}(A', B'|V)$ are supersets of V. Hence, by the inductive hypothesis, (1) reduces to $\Delta F(V) = 0$.

(2) $\Rightarrow$ (3). Let $\mathscr{C}$ be a set of CI statements and c a single a CI statement for which $\mathscr{C} \vdash c$. Let $F \in \mathscr{F}$ for which $\models^a_F \mathscr{C}$. Since $\mathscr{F}$ has the zero-density property, we have that, for each $U \in \mathscr{L}(\mathscr{C})$, $\Delta F(U) = 0$. By Theorem 11, $\mathscr{C} \vdash c$ implies $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$. Hence, by Property 22, $\models^a_F c$.

(3) $\Rightarrow$ (1). All rules of System $\mathscr{I}$ belong to System $\mathscr{A}$, except for *weak union*, which can be inferred from it (Proposition 3). $\square$

6.3. Soundness properties for saturated CI statements

We now discuss the additive implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$.

Theorem 26. *Let $\mathscr{I}$ consist of the* symmetry *and* weak union *rules. Let $\mathscr{F}$ be a class of real-valued functions. The following statements are equivalent*:

(1) *System $\mathscr{I}$ is sound for the additive implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$;*
(2) *$\mathscr{F}$ has the zero-density property on $\mathfrak{C}_{\text{sat}}$;*
(3) *System $\mathscr{A}$ is sound for the additive implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$; and*
(4) *System $\mathscr{G}$ is sound for the additive implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{F}}$.*

Proof. (1) $\Rightarrow$ (2). Let $I(A, B|C)$ be a saturated CI statement, and let $F \in \mathscr{F}$ be a real-valued function such that $\models^a_F I(A, B|C)$. We show that $\Delta F(V) = 0$, for each $V \in \mathscr{L}(A, B|C)$. The proof is similar to the corresponding part of that of Theorem 25, except that we may not use *strong union*. In the base case of the downward induction, we argued that, for $W = \{a, b\} \in \mathscr{W}(I(A, B|C))$, with $a \in A$ and $b \in B$, the CI statement $I(a, b|C \cup (A - \{a\}) \cup (B - \{b\}))$ can be inferred from $I(A, B|C)$ using *weak union* and *symmetry*. Since $S = ABC$, that CI statement equals $I(a, b|\overline{W})$. A similar argument applies to the inductive step.

(2) $\Rightarrow$ (3). Analogous to the corresponding part of the proof of Theorem 25.

Table 1
Densities of the real-valued functions considered in Example 30.

Density	$\emptyset$	a	b	c	bc	ac	ab	abc
ΔF_{10}	0.1	0	0	0	0	0	0	0
ΔF_{11}	0	−0.3	0	0	0	0	0	0
ΔF_{12}	0	0	−0.6	0	0	0	0	0
ΔF_{13}	0	0	0	0.9	0	0	0	0
ΔF_2	−0.2	0.2	0.6	0.3	0	0	0	0

(3) $\Rightarrow$ (4). System $\mathscr{G}$ can be inferred from System $\mathscr{A}$ (Proposition 3).
(4) $\Rightarrow$ (1). System $\mathscr{I}$ is a subset of System $\mathscr{G}$. $\square$

6.4. General completeness properties

We first define *completeness*, and then study it as in Section 6.2.

Definition 27. Let $\mathscr{I}$ be an inference system for CI statements, $\mathscr{F}$ be a class of real-valued functions, and $\mathfrak{C}_1$ and $\mathfrak{C}_2$ be classes of CI statements. Then, $\mathscr{I}$ is *complete* for the additive implication problem $(\mathfrak{C}_1, \mathfrak{C}_2)_{\mathscr{F}}$ if, for each set of CI statements $\mathscr{C} \subseteq \mathfrak{C}_1$, and for each single CI statement $c \in \mathfrak{C}_2$, $\mathscr{C} \models^a_{\mathscr{F}} c$ implies $\mathscr{C} \vdash_{\mathscr{I}} c$.

In order to characterize completeness for additive implication for System $\mathscr{A}$, we introduce the *Kronecker property*.

Definition 28. A class of real-valued functions $\mathscr{F}$ has the *Kronecker property* on $\Omega \subseteq 2^S$ if, for each $U \in \Omega$, there exists $F_U \in \mathscr{F}$ such that $\Delta F_U(U) \neq 0$ and, for all $X \in \Omega$, $X \neq U$, $\Delta F_U(X) = 0$.

The name *Kronecker property* is inspired by the observation that, on Ω, ΔF_U behaves as a Kronecker delta function, which is zero everywhere except in one point.

Let $\Omega^{(2)} = \{V \subset S \mid |\overline{V}| \geqslant 2\}$. Relative to a class of real-valued functions, the Kronecker property on $\Omega^{(2)}$ implies the completeness of System $\mathscr{A}$:

Theorem 29. *Let $\mathscr{F}$ be a class of real-valued functions. If $\mathscr{F}$ has the Kronecker property on $\Omega^{(2)}$, then System $\mathscr{A}$ is complete for the additive implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}}$.*

Proof. Let $\mathscr{F}$ have the Kronecker property on $\Omega^{(2)}$, and let $\mathscr{C} \models^a_{\mathscr{F}} c$. Let $\mathscr{C}$ be a set of CI statements, and let c be a single CI statement. Assume $\mathscr{C} \nvdash c$, or, equivalently, $\mathscr{L}(c) \nsubseteq \mathscr{L}(\mathscr{C})$ (Theorem 11). By Proposition 7, all sets in $\mathscr{L}(\mathscr{C})$ and $\mathscr{L}(c)$ are in $\Omega^{(2)}$. Now, let $U \in \mathscr{L}(c) - \mathscr{L}(\mathscr{C})$. Since $\mathscr{F}$ has the Kronecker property on $\Omega^{(2)}$, we know there exists $F_U \in \mathscr{F}$ such that $\Delta F_U(U) \neq 0$ and, for all $X \in \Omega^{(2)}$, $X \neq U$, $\Delta F_U(X) = 0$. From Proposition 22, it readily follows that $\models^a_F \mathscr{C}$, while $\not\models^a_F c$. Hence, $\mathscr{C} \not\models^a_{\mathscr{F}} c$. By contraposition, $\mathscr{C} \models^a_{\mathscr{F}} c$ implies $\mathscr{C} \vdash c$. $\square$

The following example illustrates the zero-density and Kronecker properties.

Example 30. Let $S = \{a, b, c\}$. Let $\mathscr{F}_1 = \{F_{10}, F_{11}, F_{12}, F_{13}\}$ and $\mathscr{F}_2 = \{F_2\}$ as defined by their densities given in Table 1. Now, $\Omega^{(2)} = \{\emptyset, a, b, c\}$. $\mathscr{F}_1$ has the Kronecker property on $\Omega^{(2)}$, as F_{10}, F_{11}, F_{12}, and F_{13} satisfy Definition 28 for, respectively, $\emptyset$, a, b, and c. Hence, by Theorem 29, System $\mathscr{A}$ is complete for the additive implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}_1}$. We next investigate if $\mathscr{F}_1$ satisfies the zero-density property. First, consider F_{10}. The zero-density property relative to F_{10} can only be broken by a constraint $I(A, B|C)$ if $\emptyset \in \mathscr{L}(A, B|C)$, i.e., if $C = \emptyset$, $A \neq \emptyset$, and $B \neq \emptyset$. However, F_{10} does not a-satisfy this constraint, as, in the defining condition, $F_{10}(C) + F_{10}(ABC) = F_{10}(AC) + F_{10}(BC)$, the first term is non-zero, while the others are zero. Similar arguments can be given for F_{11}, F_{12} and F_{13}. Hence, $\mathscr{F}_1$ also satisfies the zero-density property, and, by Theorems 25 and 29, System $\mathscr{A}$ is sound for the additive implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}_1}$.

Finally, consider $\mathscr{F}_2 = \{F_2\}$, which does *not* satisfy the Kronecker property, as it does not contain functions whose densities are non-zero on only one member of $\Omega^{(2)}$. It also does not have the zero-density property as $\models^a_{F_2} I(b, c|\emptyset)$ but $\Delta F_2(\emptyset) \neq 0$. By Theorem 25, System $\mathscr{A}$ is *not* sound for the additive implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}_2}$. Theorem 29 does not allow us to make any statements about the completeness of System $\mathscr{A}$ in this case.

Theorem 29 goes only in one direction, but has the following weak converse.

Theorem 31. *Let $\mathscr{F}$ be a class of real-valued functions. If $\mathscr{A}$ is sound for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}}$ and complete for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{sat}})_{\mathscr{F}}$, then $\mathscr{F}$ has the Kronecker property on $\Omega^{(2)}$.*

Table 2
Example of binary probability distribution over $S = \{A, B, C\}$.

A	0	0	0	0	1	1	1	1
B	0	0	1	1	0	1	1	1
C	0	1	0	1	0	1	0	1
P	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.25

Proof. If $|S| \leqslant 1$, then $\Omega^{(2)} = \emptyset$ and the statement is true. Thus, assume $|S| \geqslant 2$. For $U \in \Omega^{(2)}$, let $\mathscr{C}$ be the set of non-trivial saturated CI statements $I(A, B|C)$ which satisfy either (1) $A \subseteq U$ or (2) $C - U \neq \emptyset$. By Proposition 7, $\mathscr{L}(\mathscr{C}) \subseteq \Omega^{(2)}$. By construction, $U \notin \mathscr{L}(\mathscr{C})$. Now, let $X \in \Omega^{(2)}$, $X \neq U$. If $\overline{X}$ meets both U and $\overline{U}$, then let $X_1 = \overline{X} \cap U$ and $X_2 = \overline{X} \cap \overline{U}$. Otherwise, let X_1 and X_2 be any two nonempty disjoint sets such that $X_1 \cup X_2 = \overline{X}$. It is easily verified that, in all cases, $I(X_1, X_2|X)$ is in $\mathscr{C}$, and, hence $X \in \mathscr{L}(\mathscr{C})$. Thus, $\mathscr{L}(\mathscr{C}) = \Omega^{(2)} - \{U\}$. Now, let U_1 and U_2 be any two nonempty disjoint sets such that $U_1 \cup U_2 = \overline{U}$, and let c be $I(U_1, U_2|U)$. Since $U \in \mathscr{L}(c)$, $\mathscr{L}(c) - \mathscr{L}(\mathscr{C}) \neq \emptyset$. By Theorem 11, $\mathscr{C} \nvdash c$, and, by the completeness of $\mathscr{A}$ for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{sat}})_{\mathscr{F}}$, $\mathscr{C} \not\models^a_{\mathscr{F}} c$. Hence, there exists $F_U \in \mathscr{F}$ such that $\models^a_{F_U} \mathscr{C}$, but $\not\models^a_{F_U} c$. Since $\mathscr{A}$ is also sound for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}}$, $\mathscr{F}$ has the zero-density property on $\mathfrak{C}_{\mathrm{sat}}$ by Theorem 26. Thus, for all $X \in \mathscr{L}(\mathscr{C}) = \Omega^{(2)} - \{U\}$, $\Delta F_U(X) = 0$, but, by Proposition 22, $\Delta F_U(U) \neq 0$, since $\not\models^a_{F_U} c$. $\square$

Theorem 31 and Theorem 29 immediately yield the following corollary.

Corollary 32. *Let $\mathscr{F}$ be a class of real-valued functions. If $\mathscr{A}$ is sound for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}}$ and complete for the additive implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{sat}})_{\mathscr{F}}$, then $\mathscr{A}$ is complete for the additive implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{F}}$.*

7. The probabilistic CI implication problem

Although the theory developed in Section 6 concerns the additive implication problem for CI statements, it is also applicable to the probabilistic implication problem. The link between both is made with the notion of *multi-information functions* (Studený [6]) induced by probability measures.[4]

7.1. Preliminaries and problem statement

We first introduce some notations and terminology for stating and studying the probabilistic implication problem. A *probability model* over $S = \{s_1, \ldots, s_n\}$ is a pair (dom, P), where dom is a domain mapping associating s_i to a finite domain $dom(s_i)$, for $i = 1, \ldots, n$, and P is a probability measure having $dom(s_1) \times \cdots \times dom(s_n)$ as its sample space. For $i_1, \ldots, i_k$ a subsequence of $1, \ldots, n$ and $A = \{s_{i_1}, \ldots, s_{i_k}\} \subseteq S$, $\mathbf{a}$ is a *domain vector* of A if $\mathbf{a} \in dom(s_{i_1}) \times \cdots \times dom(s_{i_k})$. In what follows, we focus on probability measures, leaving their probability models implicit. If all domains of the underlying probability model are binary, i.e., consisting of two values, we call the probability measure *binary*. We denote the class of all probability measures by $\mathscr{P}_{\mathrm{all}}$ and the class of all binary probability measures by $\mathscr{P}_{\mathrm{bin}}$. For a probability measure P and $A \subseteq S$, we denote by P^A the marginal probability measure of P over A, i.e., for a domain vector $\mathbf{a}$ of A, $P^A(\mathbf{a}) = \sum_{\mathbf{b}} P(\mathbf{a}, \mathbf{b})$, where $\mathbf{b}$ ranges over all domain vectors of $S - A$.

We now define when a probability measure satisfies a CI statement.

Definition 33. Let $I(A, B|C)$ be a CI statement, and P be a probability measure. Then, P *m-satisfies* $I(A, B|C)$, denoted $\models^m_P I(A, B|C)$, if, for all domain vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ of A, B, and C, respectively, $P^C(\mathbf{c})P^{ABC}(\mathbf{a}, \mathbf{b}, \mathbf{c}) = P^{AC}(\mathbf{a}, \mathbf{c})P^{BC}(\mathbf{b}, \mathbf{c})$.

Relative to the *multiplicative* notion of *m-satisfaction*, we now define *probabilistic implication* for CI statements.

Definition 34. Let $\mathscr{C}$ be a set of CI statements, c be a single CI statement, and $\mathscr{P}$ be a class of discrete probability measures. Then, $\mathscr{C}$ *m-implies* c relative to $\mathscr{P}$, denoted $\mathscr{C} \models^m_{\mathscr{P}} c$, if each probability measure P in $\mathscr{P}$ that m-satisfies all the CI statements in $\mathscr{C}$ also m-satisfies c.

For $\mathscr{C}$, c, and $\mathscr{P}$ as above, we denote by $\mathscr{C}^*_{\mathscr{P}}$ the set $\{c \mid \mathscr{C} \models^m_{\mathscr{P}} c\}$, or $\mathscr{C}^*$, if the class of discrete probability measures involved is clear from the context.

Before continuing, we illustrate some of the concepts above by an example.

Example 35. Let $S = \{A, B, C\}$, and P be the binary probability distribution shown in Table 2. Clearly, $P^{AB}(0,0) = P^{AB}(0,1) = P^{AB}(1,0) = P^{AB}(1,1) = 0.25$ and $P^A(0) = P^A(1) = P^B(0) = P^B(1) = 0.50$. Hence, for each domain vector $\mathbf{u}$

---

[4] In this work, we only consider *discrete* probability measures.

of S, $P^{\emptyset}(\pi_{\emptyset}(\mathbf{u}))P^{AB}(\pi_{AB}(\mathbf{u})) = P^{A}(\pi_{A}(\mathbf{u}))P^{B}(\pi_{B}(\mathbf{u}))$, where "$\pi_X$" denotes projection onto X, always evaluates to *true*. We may therefore conclude that P m-satisfies $I(A, B|\emptyset)$. However, $P^{C}(0)P^{ABC}(0,0,0) = 0.75 \times 0.25$, while $P^{AC}(0,0)P^{BC}(0,0) = 0.50 \times 0.50$ yields a different result. Hence, P does *not* m-satisfy $I(A, B|C)$.

Given a class of discrete probability measures $\mathscr{P}$ and classes of CI statements $\mathfrak{C}_1$ and $\mathfrak{C}_2$, the *probabilistic implication problem* $(\mathfrak{C}_1, \mathfrak{C}_2)_{\mathscr{P}}$ is to decide whether $\mathscr{C} \models^{m}_{\mathscr{P}} c$, for $\mathscr{C} \subseteq \mathfrak{C}_1$ and $c \in \mathfrak{C}_2$.

Next, we define the Kullback–Leibler divergence [39].

Definition 36. Let P and Q be two probability measures satisfying $Q(\mathbf{s}) > 0$ whenever $P(\mathbf{s}) > 0$, for all domain vectors $\mathbf{s}$ of S. Then, the *Kullback–Leibler divergence* H is defined as

$$H(P|Q) = \sum_{\mathbf{s}} P(\mathbf{s}) \log \frac{P(\mathbf{s})}{Q(\mathbf{s})},$$

with $\mathbf{s}$ ranging over all domain vectors of S for which $P(\mathbf{s}) > 0$.

For P and Q as above, it is a well-know property [39] that $H(P|Q) \geqslant 0$, with $H(P|Q) = 0$ if and only if $P = Q$.

We now define *multi-information functions* induced by a probability measures [6].

Definition 37. Let P be a probability measure. The *multi-information function* induced by P is the real-valued function $M_P : 2^S \to \mathbf{R}$ defined by

$$\begin{cases} M_P(\emptyset) = 0; \\ M_P(A) = H\left(P^A \middle| \prod_{a \in A} P^a\right), & A \subseteq S,\ A \neq \emptyset. \end{cases}$$

A real-valued function $F : 2^S \to \mathbf{R}$ is *isotone* if, for all disjoint subsets A and B of S, $F(AB) \geqslant F(A)$, and *supermodular* if, for all pairwise disjoint subsets A, B, and C of S, $F(A) + F(ABC) \geqslant F(AB) + F(AC)$. Studený [6] showed that multi-information functions are isotone and supermodular. By a previous remark, multi-information functions are also *nonnegative*. The class of multi-information functions induced by the class of discrete probability measures $\mathscr{P}$ will be denoted by $\mathscr{M}_{\mathscr{P}}$.

Studený [6] showed, for a CI statement c, that $\models^{a}_{\mathscr{M}_{\mathscr{P}}} c$ if and only if $\models^{m}_{\mathscr{P}} c$. With this result, we can reduce the (multiplicative) probabilistic implication problem for CI statements to the additive implication problem for CI statements, the advantage being that additive problems are usually easier to deal with than multiplicative problems. Moreover, we can then apply the machinery developed in Section 6.

7.2. Soundness and completeness properties

Definitions 23 and 27 for soundness and completeness of an inference system for the additive implication problem $(\mathfrak{C}_1, \mathfrak{C}_2)_{\mathscr{F}}$ ($\mathfrak{C}_1$ and $\mathfrak{C}_2$ are classes of CI statements and $\mathscr{F}$ is a class of real-valued functions) can be adapted to the setting of probabilistic implication, provided $\mathscr{F}$ is replaced by a class of discrete probability measures $\mathscr{P}$, and a-implication ($\models^{a}$) is replaced by m-implication ($\models^{m}$).

Example 38. From Example 35, we may conclude that *strong union* is *not* sound for the probabilistic implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{bin}}}$. Hence, it is also not sound for the probabilistic implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{all}}}$.

It is known (e.g., [40]) that the semi-graphoid axioms are sound for the probabilistic implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{all}}}$. However, System $\mathscr{G}$ is not complete. Moreover, there cannot exist a finite set of inference rules that is sound and complete for the implication problem [6]. In addition, it is unknown whether the problem is decidable. For saturated CI statements, the situation is different. Malvestuto [26] and Geiger and Pearl [14] independently showed that System $\mathscr{S}$ is sound and complete for the probabilistic implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{sat}})_{\mathscr{P}_{\text{all}}}$ (see also [40]). Using the results of Section 6, we can bootstrap these results, as follows.

Theorem 39.

(1) *The class of multi-information functions $\mathscr{M}_{\mathscr{P}_{\text{all}}}$ has the zero-density property on $\mathfrak{C}_{\text{sat}}$.*
(2) *System $\mathscr{A}$ is sound for the probabilistic implication problem $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{all}}}$.*

Proof. Since System $\mathscr{G}$ is sound for the general probabilistic implication problem $(\mathfrak{C}_{\text{all}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{all}}}$, it is also sound for $(\mathfrak{C}_{\text{sat}}, \mathfrak{C}_{\text{all}})_{\mathscr{P}_{\text{all}}}$. The result now follows immediately from Theorem 26. $\square$

Theorem 40.

(1) *The class of multi-information functions* $\mathscr{M}_{\mathscr{P}_{\mathrm{all}}}$ *has the Kronecker property on* $\Omega^{(2)}$.
(2) *System* $\mathscr{A}$ *is complete for the probabilistic implication problem* $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{all}}}$.

Proof. By Theorem 39(2), System $\mathscr{A}$ is sound for the probabilistic implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{all}}}$. Since System $\mathscr{S}$ is already complete for the probabilistic implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{sat}})_{\mathscr{P}_{\mathrm{all}}}$, it follows in particular that System $\mathscr{A}$ is. The result now follows immediately from Theorem 31 and Corollary 32. □

It can be shown by direct calculation that the class of multi-information functions induced by the class of *binary* probability measures has also the Kronecker property on $\Omega^{(2)}$. Therefore, System $\mathscr{A}$ is also complete for the probabilistic implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{bin}}}$. Please note that completeness of System $\mathscr{A}$ with respect to a class of probability functions $\mathscr{P}$ does not automatically imply completeness with respect to a subclass $\mathscr{P}' \subseteq \mathscr{P}$.

We conclude this section by a sound-and-completeness result.

Theorem 41. *Both System* $\mathscr{A}$ *and System* $\mathscr{G}$ *are sound and complete for the probabilistic implication problems* $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{all}}}$ *and* $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{bin}}}$.

Proof. It only remains to show that System $\mathscr{G}$ is complete for both probabilistic implication problems. This is because System $\mathscr{G}$ is complete for the probabilistic implication problem $(\mathfrak{C}_{\mathrm{sat}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}}$, where $\mathscr{P}$ is an arbitrary class of probability measures, if and only if System $\mathscr{A}$ is [41]. □

Notice that, in Theorem 41, we do *not* assume that the CI statements to be derived are saturated.

8. Falsification and validation criteria

Decidability of the CI implication problem over the class of discrete probability measures is open, but it is known it has no finite sound and complete inference system [15]. Nonetheless, Theorem 40 demonstrates it has a finite complete inference system, namely System $\mathscr{A}$, which is useful since it provides a falsification criterion to decide some non-valid instances (Section 8.2). Of course, for instances which cannot so be falsified, it would be desirable to have validation criteria to decide some valid ones among these. We develop such a validation criterion, based on linear programming techniques (Section 8.3). In preparation, we explore two simplifications (Section 8.1). First, we show that we can reduce the general probabilistic implication problem to the probabilistic implication problem for elementary CI statements only. Then, we demonstrate that it suffices to consider only those variables that occur in the CI statements under consideration. In Section 9, we will give experimental evidence that the combination of the falsification and validation criteria is useful for deciding a large set of instances of the probabilistic implication problem.

8.1. Simplifications

To reduce the probabilistic CI implication problem to one on elementary CI statements only, we build upon the work in Section 5, where elementary CI statements were considered from a purely syntactical point of view.

Proposition 42. *Let* $\mathscr{C}$ *be a set of CI statements, and c be a single CI statement. Then,*

(1) $\mathscr{C}^* = \mathscr{D}_{\mathrm{el}}(\mathscr{C})^*$ *and* $c^* = \mathscr{D}_{\mathrm{el}}(c)^*$;
(2) $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c$ *if and only if, for all* $c_{\mathrm{el}} \in \mathscr{D}_{\mathrm{el}}(c)$, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \models^m_{\mathscr{P}_{\mathrm{all}}} c_{\mathrm{el}}$.

Proof. It suffices to prove the first statement for a set of CI statements $\mathscr{C}$. By Theorem 15, $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \subseteq \mathscr{C}^+_{\mathscr{G}}$. Since System $\mathscr{G}$ is sound for the probabilistic implication problem $(\mathfrak{C}_{\mathrm{all}}, \mathfrak{C}_{\mathrm{all}})_{\mathscr{P}_{\mathrm{all}}}$, we have that $\mathscr{D}_{\mathrm{el}}(\mathscr{C}) \subseteq \mathscr{C}^*$, hence $\mathscr{D}_{\mathrm{el}}(\mathscr{C})^* \subseteq \mathscr{C}^*$. Theorem 15 also yields that $\mathscr{C} \subseteq \mathscr{D}_{\mathrm{el}}(\mathscr{C})^+_{\mathscr{G}}$, from which we derive $\mathscr{C}^* \subseteq \mathscr{D}_{\mathrm{el}}(\mathscr{C})^*$. □

We now turn to the second simplification, namely that variables that do not occur in the given CI statements need not be considered for the implication problem. In particular, this is also the case for hidden variables. For a set of CI statements $\mathscr{C}'$ over a finite universe S (the set of variables), let $\mathrm{var}(\mathscr{C}')$ be the set of variables occurring in at least one CI statement of $\mathscr{C}'$. Now consider an instance "$\mathscr{C} \models^m_{\mathscr{P}} c$?" of the probabilistic implication problem relative to S. Intuitively, the variables of S outside $\mathrm{var}(\mathscr{C} \cup \{c\})$ play no role. Nevertheless, restricting this instance of the probabilistic implication problem to the relevant variables also means restricting the probability measures considered. Therefore, we need to prove formally that our intuition is correct.

Lemma 43. *Let U and V be disjoint sets of variables. Let P_U be a probability measure over U. Then there exists a probability measure P_{UV} over UV such that, for each subset X of U, $P_{UV}^X = P_U^X$.*

Proof. Designate by "0" one element in the domain of each variable in UV. Let $\mathbf{u}$ be a domain vector of U and $\mathbf{v}$ a domain vector of V. Let $\mathbf{0}_V$ be the domain vector of V of which each component is 0. We define

$$\begin{cases} P_{UV}(\mathbf{u},\mathbf{v}) = 0 \quad \text{if } \mathbf{v} \neq \mathbf{0}_V; \\ P_{UV}(\mathbf{u},\mathbf{0}_V) = P_U(\mathbf{u}). \end{cases}$$

Clearly, P_{UV} is a probability measure over UV and, for $X \subseteq U$, $P_{UV}^X = P_U^X$. $\square$

For $U \subseteq S$, $\mathscr{P}_U$ denotes class of all probability measures over U.

Proposition 44. *Let U and V be disjoint sets of variables, $\mathscr{C}$ be a set of CI statements over U, and c be a single CI statement over U. Then, $\mathscr{C} \models^m_{\mathscr{P}_U} c$ if and only if $\mathscr{C} \models^m_{\mathscr{P}_{UV}} c$.*

Proof. We first prove the "only if". Thereto, let $P_{UV} \in \mathscr{P}_{UV}$ be such that P_{UV} m-satisfies all CI statements in $\mathscr{C}$. Let $P_U = P_{UV}^U$. Obviously, $P_U \in \mathscr{P}_U$, and, for $X \subseteq U$, $P_U^X = P_{UV}^X$. Hence, P_U m-satisfies all CI statements in $\mathscr{C}$. From $\mathscr{C} \models^m_{\mathscr{P}_U} c$, it now follows that $P_U = P_{UV}^U$ m-satisfies c, and hence that P_{UV} m-satisfies c. We now turn to the "if." Thereto, let $P_U \in \mathscr{P}$ be such that P_U m-satisfies all CI statements in $\mathscr{C}$. By Lemma 43, there exists a probability function $P_{UV} \in \mathscr{P}_{UV}$ such that $P_U = P_{UV}^U$. The remainder of the proof now goes along the same lines as for the "only if." $\square$

An immediate corollary to Proposition 44 is that, for a set of CI statements $\mathscr{C}$ and a single CI statement c, both over S, $\mathscr{C} \models^m_{\mathscr{P}_S} c$ if and only if $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{var}(\mathscr{C}\cup\{c\})}} c$, i.e., the targeted simplification. We shall, therefore, disregard irrelevant variables.

8.2. Falsification criterion

Theorems 11 and 40 yield a *falsification criterion*, i.e., a sufficient condition which can falsify certain instances of the probabilistic CI implication problem:

Corollary 45. *Let $\mathscr{C}$ be a set of CI statements. If $\mathscr{L}(c) \nsubseteq \mathscr{L}(\mathscr{C})$, then $\mathscr{C} \not\models^m_{\mathscr{P}_{\mathrm{all}}} c$.*

We recall from [5] that testing for lattice inclusion is coNP-complete, that there exists a linear-time reduction to SAT for this problem, and that one can leverage SAT solvers to decide semi-lattice inclusion effectively, even if several hundreds of variables are involved. Now, if the falsified implications were, on average, only a small fraction of all those that are falsifiable, the result would be disappointing from a practical point of view. Fortunately, we are not only able to show that a large number of implications can be falsified by the "lattice-exclusion" criterion identified in Corollary 45, but also that polynomial time heuristics exist that provide good approximations of said criterion. The falsification criterion and the two heuristics we consider are formally exhibited in Algorithms 1, 2, and 3 (Fig. 7).

It follows from Proposition 7 that, if one of the two heuristics returns *false*, then $\mathscr{L}(c) \nsubseteq \mathscr{L}(\mathscr{C})$, and hence, by Corollary 45, that $\mathscr{C} \not\models^m_{\mathscr{P}_{\mathrm{all}}} c$.

Example 46. The inference rule $I(A,B|DC)$ & $I(A,D|BC) \to I(A,BD|C)$, is *not* sound relative to the class of discrete probability measures. Falsification Heuristic 1 can reject this instance of the probabilistic implication problem.

The Falsification Criterion actually leads to a *family* of polynomial time heuristics. While Falsification Heuristic 1 checks if the greatest lower bound of $\mathscr{L}(c)$ is not in $\mathscr{L}(\mathscr{C})$ and Falsification Heuristic 2 checks if the least upper bounds of $\mathscr{L}(c)$ are not in $\mathscr{L}(\mathscr{C})$, we may select additional elements in that semi-lattice located between these two extremes to derive more falsification heuristics. Finally, we observe that there is also an alternative falsification strategy. The principal idea is to construct a discrete probability measure which provides a counterexample using the decidable theory of the reals with addition. This is almost what Bouckaert and Studený have done with their racing algorithms [8]. Needless to say, this rather expensive approach can produce false positives in the sense that it may falsify instances of the probabilistic implication problem that are valid. This was not problematic for their application, as they had a different goal in mind.

8.3. Validation criterion

A validation criterion for the probabilistic implication problem accepts an instance of the problem only if the implication is valid. If the criterion does not accept the instance, however, this does not imply that the instance is invalid. One of the most prominent validation criteria is the algorithm that computes the closure of the semi-graphoid axioms [2,3], which,

Falsification Criterion

Input: a set of CI statements $\mathscr{C}$ and a single CI statement c.

Method:

if $\mathscr{L}(c) \nsubseteq \mathscr{L}(\mathscr{C})$

then return false else return unknown.

Falsification Heuristic 1

Input: a set of CI statements $\mathscr{C}$ and a single CI statement $I(A,B|C)$.

Method:

if for each CI statement $I(A',B'|C') \in \mathscr{C}$, $C' \nsubseteq C$

then return false else return unknown.

Falsification Heuristic 2

Input: a set of CI statements $\mathscr{C}$ and a single CI statement $I(A,B|C)$.

Method:

if there exists $W \in \mathscr{W}(I(A,B|C))$ such that, for all $I(A',B'|C') \in \mathscr{C}$,

$W \notin \mathscr{W}(A',B'|C')$

then return false else return unknown.

Fig. 7. The different falsification criteria and heuristics derived from the completeness of inference system $\mathscr{A}$.

unfortunately, can only validate a small fraction of the set of verifiable instances. Here, we introduce a much more powerful validation criterion, based on the lattice-theoretic framework. We harness our results to represent a set of CI statements $\mathscr{C}$ as a sparse 0–1 matrix $\mathbf{A}_{\mathscr{C}}$ in which the rows correspond to elements of $\mathscr{L}(\mathscr{C})$ and the columns correspond to certain CI statements of the form $I(a,b|C)$ that are induced from $\mathscr{C}$. Subsequently, for a given CI statement c, we will associate with the instance $\mathscr{C} \models^m_{\mathscr{P}_{\text{all}}} c$ a linear program with equality constraints involving an encoding of c in terms of its associated semi-lattice $\mathscr{L}(c)$ and the matrix $\mathbf{A}_{\mathscr{C}}$. We then show that $\mathscr{C} \models^m_{\mathscr{P}_{\text{all}}} c$ is valid if this linear program has a solution. To achieve this, we introduce *relevant* elementary CI statements:

Definition 47. An elementary CI statement c is *relevant* to a set of CI statements $\mathscr{C}$ if $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$. The set of all such CI statements is denoted $\mathscr{R}(\mathscr{C})$.

Since System $\mathscr{A}$ is complete but not sound for the probabilistic implication problem (Theorem 40), $\mathscr{C} \vdash c$ or, equivalently, $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$ (Theorem 11) is necessary but not sufficient for $\mathscr{C} \models^m_{\mathscr{P}} c$, not all relevant statements associated with $\mathscr{C}$ need be probabilistically implied by $\mathscr{C}$. By Proposition 42, however, there is a subset of $\mathscr{R}(\mathscr{C})$ that is probabilistically equivalent to $\mathscr{C}$.

With a CI statement c over the universe S, we associate a vector $\mathbf{v}_c$ over 2^S:

$$\mathbf{v}_c(U) = \begin{cases} 1 & \text{if } U \in \mathscr{L}(c); \\ 0 & \text{if } U \notin \mathscr{L}(c). \end{cases}$$

In other words, $\mathbf{v}_c$ is a representation of the *characteristic function* of $\mathscr{L}(c)$.

Apart from being the set of all elementary CI statements equivalent to the given set of CI statements $\mathscr{C}$ for derivations in System $\mathscr{A}$, $\mathscr{R}(\mathscr{C})$ has the following attractive property which justifies calling $\mathscr{R}(\mathscr{C})$ a “basis” for $\mathscr{C}$.

Proposition 48. *Let $\mathscr{C}$ be a set of CI statements, and let c be a single CI statement such that $\mathscr{C} \vdash c$. Then, for each CI statement c_{el} in $\mathscr{R}(\mathscr{C})$, there exists a nonnegative integer number $k_{c_{\text{el}}}$ such that $\mathbf{v}_c = \sum_{c_{\text{el}} \in \mathscr{R}(\mathscr{C})} k_{c_{\text{el}}} \mathbf{v}_{c_{\text{el}}}$.*

Proof. With the semi-graphoid inference rules *decomposition* and *weak union*, we can show that $\mathscr{L}(A,bD|C) = \mathscr{L}(A,b|CD) \cup \mathscr{L}(A,D|C)$. Moreover, $\mathscr{L}(A,b|CD) \cap \mathscr{L}(A,D|C) = \emptyset$; indeed, the former set contains only supersets of CD, while the latter cannot contain supersets of C. Hence, $\mathbf{v}_{I(A,bD|C)} = \mathbf{v}_{I(A,b|CD)} + \mathbf{v}_{I(A,D|C)}$. By applying this repeatedly, it follows that, for each elementary CI statement c_{el} in $\mathscr{D}_{\text{el}}(c)$, there exists a nonnegative integer $k_{c_{\text{el}}}$ such that $\mathbf{v}_c = \sum_{c_{\text{el}} \in \mathscr{D}_{\text{el}}(c)} k_{c_{\text{el}}} \mathbf{v}_{c_{\text{el}}}$. The final result follows from observing that, by Corollary 17 and Theorems 15 and 11, $\mathscr{D}_{\text{el}}(c) \subseteq \mathscr{R}(\mathscr{C})$. $\square$

Observe that $I(a,b|C) \in \mathscr{R}(\mathscr{C})$ if and only if $I(b,a|C) \in \mathscr{R}(\mathscr{C})$, as both have the same semi-lattice. They also have the same associated vector. To avoid such duplication, we define $\mathscr{R}/2(\mathscr{C})$ to be $\mathscr{R}(\mathscr{C})$ in which CI statements of the form $I(a,b|C)$ and $I(b,a|C)$ are no longer distinguished. Clearly, Proposition 48 still holds when $\mathscr{R}(\mathscr{C})$ is replaced by $\mathscr{R}/2(\mathscr{C})$.

We still need a couple of additional results before we can move on to the construction of the linear program. The first one involved the multi-information function M_P of a discrete probability measure P.

Proposition 49. *Let A, B, and C be pairwise disjoint subsets of S, and let P be a discrete probability distribution. Then, $\sum_{U \in \mathscr{L}(A,B|C)} \Delta M_P(U) \geqslant 0$.*

Proof. For a real-valued function F, we have $F(C) - F(AC) - F(BC) + F(ABC) = \sum_{U \in \mathscr{L}(A,B|C)} \Delta F(U)$ [31]. If F is supermodular, the left-hand side is nonnegative. Since M_P is supermodular [6], Proposition 49 immediately follows. □

We have already defined the vector associated with a single CI statement. We now define the vector associated with a set of CI statements $\mathscr{C}$ as the sum of the vectors associated to the CI statements in $\mathscr{C}$: $\mathbf{v}_{\mathscr{C}} = \sum_{c \in \mathscr{C}} \mathbf{v}_c$.

Proposition 50. *Let $\mathscr{C}$ be a set of CI statements, c be a single CI statement, and k_c be a positive real number. If, for each CI statement $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, there exists a nonnegative real number $k_{c_{\mathrm{el}}}$ such that*

$$\mathbf{v}_{\mathscr{C}} = k_c \mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}} \mathbf{v}_{c_{\mathrm{el}}},$$

then $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c$. Moreover, $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c_{\mathrm{el}}$ for all $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$ for which $k_{c_{\mathrm{el}}} > 0$.

Proof. Let P be a discrete probability measure such that $\models^m_P \mathscr{C}$. By Proposition 49, we have that, for each CI statement c_{el} in $\mathscr{R}$, $\sum_{U \in \mathscr{L}(c_{\mathrm{el}})} \Delta M_P(U) \geqslant 0$. Hence,

$$\sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}} \sum_{U \in \mathscr{L}(c_{\mathrm{el}})} \Delta M_P(U) \geqslant 0.$$

Again by Proposition 49, we have that $\sum_{U \in \mathscr{L}(c)} \Delta M_P(U) \geqslant 0$. Now, since P satisfies all CI statements in $\mathscr{C}$, we also have that $\sum_{c' \in \mathscr{C}} \sum_{U \in \mathscr{L}(c')} \Delta M_P(U) = 0$, by Proposition 22. Now, we have that

$$0 = \sum_{c' \in \mathscr{C}} \sum_{U \in \mathscr{L}(c')} \Delta M_P(U) = k_c \sum_{U \in \mathscr{L}(c)} \Delta M_P(U) + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}} \sum_{U \in \mathscr{L}(c_{\mathrm{el}})} \Delta M_P(U),$$

where the second equality follows from the precondition in the statement of Proposition 50. Since both terms in the right-hand side of the above equation are nonnegative, and since $k_c > 0$, necessarily $\sum_{U \in \mathscr{L}(c)} \Delta M_P(U) = 0$. Thus, by Proposition 22, $\models^m_P c$. By the same token, $\models^m_P c_{\mathrm{el}}$ for all c_{el} for which $k_{c_{\mathrm{el}}} > 0$. □

Proposition 50 still holds if the sum ranges over all elementary CI statements. There is no point in considering those not relevant to $\mathscr{C}$, however, as a careful examination of the proof reveals that their coefficients in the sum are necessarily 0, and, therefore, they cannot contribute to the solutions of the above equation.

We write the equation in Proposition 50 as a linear program (Schrijver [42])

$$\text{minimize } \mathbf{c}^T \mathbf{x} \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geqslant \mathbf{0}.$$

As we are only interested in satisfying the constraint, that is, we are only interested in the existence of a feasible solution, we set $\mathbf{c} = \mathbf{0}$ (hence, the objective function $\mathbf{c}^T\mathbf{x}$ is zero). As to the parameters of the condition, we put $\mathbf{A} = \mathbf{A}_{\mathscr{C}}$, the matrix associated with $\mathscr{C}$, defined over $\mathscr{L}(\mathscr{C}) \times \mathscr{R}/2(\mathscr{C})$ by

$$\mathbf{A}_{\mathscr{C}}[U, c_{\mathrm{el}}] = \begin{cases} 1 & \text{if } U \in \mathscr{L}(c_{\mathrm{el}}); \\ 0 & \text{if } U \notin \mathscr{L}(c_{\mathrm{el}}), \end{cases}$$

and $\mathbf{b} = \mathbf{v}_{\mathscr{C}} - k_c \mathbf{v}_c$. Clearly, if $\mathbf{x}$ is a solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$, then $\mathbf{x}$ is a vector of values $k_{c_{\mathrm{el}}}$, $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, satisfying the equation in Proposition 50.

Example 51. Let $S = \{a, b, d, e, f\}$, and

$$\mathscr{C} = \{I(a,b|\emptyset), I(a,b|de), I(d,e|a), I(d,e|b)\}.$$

Based on the results from Section 8.1 concerning the set of relevant CI statements we can exclude the variable f. Then, $\mathscr{R}/2(\mathscr{C})$ equals

$\mathbf{A}_{\mathscr{C}}$	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
$\emptyset$	1	0	0	0	1	0	0	0
a	0	0	0	0	1	1	0	0
b	0	0	0	0	1	0	1	0
d	1	1	0	0	0	0	0	0
e	1	0	1	0	0	0	0	0
ab	0	0	0	0	1	1	1	1
de	1	1	1	1	0	0	0	0

Fig. 8. The matrix $\mathbf{A}_{\mathscr{C}}$ of Example 51.

$$\big\{I(a,b|\emptyset), I(a,b|d), I(a,b|e), I(a,b|de), I(d,e|\emptyset), I(d,e|a), I(d,e|b), I(d,e|ab)\big\}.$$

We denote these eight relevant elementary CI statements by $c_1,\dots,c_8$, respectively. Let $c = c_5 = I(d,e|\emptyset)$. Hence, $\mathrm{var}(\mathscr{C} \cup c) = \{a,b,d,e\}$.

The 0–1 matrix $\mathbf{A}_{\mathscr{C}}$ is shown in Fig. 8. The rows in $\mathbf{A}_{\mathscr{C}}$ correspond to the elements in $\mathscr{L}(\mathscr{C}) = \{\emptyset, a, b, d, e, ab, de\}$ computed relative to $\mathrm{var}(\mathscr{C} \cup c)$ rather than S. While the number of possible rows in the matrix is $2^{|S|} - |S| - 1$, here 26, we have first reduced it to $2^{|\mathrm{var}(\mathscr{C}\cup c)|} - |\mathrm{var}(\mathscr{C} \cup c)| - 1$, here 11, and then further to $|\mathscr{L}(\mathscr{C})|$ as computed relative to $\mathrm{var}(\mathscr{C} \cup c)$, here only 7. The columns correspond to the relevant elementary CI statements $c_1,\dots,c_8$, respectively. Actually, these columns equal the vectors $\mathbf{v}_{c_1},\dots,\mathbf{v}_{c_8}$, respectively. Notice that, given S, the number of elementary CI statements up to symmetry is

$$\binom{|S|}{2} 2^{|S|-2},$$

here 80. We have first reduced this number to

$$\binom{|\mathrm{var}(\mathscr{C} \cup c)|}{2} 2^{|\mathrm{var}(\mathscr{C}\cup c)|-2},$$

here 24, and then further to the number of *relevant* elementary CI statements up to symmetry, here only 8. Compared to the maximum possible size of $26 \times 80 = 2080$, our matrix has a size of only $7 \times 8 = 56$.

We have that $\mathbf{v}^T_{\mathscr{C}} = (1,1,1,1,1,2,2)$. Now, choose $k_c = 1$. Then, $\mathbf{v}^T_{\mathscr{C}} - k_c \mathbf{v}^T_c = (0,0,0,1,1,1,2)$. This vector is the sum of the 2nd, 3rd, and 8th column of $\mathbf{A}_{\mathscr{C}}$. Hence, $\mathbf{A}_{\mathscr{C}}\mathbf{x} = \mathbf{v}_{\mathscr{C}} - k_c \mathbf{v}_c$ has a nonnegative solution $\mathbf{x}^T = (0,1,1,0,0,0,0,1)$. By Proposition 50, $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c$, and also $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c_2$, $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c_3$, and $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c_8$.

It is well-known that a linear program (LP) is solvable in polynomial time in the number of its variables. As calculated in Example 51, we get in the worst case an LP with an exponential number of variables (matrix columns) and constraints (matrix rows), namely

$$\binom{|S|}{2} 2^{|S|-2} \quad \text{and} \quad 2^{|S|} - |S| - 1,$$

respectively. As a rule of thumb, the more columns matrix $\mathbf{A}_{\mathscr{C}}$ has, the more difficult the corresponding LP problem becomes. An advantage of our validation strategy over a naïve approach is that $\mathbf{A}_{\mathscr{C}}$ only consists of the vectors representing the *relevant* elementary CI statements in $\mathscr{R}/2(\mathscr{C})$. This means that the actual number of variables of the LP may be very small compared to the worst case. We also emphasize that matrix $\mathbf{A}$ is always a 0–1 matrix, leading to better numerical stability and the possibility to employ existing sparse matrix data structures.[5] We will come back to algorithmic issues when we discuss the results of our experiments.

Finally, consider again Proposition 50. Unfortunately, we cannot express the condition $k_c > 0$ in a linear program.[6] Therefore, we have to make a particular choice for k_c (e.g., $k_c = 1$ in Example 51). In the remainder of this section, we shall relate different choices for k_c. Thereto, let $CI(k_c)$ denote the set of all pairs $(\mathscr{C}, c)$ such that, for each elementary CI statement

[5] In rare cases, solutions to the LPs may be inaccurate due to round-off and truncation errors. Therefore, when we obtain a solution, we convert its components into fractional form (with integer numerator and denominator). For the resulting vector $\mathbf{x}_{\mathrm{frac}}$, we check whether $\mathbf{A}\mathbf{x}_{\mathrm{frac}} = \mathbf{b}$ holds *exactly*, and we only accept $\mathbf{x}$ as a solution if this is the case.

[6] Please note that the authors of the paper [9] have developed a method to circumvent this problem. It is potential future work to combine their approach with the present one.

$c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, there exists a nonnegative real number $k_{c_{\mathrm{el}}}$ for which $\mathbf{v}_{\mathscr{C}} = k_c\mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}}\mathbf{v}_{c_{\mathrm{el}}}$. We have the following monotonicity property.

Proposition 52. *If $k_2 \geqslant k_1 > 0$, then $CI(k_2) \subseteq CI(k_1)$.*

Proof. Let $(\mathscr{C}, c) \in CI(k_2)$. Then, for each $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, there exists $k_{c_{\mathrm{el}}} \geqslant 0$ such that $\mathbf{v}_{\mathscr{C}} = k_2\mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}}\mathbf{v}_{c_{\mathrm{el}}}$. By Proposition 50, $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c$, and hence, by Theorem 40, $\mathscr{C} \vdash c$. By Proposition 48, it follows that, for each $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, there exists $k'_{c_{\mathrm{el}}} \geqslant 0$ such that $\mathbf{v}_c = \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k'_{c_{\mathrm{el}}}\mathbf{v}_{c_{\mathrm{el}}}$. Combining all of this, we obtain

$$\begin{aligned}
\mathbf{v}_{\mathscr{C}} &= k_2\mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}}\mathbf{v}_{c_{\mathrm{el}}} \\
&= k_1\mathbf{v}_c + (k_2 - k_1)\mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} k_{c_{\mathrm{el}}}\mathbf{v}_{c_{\mathrm{el}}} \\
&= k_1\mathbf{v}_c + \sum_{c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})} \big((k_2 - k_1)k'_{c_{\mathrm{el}}} + k_{c_{\mathrm{el}}}\big)\mathbf{v}_{c_{\mathrm{el}}}.
\end{aligned}$$

Since $k_2 \geqslant k_1$, we have, for each $c_{\mathrm{el}} \in \mathscr{R}/2(\mathscr{C})$, that $(k_2 - k_1)k'_{c_{\mathrm{el}}} + k_{c_{\mathrm{el}}} \geqslant 0$. Consequently, $(\mathscr{C}, c) \in CI(k_1)$. □

We can actually show that some of the inclusions in Proposition 52 are strict.

Example 53. This is an adaptation of Example 6.3 in [6]. Let $S = \{a, b, d, e\}$, $\mathscr{C} = \{I(a, d|\emptyset), I(d, e|b), I(b, e|d), I(a, e|bd), I(b, d|ae)\}$, and $c = I(a, de|b)$. It can easily be verified that

$$\begin{aligned}
\mathbf{v}_{\mathscr{C}} = 0.5\mathbf{v}_c &+ 0.5\mathbf{v}_{I(a,d|\emptyset)} + 0.5\mathbf{v}_{I(a,e|\emptyset)} + 0.5\mathbf{v}_{I(d,e|b)} + 0.5\mathbf{v}_{I(b,e|d)} \\
&+ 0.5\mathbf{v}_{I(b,d|e)} + 0.5\mathbf{v}_{I(d,e|ab)} + 0.5\mathbf{v}_{I(b,e|ad)} + 0.5\mathbf{v}_{I(b,d|ae)},
\end{aligned}$$

and that the CI statements above are in $\mathscr{R}/2(\mathscr{C})$. Hence, there is a solution to the equation in Proposition 50 with $k_c = 0.5$. One can also verify that this equation has *no* solution for $k_c > 0.5$.

In Section 9, we shall further explore the actual impact of the choice of k_c.

We conclude this section with an example of an implication that holds and can neither be falsified nor validated by our algorithms.

Example 54. This is an adaptation of Example 4.1 in [6]. Let $S = \{a, b, d, e\}$, $\mathscr{C} = \{I(a, b|\emptyset), I(a, b|d), I(a, b|e), I(d, e|ab)\}$ and $c = I(a, b|de)$. Observe that $\mathscr{L}(c) \subseteq \mathscr{L}(\mathscr{C})$, so the lattice-exclusion criterion cannot be used to falsify this instance of the implication problem. Moreover, Studený [6] showed that $\mathscr{C} \models^m_{\mathscr{P}_{\mathrm{all}}} c$ although the equation in Proposition 50 has *no* solution for $k_c > 0$.

9. Experiments

The theory of the previous sections provides the formal foundation for implementations of practical falsification and validation algorithms for the probabilistic implication problem. With the experiments in the present section, we address the following empirical questions:

(1) **Effectiveness**: What fraction of the instances of the probabilistic CI implication problem can we either falsify or validate, i.e., *how close* do we get to a decision procedure?
(2) **Efficiency**: How fast does the algorithm run? To how many variables does it scale? How much more efficient is the algorithm compared to the naïve approach[7] both in terms of time and space complexity? How does the algorithm compare to other approaches in the literature?
(3) **Structural and numerical properties**: How large is the constraint matrix **A** for different instances? What are the numerical properties of the validation algorithm? To which extent does the parameter k_c of the validation algorithm influences the effectiveness and efficiency of the algorithm?

Since we may restrict our attention to elementary CI statements, we investigate the probabilistic implication problem $(\mathfrak{C}_{\mathrm{el}}, \mathfrak{C}_{\mathrm{el}})_{\mathscr{P}_{\mathrm{all}}}$, with $\mathfrak{C}_{\mathrm{el}}$ the class of all elementary CI statements. For each experiment, we first generated instances of this

[7] *Naïve approach* refers to the straightforward application of Corollary 45 without using the simplifications developed in Section 8.1 for falsification and without restricting the elementary CI statements to the relevant ones for validation.

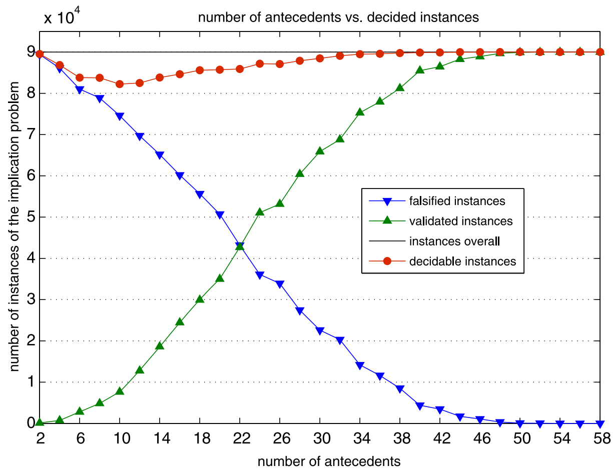

Fig. 9. Falsification and validation curves of the approximate decision algorithm for 5 variables. For each number of antecedents, 90 000 instances were considered (horizontal line). The number of decidable instances is the sum of the number of falsified instances and the number of validated instances.

Table 3
Mean and standard deviation of over 10 000 trials of the time to solve a linear program without and with optimizing the constraint matrix **A**.

Variables	Without optimizing **A**		With optimizing **A**	
	Mean	Standard deviation	Mean	Standard deviation
7	0.95 ms	0.39 ms	0.20 ms	0.33 ms
8	3.79 ms	1.83 ms	0.25 ms	0.49 ms
9	32.99 ms	51.28 ms	0.40 ms	0.77 ms
10	174.30 ms	907.53 ms	0.61 ms	0.54 ms

implication problem by randomly selecting n different sets of elementary CI statements $\mathscr{C}$ over S as antecedent, and, for each of these, m different elementary CI statements c over S as consequence, one at a time. We first applied the falsification algorithm to each of these instances "$\mathscr{C} \models^m_{\mathscr{P}_{\text{all}}} c$?". When the falsification algorithm failed, we applied the validation algorithm and took advantage of the results to create the constraint matrix $\mathbf{A}_{\mathscr{C}}$ and vector $\mathbf{b}$ from $\mathscr{C}$ and c as described in Section 8. To solve the resulting linear programs we used Gurobi,[8] a mixed integer linear programming solver that employs a version of the simplex algorithm [42,43] and the branch-and-cut method for integer programs. We did not change the standard optimization settings of the solver. Furthermore, we only accepted a solution if its rational expansion solved the respective constraints. For our purposes, this is entirely unproblematic, because the objective is to validate as many instances of the implication problem as possible while ruling out false positives. All experiments were run on a dual-core 3.2 GHz Linux PC with 3 GB RAM.

For a set of elementary CI statements $\mathscr{C}$, we build the optimized constraint matrix $\mathbf{A}_{\mathscr{C}}$ bottom-up using System $\mathscr{E}$. First, the closure under the *strong union* rule is computed. Thereafter, the closure under the *elementary contraction* and *strong contraction rules* is computed. Then, we again compute the closure under the *strong union* rule. Although this procedure might involve an exponential number of steps in the size of S, the naïve method of building the constraint matrix by using the set of *all* elementary CI statements over S is at least as complex.

Fig. 9 shows the number of instances of the probabilistic implication problem that could be falsified or validated by the algorithms for 5 variables. For each $\ell = 2, \ldots, 58$ (the number of antecedents), we randomly created $n = 4500$ different sets of ℓ elementary CI statements, and, for each of those, randomly selected $m = 20$ different elementary CI statements as consequences, one at a time, resulting in 90 000 instances for each value of ℓ. The results show that (1) only a small fraction of the instances could not be decided, and, (2) for larger values of ℓ (for 5 variables, $\ell > 40$), all instances could either be falsified or validated. The behavior of the algorithm was similar over all tested numbers of variables (see below).

Table 3 reveals the computational efficiency gained in optimizing the constraint matrix **A** by using only *relevant* elementary CI statements, compared to the naïve approach, where all elementary CI statements are used. The times provided are for 7–10 variables, averaged over 10 000 trials, for $n = 100$ sets of $\ell = 10$ antecedents, and $m = 100$ different consequences, one at a time. Table 4 shows the average time to construct and solve the linear program for sets of $\ell = 10$ antecedents for 5–15 variables, averaged over 100 000 trails ($n = 1000$, $m = 100$). The high standard deviations show that construction and solving times exhibit a high variance depending on the input. Table 5, finally, compares the size of the optimized constraint matrix to those of the constraint matrix resulting from the naïve approach.

[8] See http://www.gurobi.com.

Table 4
Mean and standard deviation over 100 000 trials of the time needed to construct, respectively, solve the linear program using only *relevant* elementary CI statements.

Variables	Construction time		Solving time	
	Mean	Standard deviation	Mean	Standard deviation
5	0.50 ms	0.65 ms	0.18 ms	0.10 ms
6	1.71 ms	2.19 ms	0.16 ms	0.15 ms
7	3.94 ms	0.51 ms	0.13 ms	0.14 ms
8	8.68 ms	1.48 ms	0.10 ms	0.19 ms
9	16.19 ms	3.71 ms	0.26 ms	0.32 ms
10	65.74 ms	9.11 ms	0.53 ms	0.70 ms
11	180.98 ms	26.18 ms	1.00 ms	1.28 ms
12	640.62 ms	54.04 ms	2.23 ms	2.48 ms
13	2150.81 ms	39.24 ms	6.09 ms	5.03 ms
14	6999.10 ms	471.01 ms	10.19 ms	9.81 ms
15	25 968.52 ms	497.67 ms	27.12 ms	21.08 ms

Table 5
Mean and standard deviation over 100 000 trials of the size of the optimized constraint matrix **A**, compared to the maximal possible values in parenthesis.

Variables	Number of rows		Number of columns	
	Mean	Standard deviation	Mean	Standard deviation
5	19.53 (26)	2.02	38.63 (80)	8.76
6	35.14 (57)	5.73	70.12 (240)	18.99
7	60.66 (120)	11.85	107.05 (672)	27.38
8	98.46 (247)	23.12	152.36 (1792)	44.09
9	146.72 (502)	41.66	188.32 (4608)	58.49
10	244.42 (1013)	68.01	319.61 (11 520)	95.17
11	355.02 (2036)	102.74	428.29 (28 160)	128.63
12	531.19 (4083)	176.39	618.84 (67 582)	205.68
13	823.41 (8178)	347.51	923.39 (159 744)	388.19
14	1154.84 (16 369)	672.25	1309.33 (372 736)	774.52
15	1930.60 (32 752)	960.26	2089.80 (860 160)	1046.63

To investigate the impact of the choice of k_c in Proposition 50, we conducted experiments for sets of $\ell = 10$, 20, and 30 antecedents, for 5–10 variables, averaged over 1 000 000 trials ($m = 1000$, $n = 1000$). Each of these was run with $k_c = 1.0$ and $k_c = 10^{-6}$. Despite the large difference between both values, at most 1176 additional instances out of a total of 1 000 000 (for 20 antecedents and 6 variables) could be validated. Hence, lowering k_c does *not* provide an added benefit, also because small values of k_c may lead to numerical instabilities.

We also want to empirically verify that (1) the lattice-exclusion criterion can falsify a large fraction of all falsifiable instances, and (2) Heuristics 1 and 2 are good approximations of the full-blown lattice-exclusion criterion. To make our results comparable to results from other approaches, we adopted the experimental setup for the *racing algorithm* of Bouckaert and Studený [8] (also using 5 attributes). One thousand sets of antecedents each were generated by randomly selecting 3 up to 10 elementary CI statements, resulting in a total of 8000 sets of antecedents. The *falsification algorithm* and the heuristics were run on these sets with each of the remaining elementary CI statements as a consequence, one at a time. This resulted in 77 000 instances of the probabilistic implication problem for sets with 3 antecedents, down to 70 000 instances for sets with 10 antecedents.

The rejection procedure of the *racing algorithm* is rooted in the theory of imsets: an instance is rejected if one of the constructed supermodular functions is a counter-model. It has double-exponential running time and may reject valid instances of the probabilistic CI implication problem. This is because the racing algorithm has the purpose of deciding the implication problem for imsets and, thus, the additive implication problem relative to the class of supermodular functions. The *falsification algorithm* based on Corollary 45 only rejects invalid instances of the CI implication problem. Fig. 10, left, shows the results. The falsification procedure of the racing algorithm rejects more instances than our approach, but among them are possibly also valid instances. Fig. 10, right, shows the rejection curves for the falsification algorithm, Heuristics 1 and 2, and the combination of both. The latter compares favorably with the falsification algorithm: 95% of the falsifications for 3 antecedents down to 77% for 10 antecedents are found. Also, Heuristic 2 is more effective than Heuristic 1.

Subsequent to the conference paper [25] in which we had first introduced the linear programming formulation of the validation algorithm based on Proposition 50, Bouckaert et al. [9] leveraged the theory of imsets [6] to also include the construction of linear programs to verify various instances of the CI implication problem. Their approach is almost identical to ours except that they do not optimize the constraint matrix **A** by considering only the relevant CI statements as columns and the elements in the semi-lattice of $\mathscr{C}$ as rows. In addition, the earlier version of our LP formulation did not take into account the parameter k_c of Proposition 50.

We adopted the same experimental setup as Bouckaert et al. [9] for an empirical comparison of the two linear programming formulations. For $|S| = 4$–10, we randomly generated 1000 sets each consisting of 3 elementary CI statements,

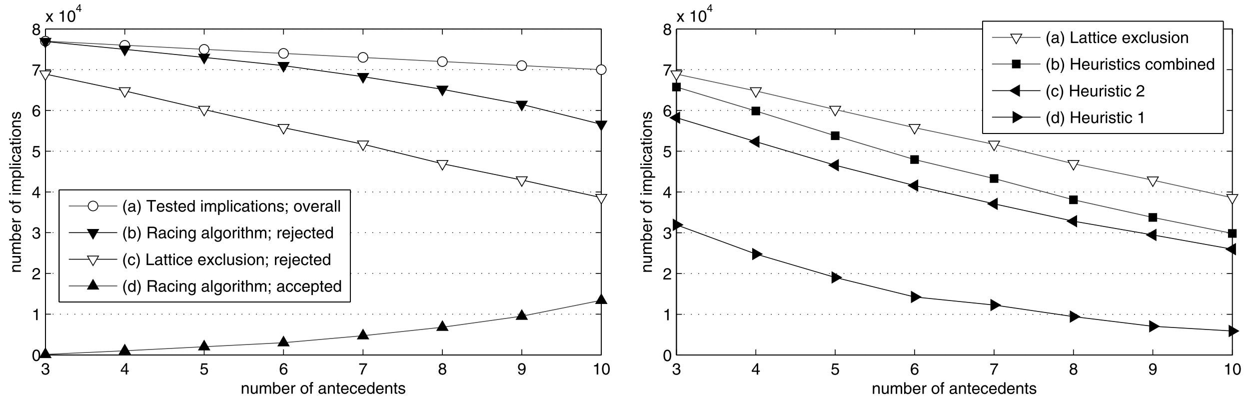

Fig. 10. Left: Rejection and acceptance curves of the racing and falsification algorithms. Right: Falsifications based on the lattice-exclusion criterion and the heuristics.

Table 6
Comparison of the LP formulation of Bouckaert et al. [9] and the one presented here.

$\|S\|$	4	5	6	7	8	9	10
Bouckaert et al. [9]							
LPs/experiment	24	80	240	672	1792	4608	11 520
Time/experiment	4 ms	18 ms	115 ms	658 ms	5037 ms	150.38 s	3862.1 s
Niepert et al., this paper							
LPs/experiment	**8.34**	**10.85**	**17.45**	**24.5**	**35.45**	**52.31**	**79.79**
Time/experiment	**1.03 ms**	**1.74 ms**	**2.68 ms**	**4.06 ms**	**7.18 ms**	**15.42 ms**	**43.01 ms**
Time to construct **A**	0.2 ms	0.44 ms	0.93 ms	1.36 ms	3.28 ms	8.1 ms	23.06 ms
Nr. rows **A**	6.06	8.84	14.56	21.36	32.31	49.02	75.18
Nr. columns **A**	8.34	10.85	17.45	24.5	35.45	52.31	79.79
Time to solve LP	0.1 ms	0.12 ms	0.1 ms	0.11 ms	0.11 ms	0.14 ms	0.25 ms
Overall nr. LPs	8342	10 854	17 449	24 501	35 450	52 306	79 790

then constructed the set of relevant CI statements and the constraint matrix **A** for each set of antecedents, and finally ran the validation algorithm with each of the possibly implied elementary CI statements as a consequence. Because we restrict ourselves to relevant CI statements, the number of LPs/experiment in Table 6 is much smaller for our approach than for the approach of Bouckaert et al. The table lists also the other statistics for the two LP formulations. The average time to construct the constraint matrix **A** increases from 0.2 ms for 4 variables to 23 ms for 10 variables. Since the size of the constraint matrix **A** remains small (on average about 80×80 for 10 variables), the average time to solve a single LP is only 0.25 ms for 10 variables which is to be compared with the average time of 3862.1 s/11520 = 300 ms for the constraint matrix consisting of *all* elementary CI statements used in Bouckaert et al.'s LP formulation, which is several orders of magnitudes slower than ours. Our procedure needs only 43.01 ms for 10 variables compared to the 3862.1 s of Bouckaert et al.'s approach.

10. Discussion and future work

Logical inference algorithms for probabilistic conditional independence statements have several important applications from checking consistency during knowledge elicitation to constraint-based structure learning of graphical models [44]. As it is generally known that there does not exist a finite, sound and complete axiomatization for probabilistic conditional independence implication [6], we can try to do the next best thing, which is enclosing the problem between a "lower bound", that is, a finite set of inference rules that is sound, and an "upper bound", a finite set of inference rules that is complete. It is well-known that the semi-graphoid axiom system, referred to in this paper as System $\mathcal{G}$ is sound (e.g., [40]). One of the main contributions of the present paper is introducing another finite set of inference rules, System $\mathcal{A}$, which is shown to be complete for the probabilistic implication problem. While a sound but not complete finite system of inference rules can always be transformed into a "tighter lower bound" by adding true logical implications that are not provable, an analogous strategy is not obvious for transforming a complete but not sound system of inference rules into a "tighter upper bound". It is therefore justified to ask whether System $\mathcal{A}$ is in some sense optimal as an "upper bound". This is work for future research. Of course, the above question could in principle be asked about any complete but not sound axiomatization system. It is therefore important to observe the connection between System $\mathcal{A}$ and the lattice-inclusion property, which plays a pivotal role in the present paper.

The theory underlying the present paper is based on a study of the additive implication problem, and, therefore, its scope of applications is much broader than only the probabilistic implication problem. Other application areas include database constraints, data mining, and reasoning about uncertainty [28,29,32,33]. Within the context of the probabilistic implication problem, the theory is also applicable to other types of (sets of) CI statements, such as stable sets of CI statements [5].

An important contribution of this paper are the falsification and validation heuristics for the probabilistic implication problem. While these heuristics are rooted in the lattice-theoretical framework, the supermodularity of the multi-information function also played a pivotal role in the development of the validation criterion. Therefore, a more thorough study of additive implication with respect to the class of supermodular functions is called for.

Finally, the lattice-inclusion property and the heuristics developed in this paper can be utilized to store information about conditional independencies more efficiently, using non-redundant representations [5]. This is in line with previous work on more efficient maintenance of conditional independence information [35,36]. Overall, we believe that we have convincingly argued that the lattice-theoretic framework for reasoning about conditional independence is a novel and powerful tool.

Acknowledgements

We thank Remco Bouckaert for providing us with the source code of the *racing algorithm*. We also want to thank Milan Studený for his continuous support of the present work and, in particular, the information he provided us concerning the complete axiomatization of the implication problem for four attributes. We also thank the anonymous reviewers of several conferences and workshops (UAI, PGM) and especially those of the *Artificial Intelligence* journal whose detailed comments and suggestions have helped to improve the clarity and quality of the article.

References

[1] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, Cambridge, MA, 2009.
[2] A. Dawid, Conditional independence in statistical theory, J. R. Stat. Soc. 41B (1979) 1–31.
[3] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., 1988.
[4] Y. Xiang, S.K.M. Wong, N. Cercone, Critical remarks on single link search in learning belief networks, in: Proceedings of the 12th Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 1996, pp. 564–571.
[5] M. Niepert, D. Van Gucht, M. Gyssens, Logical and algorithmic properties of stable conditional independence, Internat. J. Approx. Reason. 51 (5) (2010) 531–543.
[6] M. Studený, Probabilistic Conditional Independence Structures, Springer-Verlag, 2005.
[7] R. Hemmecke, J. Morton, A. Shiu, B. Sturmfels, O. Wienand, Three counterexamples on semigraphoids, Combin. Probab. Comput. 17 (2) (2008) 239–257.
[8] R. Bouckaert, M. Studený, Racing algorithms for conditional independence inference, Internat. J. Approx. Reason. 45 (2) (2007) 386–401.
[9] R. Bouckaert, R. Hemmecke, S. Lindner, M. Studený, Efficient algorithms for conditional independence inference, J. Mach. Learn. Res. 11 (2010) 3453–3479.
[10] S.M.M.S. Renooij, H.J.M. Tabachneck-Schijf (Eds.), Proceedings of the 6th UAI Bayesian Modelling Applications Workshop, 2008.
[11] J.Y. Halpern, Reasoning about Uncertainty, The MIT Press, 2003.
[12] M.J. Druzdzel, L.C. van der Gaag, Elicitation of probabilities for belief networks: Combining qualitative and quantitative information, in: UAI, 1995, pp. 141–148.
[13] R. Dechter, Constraint Processing, Morgan Kaufmann Publishers Inc., 2003.
[14] D. Geiger, J. Pearl, Logical and algorithmic properties of conditional independence and graphical models, Ann. Statist. 21 (4) (1993) 2001–2021.
[15] M. Studený, Conditional independence statements have no complete characterization, in: Transactions of the 11th Prague Conference on Information Theory, 1992, pp. 377–396.
[16] J. Pearl, A. Paz, Graphoids, graph-based logic for reasoning about relevance relations, Adv. Artif. Intell. II (1987) 357–363.
[17] A.P. Dawid, Separoids: A mathematical framework for conditional independence and irrelevance, Ann. Math. Artif. Intell. 32 (1–4) (2001) 335–372.
[18] G. Shafer, Advances in the understanding and use of conditional independence, Ann. Math. Artif. Intell. 21 (1997) 1–11.
[19] J.T.A. Koster, A graphical characterization of lattice conditional independence models, Ann. Math. Artif. Intell. 21 (1997) 13–26.
[20] S.A. Andersson, D. Madigan, M.D. Perlman, C.M. Triggs, Gibbs and Markov properties of graphs, Ann. Math. Artif. Intell. 21 (1997) 27–50.
[21] S.L. Lauritzen, F.V. Jensen, Local computation with valuations from a commutative semigroup, Ann. Math. Artif. Intell. 21 (1997) 51–69.
[22] M. Studený, Semigraphoids and structures of probabilistic conditional independence, Ann. Math. Artif. Intell. 21 (1997) 71–89.
[23] F. Matúš, Conditional independence structures examined via minors, Ann. Math. Artif. Intell. 21 (1997) 99–128.
[24] M. Studený, Structural imsets, an algebraic method for describing conditional independence structures, in: Proceedings of the Conference on Information Processing and Management of Uncertainty (IPMU), 2004, pp. 1323–1330.
[25] M. Niepert, Logical inference algorithms and matrix representations for probabilistic conditional independence, in: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, 2009, pp. 428–435.
[26] F.M. Malvestuto, A unique formal system for binary decompositions of database relations, probability distributions, and graphs, Inform. Sci. 59 (1992) 21–52.
[27] M. Niepert, D. Van Gucht, M. Gyssens, On the conditional independence implication problem: A lattice-theoretic approach, in: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, 2008, pp. 435–443.
[28] B. Sayrafi, A measure-theoretic framework for constraints and bounds on measurements of data, Ph.D. thesis, Indiana University, 2005.
[29] B. Sayrafi, D. Van Gucht, Inference systems derived from additive measures, in: Workshop on Causality and Causal Discovery, London, Canada, 2004.
[30] B. Sayrafi, D. Van Gucht, M. Gyssens, Measures in databases and data mining, Tech. Report TR602, Indiana University Computer Science, 2004.
[31] B. Sayrafi, D. Van Gucht, Differential constraints, in: Proceedings of the ACM Symposium on Principles of Database Systems, 2005, pp. 348–357.
[32] B. Sayrafi, D. Van Gucht, M. Gyssens, The implication problem for measure-based constraints, Inf. Syst. 33 (2) (2008) 221–239.
[33] B. Sayrafi, D. Van Gucht, P.W. Purdom, On the effectiveness and efficiency of computing bounds on the support of item-sets in the frequent item-sets mining problem, in: OSDM '05: Proceedings of the 1st International Workshop on Open Source Data Mining, 2005, pp. 46–55.
[34] F. Matúš, Ascending and descending conditional independence relations, in: Transactions of the 11th Prague Conference on Information Theory, 1992, pp. 189–200.
[35] P. de Waal, L.C. van der Gaag, Stable independence and complexity of representation, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 112–119.
[36] P. de Waal, L.C. van der Gaag, Stable independence in perfect maps, in: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, 2005, pp. 161–168.

[37] G.A. Grätzer, General Lattice Theory, second ed., Birkhäuser-Verlag, 1998.
[38] J.H. van Lint, R.M. Wilson, A Course in Combinatorics, second ed., Cambridge University Press, 2001.
[39] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86.
[40] J. Pearl, A. Paz, Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z?, in: Advances in Artificial Intelligence II, Seventh European Conference on Artificial Intelligence, 1986, pp. 357–363.
[41] M. Gyssens, M. Niepert, D. Van Gucht, About the completeness of the semigraphoid axioms for the derivation of arbitrary conditional independence statements from saturated ones, Tech. rep., Indiana University, 2011.
[42] A. Schrijver, Theory of Linear and Integer Programming, John Wiley & Sons, Inc., 1986.
[43] G. Dantzig, Linear Programming and Extensions, Princeton University Press, 1963.
[44] P. Gandhi, F. Bromberg, D. Margaritis, Learning Markov network structure using few independence tests, in: Proceedings of the SIAM International Conference on Data Mining, 2008, pp. 680–691.