Information Processing Letters 114 (2014) 628–633

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On the completeness of the semigraphoid axioms for deriving arbitrary from saturated conditional independence statements

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

[a] Hasselt University and Transnational Univ. of Limburg, Faculty of Sciences, Martelarenlaan 42, B-3500 Hasselt, Belgium
[b] Univ. of Washington, Dept. of Computer Science & Engineering, Paul G. Allen Center, AC101, 185 Stevens Way, Seattle, WA 98195-2350, USA
[c] Indiana University, Computer Science Division, Lindley Hall, 150 S Woodlawn Ave., Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Received 28 January 2014
Received in revised form 14 April 2014
Accepted 20 May 2014
Available online 24 May 2014
Communicated by Jef Wijsen

Keywords:
Conditional independence statement
Saturated conditional independence statement
Axiom system
Semigraphoid axioms
Soundness
Completeness
Databases

ABSTRACT

Conditional independence (CI) statements occur in several areas of computer science and artificial intelligence, e.g., as embedded multivalued dependencies in database theory, disjunctive association rules in data mining, and probabilistic CI statements in probability theory. Although, syntactically, such constraints can always be represented in the form $I(A, B|C)$, with A, B, and C subsets of some universe S, their semantics is very dependent on their interpretation, and, therefore, inference rules valid under one interpretation need not be valid under another. However, all aforementioned interpretations obey the so-called semigraphoid axioms. In this paper, we consider the restricted case of deriving arbitrary CI statements from so-called saturated ones, i.e., which involve all elements of S. Our main result is a necessary and sufficient condition under which the semigraphoid axioms are also complete for such derivations. Finally, we apply these results to the examples mentioned above to show that, for these semantics, the semigraphoid axioms are both sound and complete for the derivation of arbitrary CI statements from saturated ones.

1. Introduction

In numerous areas of computer science, the presence of constraints allows problems to be "decomposed" into simpler ones. Often, these constraints are *conditional independence* (CI) *statements* of the form $I(A, B|C)$, with A, B, and C pairwise disjoint subsets of some finite universe S [1].

In database theory, S could be a relation schema, and $I(A, B|C)$ an *embedded multivalued dependency* [2], meaning that the projection of the relation onto $A \cup B \cup C$ can be losslessly decomposed into its projections onto $A \cup C$ and $B \cup C$.

In data mining, S could be a set of items, and $I(A, B|C)$ a *disjunctive association rule* [3] meaning that C can only be contained in a basket if either A or B is contained in that basket.

In probability theory, S could be a set of variables, and $I(A, B|C)$ a *probabilistic CI statement* [4–6]. A probability distribution P satisfies $I(A, B|C)$ if A and B are independent conditional upon C. Because reasoning over the full joint probability distribution is almost always intractable, the presence of probabilistic CI statements may facilitate the decomposition of joint probability distributions into smaller parts which are then processed in sophisticated ways to compute a-posteriori probabilities.

Other examples identified by Studený [1] of areas in which conditional independence constraints arise, include the theory of ordinal conditional functions [7], the Dempster–Shafer theory of belief functions [8,9], and possibility theory [10].

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

http://dx.doi.org/10.1016/j.ipl.2014.05.010
0020-0190/

Sayrafi and two of the present authors introduced measure-based constraints [11,12], which encompass several of the above-cited examples (see also the work of Dalkilic and Robertson [13] and Lee [14]). A measure M is an increasing supermodular function associating non-negative real numbers[1] to the subsets of a universe S; M satisfies the *secondary constraint* $I(A, B|C)$ if $M(A \cup B \cup C) + M(C) = M(A \cup C) + M(B \cup C)$.

In view of this wide range of applications, a deeper theoretical understanding of the mathematical and algorithmic properties of conditional independence is required. Especially Studený [15] has brought this issue to the forefront, leading to an impressive body of work on algebraic representations of conditional independence structures with links to algebraic geometry [15,16], supermodular functions on sets, and algorithms for reasoning with conditional independencies [17–19]. A central notion in reasoning about conditional independence is the CI implication problem, i.e., deciding whether a set of CI statements implies a single CI statement relative to the semantics given to CI statements in the application under consideration. This was at the center of a study by the present authors and Sayrafi [19], in which soundness and completeness of axiom systems for CI implication was investigated.

As was already observed by Studený [1], soundness and completeness of such systems critically depends on the application under consideration. Also, it is known, e.g., that, for probabilistic conditional independence, no finite, sound and complete inference system exists [20] even though it remains open whether the probabilistic CI implication problem for the class of all discrete probability measures is decidable. For the case of embedded multivalued dependencies, it has even been proven that the implication problem is undecidable [21,22].

Despite all these differences, there is a finite inference system that is *sound* for all applications mentioned above, namely the *semigraphoid* axiom system [5], which is here referred to as System $\mathcal{G}$. The present authors and Sayrafi [19] investigated extensions of the System $\mathcal{G}$, and proposed a finite inference system referred to as System $\mathcal{A}$. It was shown that System $\mathcal{A}$, although not sound, is complete for the general probabilistic implication problem. What makes System $\mathcal{A}$ so attractive, also beyond the case of probabilistic conditional independence, is the existence of a set-theoretical characterization of derivability under System $\mathcal{A}$ [19], in terms of so-called meet semilattices [23].

The present authors and Sayrafi [19] have also looked to cases where some CI statements are *saturated*. Saturated CI statements involve all variables under consideration. In many cases, restricting the CI implication problem to *saturated* CI statements makes it not only decidable, but also axiomatizable, among other very desirable properties.[2] This is, e.g., the case for saturated embedded multivalued dependencies (called "multivalued dependencies" for short) and saturated probabilistic CI statements. For the latter case, it is well known that the semigraphoid System $\mathcal{G}$ is sound and complete for the derivation of saturated CI statements from saturated CI statements. Unfortunately, System $\mathcal{G}$ also allows for the derivation of unsaturated CI statements from saturated ones. To circumvent this caveat, Malvestuto [26] and Geiger and Pearl [27] proposed an alternative sound and complete set of inference rules for this purpose, referred to as System $\mathcal{S}$, and which is subsumed by the semigraphoid rules, but which has the additional advantage that all the intermediate CI statements derived are also saturated. The present authors and Sayrafi [19] were able to generalize this result by showing that it follows from their theoretical framework that System $\mathcal{A}$ is sound and complete for the derivation of *arbitrary* CI statements from saturated ones.

Since in all applications mentioned above, System $\mathcal{G}$ is sound for the CI implication problem—and hence in particular for the derivation of arbitrary CI statements from saturated ones, the above result begs the question under which conditions System $\mathcal{G}$ is also *complete* for the derivation of arbitrary CI statements from saturated ones. It is shown that a necessary and sufficient condition for this to be the case is that System $\mathcal{A}$ is complete for the derivation of arbitrary CI statements from saturated ones, *and this regardless of the context in which the CI statements are interpreted*. In addition, Theorem 1 and Corollary 1 state an analogous characterization for soundness. These results hinge on a "normal form" result for such derivations (Proposition 2) which says that whenever an arbitrary CI statement can be derived from saturated ones under System $\mathcal{A}$, a "saturated version" of that CI statement can be derived under System $\mathcal{S}$. We show that, for our initial examples, the results of this paper yield soundness and completeness of both Systems $\mathcal{A}$ and $\mathcal{G}$ for deriving arbitrary CI statements from saturated ones (Proposition 3).

2. Preliminaries

Throughout the paper, we consider a finite universe S, which we shall often leave implicit. 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. For $A \subseteq S$, we write $\overline{A}$ for $S - A$, the complement of A with respect to S.

We begin by defining conditional independence (CI) statements as a purely syntactic notion, without being concerned with the semantics.

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*.

Example 1. Let $S = abcdefgh$ be the universe. Then $I(adef, bgh|c)$ and $I(a, b|c)$ are both examples of CI statements. The former is saturated, whereas the latter is not.

[1] $M : 2^S \to \mathbb{R}^{\geq 0}$ is *increasing* if $M(A \cup B) \geq M(A)$ and *supermodular* if $M(A \cup B \cup C) + M(C) \geq M(A \cup C) + M(B \cup C)$. By swapping the inequalities, decreasing supermodular, increasing submodular, and decreasing supermodular measures are obtained, which lead to essentially the same theory.

[2] See [24,25] for examples of very recent work in this area.