On the Practical Irrelevance of Diverging Implication between Probabilistic Conditional Independence and Embedded Multivalued Dependency

C.J. Butz[1] and P. Lingras[2]

[1] Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: butz@cs.uregina.ca

[2] Department of Math and Computer Science, Saint Mary's University
Halifax, Nova Scotia, Canada, B3H 3C3.
E-mail: Pawan.Lingras@stmarys.ca

Abstract. Bayesian networks serve as the basis for developing probabilistic expert systems and have been applied widely in artificial intelligence. Previous research has argued that Bayesian networks and relational databases are *different* by showing that the logical implication of *conditional independence* (CI) and *embedded multivalued dependency* (EMVD) do not always coincide. In this paper, we show that this theoretical difference has no practical impact when designing probabilistic expert systems. Therefore, this work adds to the mounting evidence clearly indicating that the implementation of probabilistic expert systems can take advantage of conventional relational database management systems.

1 Introduction

Bayesian networks [3,9,10,12,19,20,26] are an established framework for uncertainty management in artificial intelligence. For instance, they have been applied in building intelligent agents, such as Office Assistant, and adaptive user interfaces by Microsoft [5,7,13], process control by NASA [6,8] and Lockheed [14], software diagnosis by Hewlett Packard [2,16] and Nokia [11], and medical diagnosis such as the Heart Disease Program at the Massachusetts Institute of Technology [15] and the Pathfinder Project for lymph-node diseases at Stanford University [4]. A Bayesian network consists of a *directed acyclic graph* (DAG) and a corresponding set of *conditional probability tables* (CPTs). By the *probabilistic conditional independencies* (CIs) [24] encoded in the DAG, the product of the CPTs is a joint probability distribution. Thereby, Bayesian networks provide a clear semantic modelling tool, which facilitate the acquisition of probabilistic knowledge.

Before Bayesian networks were proposed, the *relational database model* [1,18] already established itself as the basis for designing and implementing database

systems. Data dependencies [1], such as *embedded multivalued dependency* (EMVD), (nonembedded) *multivalued dependency* (MVD) and *join dependency* (JD), are used to provide an economical representation of a universal relation. As in the study of Bayesian networks, two of the most important results are the ability to specify a universal relation as a *lossless* join of several smaller relations, and the development of efficient methods to only access the relevant portions of the database in query processing.

Studeny [22] has argued that Bayesian networks and relational databases are *different* by showing that the logical implication of CI does not always coincide with that of EMVD. In [24], we responded by pointing out that Studeny's counter-example is based on classes of dependencies without a complete axiomatization. In the design of probabilistic expert systems and relational databases, usually only classes of dependencies with complete axiomatizations are considered. Thus, we concluded in [24] that Studeny's point was moot. However, one may still believe that there still exists a practical difference, if one considers situations like those given in [22].

In this paper, our objective is to show that there is *no* practical consequence even if one considers the theoretical difference given in [22]. We establish this result by formalizing the transformation between traditional relations and probabilistic relations with the notion of *obtainable*. Moreover, Lee [17] has proven that EMVD is a necessary condition for CI. We use these two concepts to show that the only probabilistic relations, considered in practice when designing a probabilistic expert system, must necessarily be obtained from traditional relations agreeing with the logical implication of CI. Hence, this work is yet another example [23–25] emphasizing the intrinsic relationship between Bayesian networks and relational databases.

This paper is organized as follows. Section 2 reviews background knowledge. Relationships between traditonal and probabilistic relations, as well as EMVD and CI, are discussed in Section 3. In Section 4, we establish our main result, namely, the theoretical difference in [22] has no consequence in practice. Conclusions are made in Section 5.

2 Background Knowledge

In this section, we review EMVD, CI, and the logical implication of CI and EMVD.

2.1 Embedded Multivalued Dependency

Here we give a quick introduction to EMVD, subclasses of which play an important role in relational database schema design [1,18].

---

[1] Constraints are traditionally called *dependencies* in relational databases, but are referred to as *independencies* in Bayesian networks. Henceforth, we will use the terms *dependency* and *independency* interchangeably.

A *relation scheme* $R = \{A_1, A_2, \ldots, A_m\}$ is a finite set of *attributes* (attribute names). Corresponding to each attribute A_i is a nonempty finite set $dom(A_i)$, $1 \le i \le m$, called the *domain* of A_i. Let $D = dom(A_1) \cup dom(A_2) \ldots \cup dom(A_m)$. A *relation* r on the relation scheme R, written $r(R)$, is a finite set of mappings $\{t_1, t_2, \ldots, t_s\}$ from R to D with the restriction that for each mapping $t \in r$, $t(A_i)$ must be in $dom(A_i)$, $1 \le i \le m$, where $t(A_i)$ denotes the value obtained by restricting the mapping to A_i. The mappings are called *tuples* and $t(A)$ is called the A-value of t. We use $t(X)$ in the obvious way and call it the X-value of the tuple t, where $X \subseteq R$ is an arbitrary set of attributes.

Example 1. Consider four binary attributes A, B, C and D. One traditonal relation $r(ABCD)$ is shown on the left of Fig. 1.

$$r(ABCD) = \begin{array}{|c c c c|}\hline A & B & C & D \\\hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\\hline\end{array} \quad \pi_{ABC}(r) = \begin{array}{|c c c|}\hline A & B & C \\\hline 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\\hline\end{array} = \begin{array}{|c c|}\hline A & B \\\hline 0 & 0 \\ 1 & 0 \\ 1 & 1 \\\hline\end{array} \bowtie \begin{array}{|c c|}\hline B & C \\\hline 0 & 0 \\ 0 & 1 \\ 1 & 1 \\\hline\end{array}$$

Fig. 1. The traditional relation $r(ABCD)$ satisfies the EMVD $B \rightarrow\rightarrow A|C$, since $\pi_{ABC}(r) = \pi_{AB}(r) \bowtie \pi_{BC}(r)$.

Let r be a relation on R and X a subset of R. The *projection of r onto X*, written $\pi_X(r)$, is defined as:

$$\pi_X(r) = \{\ t(X) \mid t \in r\ \}. \tag{1}$$

The *natural join* of two relations $r_1(X)$ and $r_2(Y)$, written $r_1(X) \bowtie r_2(Y)$, is defined as:

$$r_1(X) \bowtie r_2(Y) = \{\ t(XY) \mid t(X) \in r_1(X) \text{ and } t(Y) \in r_2(Y)\ \}. \tag{2}$$

Let X, Y, Z be disjoint subsets of attributes in R. We say relation $r(R)$ satisfies the *embedded multivalued dependency* (EMVD) $X \rightarrow\rightarrow Y|Z$, if the projection $\pi_{XYZ}(r)$ of $r(R)$ satisfies the condition:

$$\pi_{XYZ}(r) = \pi_{XY}(r) \bowtie \pi_{XZ}(r).$$

Example 2. The traditonal relation $r(ABCD)$ on the left of Fig. 1 satisfies the EMVD $B \rightarrow\rightarrow A|C$, since $\pi_{ABC}(r) = \pi_{AB}(r) \bowtie \pi_{BC}(r)$.

2.2 Probabilistic Conditional Independence

Here we briefly review the notion of probabilistic conditional independence, sub-classes of which play an instrumental role in designing the schema of a probabilistic network [19]. The reader should note that we express probabilistic conditional independence in terms of our probabilistic database model [23–25] which serves as a unified approach for both Bayesian networks and relational databases.

Let $R = \{v_1, v_2, \ldots, v_m\}$ be a finite set of variables. Each variable v_i has a finite domain, denoted $dom(v_i)$, representing the values that v_i can take on. For a subset $X = \{v_i, \ldots, v_j\}$ of R, we write $dom(X)$ for the Cartesian product of the domains of the individual variables in X, namely, $dom(X) = dom(v_i) \times \ldots \times dom(v_j)$. Each element $x \in dom(X)$ is called a *configuration* of X.

A *joint probability distribution* [20] on $dom(R)$ is a function p on $dom(R)$ such that the following two conditions both hold: (i) $0 \leq p(t) \leq 1$, for each configuration $t \in dom(R)$, and (ii) $\sum_{t \in dom(R)} p(t) = 1.0$. A *potential* on $dom(R)$ is a function ϕ on $dom(R)$ such that the following two conditions both hold: (i) $0 \leq \phi(t)$, for each configuration $t \in dom(R)$, and (ii) $\phi(t) > 0$, for at least one configuration $t \in dom(R)$. For brevity, we refer to ϕ as a potential on R rather than $dom(R)$, and we call R, not $dom(R)$, its domain [20].

We now introduce the fundamental notion of *probabilistic conditional independency* (CI) [24]. Let X, Y and Z be disjoint subsets of variables in R. Let x, y, and z denote arbitrary values of X, Y and Z, respectively. We say Y and Z are *conditionally independent* given X under the joint probability distribution p, denoted $I_p(Y, X, Z)$, if

$$p(y \mid x, z) = p(y \mid x), \tag{3}$$

whenever $p(x, z) > 0$. This conditional independency $I_p(Y, X, Z)$ can be equivalently written as

$$p(y, x, z) = \frac{p(y, x) \cdot p(x, z)}{p(x)}. \tag{4}$$

We write $I_p(Y, X, Z)$ as $I(Y, X, Z)$ if the joint probability distribution p is understood.

The above notions of probability tables and probabilistic conditional independence can be conveniently expressed in our probabilistic relational database model. For our purposes here we simply illustrate the main ideas and refer to the reader to [23–25] for a more thorough discussion of our model.

A potential $\phi(R)$ can be represented as a *probabilistic* relation $\mathbf{r}(R, A_\phi)$, where the column labeled by A_ϕ stores the probability value. The relation $\mathbf{r}(A_1, A_2, \ldots, A_m, A_\phi)$ representing a potential $\phi(A_1, A_2, \ldots, A_m)$ contains tuples of the form $\mathbf{t} = \ <t, \phi(t)>$. Note that $\mathbf{r}(A_1, A_2, \ldots, A_m, A_\phi)$ only contains tuples with *positive* probability. For convenience we will write $\mathbf{r}(R, A_\phi)$ as $\mathbf{r}(R)$ and say relation $\mathbf{r}$ is on R with the attribute A_ϕ understood by context. That is, relations denoted by boldface represent probability distributions.

Example 3. One probabilistic relation $\mathbf{r}(ABCD)$ is depicted in the top left of Fig. 2.

$$\mathbf{r}(ABCD) \;=\; \begin{array}{|cccc|c|} \hline A & B & C & D & A_{p(ABCD)} \\ \hline 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 1 & 0.1 \\ 0 & 0 & 1 & 1 & 0.2 \\ 1 & 0 & 0 & 0 & 0.1 \\ 1 & 0 & 1 & 0 & 0.1 \\ 1 & 1 & 1 & 1 & 0.4 \\ \hline \end{array} \qquad \tau_{ABC}(\mathbf{r}) \;=\; \begin{array}{|ccc|c|} \hline A & B & C & A_{p(ABC)} \\ \hline 0 & 0 & 0 & 0.2 \\ 0 & 0 & 1 & 0.2 \\ 1 & 0 & 0 & 0.1 \\ 1 & 0 & 1 & 0.1 \\ 1 & 1 & 1 & 0.4 \\ \hline \end{array}$$

$$= \begin{array}{|cc|c|} \hline A & B & A_{p(AB)} \\ \hline 0 & 0 & 0.4 \\ 1 & 0 & 0.2 \\ 1 & 1 & 0.4 \\ \hline \end{array} \otimes \begin{array}{|cc|c|} \hline B & C & A_{p(BC)} \\ \hline 0 & 0 & 0.3 \\ 0 & 1 & 0.3 \\ 1 & 1 & 0.4 \\ \hline \end{array} = \begin{array}{|ccc|c|} \hline A & B & C & A_{\frac{p(AB)p(BC)}{p(B)}} \\ \hline 0 & 0 & 0 & (0.4)(0.3)/(0.6) = 0.2 \\ 0 & 0 & 1 & (0.4)(0.3)/(0.6) = 0.2 \\ 1 & 0 & 0 & (0.2)(0.3)/(0.6) = 0.1 \\ 1 & 0 & 1 & (0.2)(0.3)/(0.6) = 0.1 \\ 1 & 1 & 1 & (0.4)(0.4)/(0.4) = 0.4 \\ \hline \end{array}$$

Fig. 2. The probabilistic relation $\mathbf{r}(ABCD)$ satisfies the CI $B \Rightarrow\Rightarrow A|C$, since $\tau_{ABC}(\mathbf{r}) = \tau_{AB}(\mathbf{r}) \otimes \tau_{BC}(\mathbf{r})$.

The *projection* operator π in relational databases corresponds to the *marginalization* operator τ in probabilistic databases. Whereas π ignores duplicate tuples, τ adds them.

Example 4. Given the probabilistic relation $\mathbf{r}(ABCD)$ in the top left of Fig. 2, the marginalization $\tau_{ABC}(\mathbf{r})$ of $\mathbf{r}(ABCD)$ onto variables ABC is shown in the top right of Fig. 2.

Probabilistic conditional independency is now defined in our probabilistic database model. A probabilistic relation $\mathbf{r}(XYZW)$ satisfies the CI $X \Rightarrow\Rightarrow Y|Z$, if

$$\tau_{XYZ}(\mathbf{r}) = \tau_{XY}(\mathbf{r}) \otimes \tau_{XZ}(\mathbf{r}), \qquad (5)$$

where the operator $\otimes$ corresponds to the right side of Equation (4).

Example 5. The probabilistic relation $\mathbf{r}(ABCD)$ on the top left of Fig. 2 satisfies the CI $B \Rightarrow\Rightarrow A|C$, since the marginal $\tau_{ABC}(\mathbf{r})$ can be written as $\tau_{ABC}(\mathbf{r}) = \tau_{AB}(\mathbf{r}) \otimes \tau_{BC}(\mathbf{r})$.

2.3 Logical Implication of CI and EMVD

Before we study the implication problem in detail, let us first introduce some basic notions. Here we will use the terms *relation* and *joint probability distribution* interchangeably.

Let $\sum$ be a set of dependencies defined on a set of attributes R. By $SAT_R(\sum)$, we denote the set of all relations on R that satisfy all of the dependencies in $\sum$. We write $SAT_R(\sum)$ as $SAT(\sum)$ when R is understood, and $SAT(\sigma)$ for $SAT(\{\sigma\})$, where σ is a single dependency. We say $\sum$ *logically implies* σ, written $\sum \models \sigma$, if $SAT(\sum) \subseteq SAT(\sigma)$. In other words, σ is logically implied by $\sum$, if every relation which satisfies $\sum$ also satisfies σ. That is, there is no counter-example relation such that all of the dependencies in $\sum$ are satisfied but σ is not.

The *implication problem* is to test whether a given set $\sum$ of dependencies logically implies another dependency σ, namely,

$$\sum \models \sigma. \tag{6}$$

Since we advocate that our probabilistic database model is a *generalization* of the relational database model, an immediate question to answer is:

Do the implication problems coincide in these two database models?

That is, we would like to know whether the following proposition holds:

$$\mathbf{C} \models \mathbf{c} \Longleftrightarrow C \models c, \tag{7}$$

given sets $\mathbf{C}$ and $\mathbf{c}$ of CIs and *corresponding* sets C and c of EMVDs.

Studeny [22] argued that the Bayesian network community could not take advantage of work in the relational database community by showing that Equation (7) does not always hold. In particular, he presented the following example violating Equation (7).

Example 6. Consider the following sets $\mathbf{C}$ and $\mathbf{c}$ of CIs, and corresponding sets C and c of EMVDs:

$$\mathbf{C} = \{A_3 A_4 \Rightarrow\Rightarrow A_1 | A_2, \ A_1 \Rightarrow\Rightarrow A_3 | A_4, \ A_2 \Rightarrow\Rightarrow A_3 | A_4, \ \emptyset \Rightarrow\Rightarrow A_1 | A_2\},$$
$$C = \{A_3 A_4 \rightarrow\rightarrow A_1 | A_2, \ A_1 \rightarrow\rightarrow A_3 | A_4, \ A_2 \rightarrow\rightarrow A_3 | A_4, \ \emptyset \rightarrow\rightarrow A_1 | A_2\},$$
$$\mathbf{c} = \emptyset \Rightarrow\Rightarrow A_3 | A_4,$$
$$c = \emptyset \rightarrow\rightarrow A_3 | A_4.$$

Then

$$\mathbf{C} \models \mathbf{c}, \tag{8}$$

yet

$$C \not\models c, \tag{9}$$

Studeny proved Equation (8) in [21]. To show Equation (9), Studeny [22] presented the relation $r(A_1 A_2 A_3 A_4)$ in Fig. 3. It can be verified that relation $r(A_1 A_2 A_3 A_4)$ satisfies all the EMVDs in C but does not satisfy the EMVD c.

$$r(A_1A_2A_3A_4) \;=\; \begin{array}{|cccc|}
\hline
A_1 & A_2 & A_3 & A_4 \\
\hline
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
\hline
\end{array}$$

Fig. 3. $C \not\models c$ as the traditional relation r satisfies all of the EMVDs in C but does not satisfy the EMVD c, where C and c are given in Example 6.

In [24], we pointed out that this counter-example should not be considered as there is no finite complete axiomatization for CI in general. Neveretheless, even if one decides to design a probabilistic expert system using input sets **C** like these, we can still show, in the remainder of this paper, that this theoretical difference has no bearing on the design of probabilistic expert systems in practice.

3 Obtainable Probabilistic and Traditional Relations

In this section, we first define a method, called *obtainable*, for transforming between relations and probabilistic relations. We then recall a known result discussing a relationship between EMVD and CI.

Given a probabilistic relation $\mathbf{r}(R)$, the unique traditional relation $r(R)$ *obtainable* from $\mathbf{r}(R)$ is defined as:

$$r(R) = \pi_R(\mathbf{r}). \tag{10}$$

Example 7. Given the probabilistic relation $\mathbf{r}(R)$ in Fig. 4, the traditional relation $r(R)$ obtainable from $\mathbf{r}(R)$ is shown in Fig. 3.

$$\mathbf{r}(A_1A_2A_3A_4) \;=\; \begin{array}{|ccccc|}
\hline
A_1 & A_2 & A_3 & A_4 & A_p \\
\hline
0 & 0 & 0 & 0 & 0.10 \\
0 & 0 & 0 & 1 & 0.10 \\
0 & 1 & 0 & 0 & 0.20 \\
1 & 0 & 0 & 0 & 0.20 \\
1 & 1 & 0 & 0 & 0.20 \\
1 & 1 & 1 & 0 & 0.20 \\
\hline
\end{array}$$

Fig. 4. For Example 6, a probabilistic relation $\mathbf{r}(R)$ not satisfying **C**. The traditional relation $r(R)$ obtained from $\mathbf{r}(R)$ satisfies the corresponding C, but not c, of EMVDs.

Given a traditional relation $r(R)$, a probabilistic relation $\mathbf{r}(R)$ is *obtainable* from $r(R)$, if the following two conditions hold:

(i) $r(R) = \pi_R(\mathbf{r})$, and
(ii) the probability values of $\mathbf{r}(R)$ sum to one.

Example 8. Given the traditional relation $r(R)$ in Fig. 3, one obtainable probabilistic relation $\mathbf{r}(R)$ is depicted in Fig. 4. On the contrary, neither of the probabilistic relations in Fig. 5 are obtainable from $r(R)$ in Fig. 3, since condition (i) is violated in both cases.

A_1	A_2	A_3	A_4	A_p	A_1	A_2	A_3	A_4	A_p
0	0	0	0	0.10	0	0	0	0	0.10
0	0	0	1	0.10	0	0	0	1	0.10
0	1	0	0	0.20	0	1	0	0	0.20
1	0	0	0	0.20	1	0	0	0	0.20
1	1	0	0	0.20	1	1	0	0	0.40
1	1	1	0	0.10					
1	1	1	1	0.10					

Fig. 5. Neither of these two probabilistic relations are obtainable from the traditional relation $r(R)$ in Fig. 3.

The next result discusses a relationship between EMVD and CI.

Lemma 9. [17] If a probabilistic relation $\mathbf{r}(R)$ satisfies a CI $X \Rightarrow\Rightarrow Y|Z$, then the traditional relation $r(R)$ obtainable from $\mathbf{r}(R)$ necessarily satisfies the EMVD $X \rightarrow\rightarrow Y|Z$.

Lemma 9 indicates that EMVD is a necessary condition for CI.

4 Practical Irrelevance of Diverging Implication

In this section, we emphasize the intrinsic relationship between Bayesian networks and relational databases by showing that the theoretical difference of

$$\mathbf{C} \models \mathbf{c}$$

and

$$C \not\models c$$

has *no* practical consequence.

Consider any instance of this theoretical difference. All traditional relations $r(R)$ on R can be partitioned into two classes:

(1) those agreeing with $C \models c$,
(2) those disagreeing with $C \models c$.

The next result shows that $\mathbf{r}(R)$ does not satisfy $\mathbf{C}$, for any probabilistic relation $\mathbf{r}(R)$ obtainable from any traditional relation $r(R)$ in class (2).

Theorem 10. Suppose $\mathbf{C} \models \mathbf{c}$ and $C \not\models c$ for some instantiation of the implication problem. For any traditional relation $r(R)$ satisfying C but not c, consider any probabilistic relation $\mathbf{r}(R)$ obtainable from $r(R)$. Then $\mathbf{r}(R) \notin SAT_R(\mathbf{C})$.

Proof. By contradiction, suppose a probabilistic relation $\mathbf{r}(R)$ satisfies all the CIs in $\mathbf{C}$, where $\mathbf{r}(R)$ was obtained from a traditional relation $r(R)$ satisfying C but not c. Since $\mathbf{C} \models \mathbf{c}$, $\mathbf{r}(R)$ also satisfies $\mathbf{c}$. By Lemma 9, as $\mathbf{r}(R)$ was obtained from $r(R)$, the traditional relation $r(R)$ satisfies both C and c. A contradiction. Therefore, all probabilstic relations $\mathbf{r}(R)$ satisfying $\mathbf{C}$ must be obtainable from traditional relations $r(R)$ satisfying both C and c.

Theorem 10 is important since it indicates that the only probabilistic relations $\mathbf{r}(R)$ satisfying $\mathbf{C}$ must be obtained from those traditional relations $r(R)$ in class (1), i.e., those $r(R)$ satisfying both C and c.

Example 11. The probabilistic relation $\mathbf{r}(R)$ illustrated in Fig. 6 satisfies the given CIs $\mathbf{C}$ in Example 6. In addition, it can be verified that the traditional relation $r(R)$ obtained from $\mathbf{r}(R)$ satisfies the EMVDs C and c in Example 6.

$$\mathbf{r}(A_1 A_2 A_3 A_4) \quad = \quad$$

A_1	A_2	A_3	A_4	A_p
0	0	0	0	0.0625
0	0	0	1	0.0625
0	0	1	0	0.0625
0	0	1	1	0.0625
0	1	0	0	0.0625
0	1	0	1	0.0625
0	1	1	0	0.0625
0	1	1	1	0.0625
1	0	0	0	0.0625
1	0	0	1	0.0625
1	0	1	0	0.0625
1	0	1	1	0.0625
1	1	0	0	0.0625
1	1	0	1	0.0625
1	1	1	0	0.0625
1	1	1	1	0.0625

Fig. 6. For Example 6, a probabilistic relation $\mathbf{r}(R)$ satisfying $\mathbf{C}$. The traditional relation $r(R)$ obtained from $\mathbf{r}(R)$ satisfies the corresponding C and c of EMVDs.

On the other hand, now consider those probabilistic relations obtainable from traditional relations in class (2).

Example 12. Recall the tradition relation $r(R)$ in Fig. 3 that satisfies C but not c. One probabilistic relation $\mathbf{r}(R)$, obtainable from $r(R)$, is illustrated in Fig. 4. It can be verified that $\mathbf{r}(R)$ does *not* satisfy all CIs in $\mathbf{C}$. By definition,

$$\mathbf{r}(R) \notin SAT(\mathbf{C}). \tag{11}$$

Thus, the probabilistic relation $\mathbf{r}(R)$ in Fig. 4 is not of interest when designing a probabilistic expert system for the CIs $\mathbf{C}$.

Our goal is to design the schema of a probabilistic expert system for a given set $\mathbf{C}$ of CIs. Studeny [22] has suggested that work on database design is of no value to the Bayesian network community, since it may be the case that

$$\mathbf{C} \models \mathbf{c}, \tag{12}$$

while for the corresponding EMVDs

$$C \not\models c. \tag{13}$$

The key point, as illustrated in Fig. 6, is that whenever a probabilistic relation satisfies $\mathbf{C}$, the unique obtainable traditional relation $r(R)$ must satisfy the corresponding EMVDs C and c. On the contrary, as depicted in Fig. 4, $\mathbf{r}(R)$ will not satisfy $\mathbf{C}$, for any probabilistic relation $\mathbf{r}(R)$ obtainable from a traditional relation $r(R)$ satisfying C but not c. Since $\mathbf{r}(R) \notin SAT_R(\mathbf{C})$, it is of no interest when designing the probabilistic expert system. Our main result, Theorem 10, ensures that the same result holds for all instances of Equations (12) and (13).

5 Conclusions

In schema design of a probabilistic network on variables R, only probability distributions $\mathbf{r}(R)$ satisfying the given probabilistic conditional independencies $\mathbf{C}$ are of interest, that is, those $\mathbf{r}(R) \in SAT_R(\mathbf{C})$. Similarly, in schema design of a relational database on attributes R, only traditional relations $r(R)$ satisfying the given EMVDs C are of interest. It has been argued [22] that Bayesian networks are *different* from relational databases by showing an instance where $\mathbf{C} \models \mathbf{c}$ and $C \not\models c$, for corresponding sets of CIs and EMVDs. On the contrary, our main result (Theorem 10) indicates that $\mathbf{r}(R) \notin SAT_R(\mathbf{C})$, for any probabilistic relation obtained from any traditional relational $r(R)$ satisfying C but not c. Therefore, the fact that $C \not\models c$ has no practical consequence on the design of a probabilistic network satisfying CIs $\mathbf{C}$. The work here is then yet one more example [23–25] emphasizing the intrinsic relationship between Bayesian networks and relational databases.

References

1. Abiteboul, S., Hull, R. and Vianu, V.: Foundations of Databases. Addison-Wesley, Don Mills (1995)

2. Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleyner, G., Mueller, M. and Singhal, S.: http://www.hpl.hp.com/techreports/2001/HPL-2001-23R1.pdf, April 25, 2005
3. Castillo, E., Gutiérrez, J. Hadi, A.: Expert Systems and Probabilistic Network Models. Springer, New York (1997)
4. Heckerman, D., Horvitz, E. and Nathwani, B.: Towards normative expert systems: Part I the Pathfinder project. Methods of Information in Medicine **31**(2) (1992) 90–105
5. Horvitz, E.: Agents with Beliefs: Reflections on Bayesian Methods for User Modelling. In: Proceedings of the Sixth International Conference on User Modelling (1997) 441–442
6. Horvitz, E. and Barry, E.M.: Display of Information for Time Critical Decision Making. In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco (1995) 296-305
7. Horvitz, E., Breese, J., Heckerman, D., Hovel, D. and Rommelse, K.: The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. Madison, WI (1998) 256-265
8. Horvitz, E., Srinivas, S., Rouokangas, C. and Barry, M.: A decision-theoretic approach to the display of information for time-critical decisions: The Vista project. In: Proceedings of SOAR-92 Conference on Space Operations Automation and Research, National Aeronautics and Space Administration (1992)
9. Jensen, F.V.: An Introduction to Bayesian Networks. UCL Press, London (1996)
10. Jensen, F.V., Lauritzen, S.L. and Olesen, K.G.: Bayesian updating in causal probabilistic networks by local computations. Comput. Stat. Quarterly **4** (1990) 269–282
11. http://www.nokia.com/nokia/0,,53720,00.html, April 25, 2005
12. Lauritzen, S.L. and Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems. J. Roy. Statist. Soc. B. **50**(2) (1988) 157–244
13. Lumière Project: Bayesian Reasoning for Automated Assistance. http://research.microsoft.com/∼ horvitz/lum.htm, April 27, 2005
14. Lockheed, Lockheed Martin Autonomous Control Logic to Guide Unmanned Underwater Vehicle, Press Release, Lockheed Martin Missiles and Space Communications Office, http://lmms.external.lmco.com/newsbureau/pressreleases/1996/9604.html, April 17, 1996
15. Long, W.: Medical diagnosis using a probabilistic causal network. Applied Artificial Intelligence **3** (1989) 367–383
16. Skaanning, C., Jensen, F.V., Kjaerulff, U., Parker, L., Pelletier, P. and Rostrup-Jensen, L.: http://www.cs.auc.dk/research/DSS/papers/skaanning98a.doc, April 25, 2005
17. Lee, T.T.: An information-theoretic analysis of relational databases - Part I: data dependencies and information metric. IEEE Transactions on Software Engineering. **SE-13** (10) (1987) 1049–1061
18. Maier, D.: The Theory of Relational Databases. Principles of Computer Science. Computer Science Press, Rockville, Maryland (1983)
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers, San Francisco (1988)
20. Shafer, G.: Probabilistic Expert Systems. Society for the Institute and Applied Mathematics, Philadelphia (1996)

21. Studeny, M.: Multiinformation and the problem of characterization of conditional-independence relations. Problems of Control and Information Theory. **18**(1) (1989) 3–16
22. Studeny, M.: Conditional independence relations have no finite complete characterization. In: Proceedings of the Eleventh Prague Conference on Information Theory, Statistical Decision Foundation and Random Processes, (1990) 377–396
23. Wong, S.K.M. and Butz, C.J.: Constructing the dependency structure of a multi-agent probabilistic network. IEEE Trans. Knowl. Data Eng. **13**(3) (2001) 395–415
24. Wong, S.K.M., Butz, C.J. and Wu, D.: On the implication problem for probabilistic conditional independency. IEEE Trans. Syst. Man Cybern. SMC-A **30**(6) (2000) 785–805
25. Wong, S.K.M., Butz, C.J. and Xiang, Y.: A method for implementing a probabilistic model as a relational database. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QC (1995) 556–564
26. Xiang, Y.: Probabilistic Reasoning in Multiagent Systems: A graphical models approach, Cambridge University Press, New York (2002)