Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering
Shaw Conference Center, Edmonton, Alberta, Canada May 9-12 1999

On Data and Probabilistic Dependencies

C.J. Butz, S.K.M. Wong and Y.Y. Yao
Department of Computer Science
University of Regina
Regina, SK S4S 0A2

Abstract

Data dependencies have been extensively studied in relational databases as they play a key role in the normalization process. On the other hand, probabilistic reasoning systems would not be practical without the notion of probabilistic conditional independence. In this paper, we present a detailed comparison of these two types of (in)dependencies. While past research has shown that multivalued dependency is a necessary but not sufficient condition for conditional independence, here we show in particular that functional dependency is a sufficient but not necessary condition for conditional independence.

1 Introduction

Due to the dependence of application programs on the physical storage of data in existing database models, Codd [3] proposed the *relational data model* [6] for the logical management of large data banks in transaction processing systems. Data dependencies were extensively studied since they played a key role in the normalization process. A *data dependency* is a restriction on the permissible sets of data that are allowed to appear in the database. Such constraints are useful in attenuating data redundancy and avoiding update anomalies. In particular, a data dependency can be used to provide an economical representation of a relation using projections of that relation. Codd [3] proposed the notion of *functional dependency* (FD) for the lossless decomposition of a relation into two projections. It is well known, however, that FD is a sufficient but not necessary condition for the lossless decomposition of a relation into two projections [6]. The notion of *multivalued dependency* (MVD) was proposed by Fagin [4] as a necessary and sufficient condition for the lossless decomposition of a relation into two projections. Rissanen [9] generalized MVD by introducing *join dependency* (JD). JDs can be used to losslessly decompose a relation into n-ary projections. However, Aho, Beeri and Ullman [1] demonstrated that arbitrary JDs may have problems including the intermediate join of two projections not being equal to the projection of the universal relation onto those attributes. A culminating result by Beeri et al. [2] was that the class of *acyclic* database schemes (acyclic hypergraphs) possess a number of desirable properties in database applications. This important class of join dependency is called *acyclic join dependency* (AJD).

On the other hand, probabilistic reasoning [5, 7, 8] has become an established framework for the management of *uncertain* knowledge. This approach assumes that knowledge can be represented as a joint probability distribution. However, a domain expert may have difficulty in specifying the required probability values for such a large frame. The notion of *probabilistic conditional independence* is extensively utilized to factorize the joint distribution into the product of *conditional probability tables*. The probability values of the joint distribution can then be obtained indirectly by the expert specifying the corresponding conditional probability values. To facilitate probabilistic inference in practice, however, it is useful to represent the joint distribution as the product of *marginal distributions* defined over an acyclic hypergraph.

In this paper, we explicitly demonstrate the relationships between the (in)dependencies used in the above two knowledge systems. The basis of this exposition is our *generalized relational data model* [11, 12]. Within this model probabilistic notions can be conveniently expressed in familiar relational terminology. Thus our model facilitates a comparison of data dependencies and probabilistic conditional independencies. While past research seems to indicate that data dependencies are a necessary but not sufficient condition for corresponding probabilistic notions, in this paper we show that *functional dependency* is a sufficient but not necessary condition for probabilistic conditional independence. These results may also be useful in knowledge discovery from database algorithms

0-7803-5579-2/99/$10.00

which utilize probabilistic notions.

This paper is organized as follows. Section 2 contains basic notions including the traditional relational data model and probabilistic models. In Section 3, we present a generalized relational data model. A detailed comparison of data dependencies and probabilistic independencies is provided in Section 4. The conclusion is presented in Section 5.

2 Basic Notions

We begin by defining some pertinent notions: relational databases and probabilistic reasoning systems with an emphasis on (in)dependencies. Henceforth, we may use the terms dependency and independency interchangeably.

2.1 Relational Databases

Let $\mathcal{N} = \{A_1, A_2, \ldots, A_m\}$ be a finite set of attributes. Each attribute $A \in \mathcal{N}$ is associated with a finite set D_A of permissible values called the domain of A. Given $X \subseteq \mathcal{N}$, we define a X-tuple t (or simply a *tuple* if X is understood) to be a function from X to $D_{A_1} \cup D_{A_2} \cup \ldots \cup D_{A_m}$ with the restriction that $t[A_i] \in D_{A_i}$ for all $A_i \in X$, where $t[A_i]$ denotes the value obtained by restricting the mapping to A_i. Thus a tuple is a mapping that associates a value with each attribute in X, i.e., $t[X] = \{t[A_1], t[A_2], \ldots, t[A_m]\}$ where $X = \{A_1, A_2, \ldots, A_m\}$. If $Y \subseteq X$ and t is a X-tuple, then $t[Y]$ denotes the Y-tuple obtained by restricting the mapping to Y. A *relation over* X, denoted $r[X]$, is a finite set of X-tuples. (We write $r[X]$ as r if X is understood.)

The *projection* of $r[\mathcal{N}]$ onto $X \subseteq \mathcal{N}$ is defined as $r[X] = \{t[X] \mid t \in r[\mathcal{N}]\}$. That is, $r[X]$ is the set of all tuples $t[X]$ such that t is in $r[\mathcal{N}]$. The *natural join* of two relations $r_1[X]$ and $r_2[Y]$, denoted $r_1[X] \bowtie r_2[Y]$, is defined as $r_1[X] \bowtie r_2[Y] = \{t[XY] \mid t[X] \in r_1[X], t[Y] \in r_2[Y]\}$, where we have written $X \cup Y$ as XY. That is, $r_1[X] \bowtie r_2[Y]$ denotes the set of tuples $t[XY]$ such that $t[X]$ is in r_1 and $t[Y]$ is in r_2.

Let $r[\mathcal{N}]$ be a relation over a set of attributes $\mathcal{N}$ and $X, Y \subseteq \mathcal{N}$. We say that the *functional dependency* (FD) $X \rightarrow Y$ is satisfied by $r[\mathcal{N}]$ if every two tuples of $r[\mathcal{N}]$ which have the same projection on X also have the same projection on Y. That is, the FD $X \rightarrow Y$ is satisfied by $r[\mathcal{N}]$ if and only if each X-value in $r[\mathcal{N}]$ is associated with precisely one Y-value. We call the FD $X \rightarrow Y$ *trivial* if $Y \subseteq X$. A trivial FD is satisfied by every $r[\mathcal{N}]$, where $XY \subseteq \mathcal{N}$. The FD $X \rightarrow Y$ is a sufficient but not a necessary condition for $r[\mathcal{N}]$ to be losslessly decomposed as $r[\mathcal{N}] = r[XY] \bowtie r[X(\mathcal{N} - XY)]$. For example, no nontrivial FDs are satisfied by $r[\{A_1, A_2, A_3\}]$ in Figure 1, yet $r[\{A_1, A_2, A_3\}]$ can still be losslessly decomposed into two projections.

A_1	A_2	A_3
0	0	0
0	0	1
1	0	0
1	0	1
1	1	1

$=$

A_1	A_2
0	0
1	0
1	1

$\bowtie$

A_2	A_3
0	0
0	1
1	1

Figure 1: No nontrivial FDs are satisfied by $r[\{A_1, A_2, A_3\}]$, yet $r[\{A_1, A_2, A_3\}]$ can still be losslessly decomposed into two projections.

Let $X, Y, Z \subseteq \mathcal{N}$ such that $Y \cap Z \subseteq X$, and $r[\mathcal{N}]$ a relation over $\mathcal{N}$. We say that the *multivalued dependency* (MVD), written $X \twoheadrightarrow Y \mid Z$, is satisfied by $r[\mathcal{N}]$, if and only if $r[XYZ] = r[XY] \bowtie r[XZ]$. The MVD $X \twoheadrightarrow Y \mid Z$ is called *nonembedded* in the special case where $XYZ = \mathcal{N}$. If $XYZ \subset \mathcal{N}$, then the MVD $X \twoheadrightarrow Y \mid Z$ is called *embedded*.

The MVD $X \twoheadrightarrow Y \mid (\mathcal{N} - XY)$ is a necessary and sufficient condition for $r[\mathcal{N}]$ to be losslessly decomposed as $r[\mathcal{N}] = r[XY] \bowtie r[X(\mathcal{N} - XY)]$. Thereby, the FD $X \rightarrow Y$ logically implies the MVD $X \twoheadrightarrow Y \mid (\mathcal{N} - XY)$, but the converse is not necessarily true.

Multivalued dependency is a special case of a more general kind of data dependency, called join dependency. We say that the *join dependency* (JD), written $\bowtie \mathcal{H}$, is satisfied by a relation $r[\mathcal{N}]$, if

$$r[\mathcal{N}] = r[h_1] \bowtie r[h_2] \bowtie \ldots \bowtie r[h_n],$$

where $h_i \subseteq \mathcal{N}$ and $\cup_{i=1}^{n} h_i = \mathcal{N}$. The database scheme $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ is in fact a hypergraph. We say that $\bowtie \mathcal{H}$ is an *acyclic* join dependency (AJD) if $\mathcal{H}$ is an *acyclic hypergraph* [2, 6]. (An acyclic hypergraph in fact represents a chordal undirected graph. Each maximal clique in the graph corresponds to a hyperedge in the acyclic hypergraph.) It has been demonstrated that an AJD has many desirable properties and plays an important role in database design [2].

2.2 Probabilistic Models

Let $\mathcal{N} = \{A_1, A_2, \ldots, A_m\}$ denote a finite set of discrete variables. Let V_A denote the finite frame (state space) of a variable $A \in \mathcal{N}$. We call an element of V_A a *configuration* of A.

A *joint probability distribution* [5, 8] over V_A is a function ϕ on V_A, assigning to each configuration $c \in V_A$ a real number $0 \leq \phi(c) \leq 1$ such that $\sum_{c \in V_A} \phi(c) = 1$. In general, a *potential* [5] is a function ψ on V_A such that $\psi(c)$ is a nonnegative real number and $\sum_{c \in V_A} \psi(c)$ is positive, i.e., at least one $\psi(c) > 0$. Each potential ψ can be transformed to a joint probability distribution ϕ through *normalization*, that is, by setting $\phi(c) = \psi(c) / \sum_{v \in V_A} \psi(v)$.

Let $X \subseteq \mathcal{N}$. We define V_X to be the Cartesian product of the frames of the variables in X. We call V_X the frame of X and its elements configurations of X. Let $Y \subseteq X \subseteq \mathcal{N}$. If $\mathbf{c}$ is a configuration of X, i.e., $\mathbf{c} \in V_X$, we write $\mathbf{c}^{\downarrow Y}$ for the configuration of Y obtained by deleting the values of the variables in X and not in Y. For example, let $Y = \{A_1, A_2\}, X = \{A_1, A_2, A_3, A_4\}$, and $\mathbf{c} = (c_1, c_2, c_3, c_4)$, where $c_i \in V_{A_i}$. Then, $\mathbf{c}^{\downarrow Y} = (c_1, c_2)$.

If X and Y are disjoint subsets of $\mathcal{N}$, $\mathbf{c}_X$ is a configuration of X, and $\mathbf{c}_Y$ is a configuration of Y, then we write $(\mathbf{c}_X \circ \mathbf{c}_Y)$ for the configuration of XY obtained by *concatenating* $\mathbf{c}_X$ and $\mathbf{c}_Y$. In other words, $(\mathbf{c}_X \circ \mathbf{c}_Y)$ is the unique configuration of XY such that $(\mathbf{c}_X \circ \mathbf{c}_Y)^{\downarrow X} = \mathbf{c}_X$ and $(\mathbf{c}_X \circ \mathbf{c}_Y)^{\downarrow Y} = \mathbf{c}_Y$.

Consider a distribution ϕ_X on a set X of variables. The *marginal* of ϕ_X on $Y \subseteq X$, denoted $\phi_X^{\downarrow Y}$, is defined as follows:

$$\phi_X^{\downarrow Y}(\mathbf{c}_Y) = \sum_{\mathbf{c}_{X-Y}} \phi_X(\mathbf{c}_Y \circ \mathbf{c}_{X-Y}),$$

where $\mathbf{c}_Y$ is a configuration of Y, $\mathbf{c}_{X-Y}$ is a configuration of $X - Y$, and $\mathbf{c}_Y \circ \mathbf{c}_{X-Y}$ is a configuration of X. For convenience, we write $\phi_X^{\downarrow Y}$ as $\phi(\mathbf{c}_Y)$ if no confusion arises.

Let X, Y, Z be disjoint subsets of $\mathcal{N}$, and V the frame of XYZ. We say that Y and Z are *conditionally independent* given X under ϕ, if for all configurations $\mathbf{c} \in V$,

$$\phi(\mathbf{c}) = \frac{\phi(\mathbf{c}^{\downarrow XY}) \cdot \phi(\mathbf{c}^{\downarrow XZ})}{\phi(\mathbf{c}^{\downarrow X})}, \qquad (1)$$

or equivalently

$$\phi(\mathbf{c}^{\downarrow Y} \mid \mathbf{c}^{\downarrow XZ}) = \phi(\mathbf{c}^{\downarrow Y} \mid \mathbf{c}^{\downarrow X}). \qquad (2)$$

Let V be the frame of the finite set of variables $\mathcal{N} = \{A_1, A_2, \ldots, A_m\}$. By the chain rule, a joint probability distribution ϕ over V can always be factorized as:

$$\begin{aligned} \phi(\mathbf{c}) &= \phi(\mathbf{c}^{\downarrow\{A_1\}}) \cdot \phi(\mathbf{c}^{\downarrow\{A_2\}} \mid \mathbf{c}^{\downarrow\{A_1\}}) \cdot \ldots \\ &\quad \cdot \phi(\mathbf{c}^{\downarrow\{A_m\}} \mid \mathbf{c}^{\downarrow\{A_1, A_2, \ldots, A_{m-1}\}}), \end{aligned}$$

where $\mathbf{c} \in V$. The above equation is an identity. However, one can use known conditional independencies to obtain a simpler representation of a joint probability distribution.

For example, let ϕ be a joint probability distribution over the frame of a set of variables $\mathcal{N} = \{A_1, A_2, A_3, A_4, A_5, A_6\}$. Consider the following known conditional independencies:

$$\begin{aligned} \phi(\mathbf{c}^{\downarrow\{A_3\}} | \mathbf{c}^{\downarrow\{A_1, A_2\}}) &= \phi(\mathbf{c}^{\downarrow\{A_3\}} | \mathbf{c}^{\downarrow\{A_1\}}), \\ \phi(\mathbf{c}^{\downarrow\{A_4\}} | \mathbf{c}^{\downarrow\{A_1, A_2, A_3\}}) &= \phi(\mathbf{c}^{\downarrow\{A_4\}} | \mathbf{c}^{\downarrow\{A_2, A_3\}}), \\ \phi(\mathbf{c}^{\downarrow\{A_5\}} | \mathbf{c}^{\downarrow\{A_1, A_2, A_3, A_4\}}) &= \phi(\mathbf{c}^{\downarrow\{A_5\}} | \mathbf{c}^{\downarrow\{A_2, A_3\}}), \\ \phi(\mathbf{c}^{\downarrow\{A_6\}} | \mathbf{c}^{\downarrow\{A_1, A_2, A_3, A_4, A_5\}}) &= \phi(\mathbf{c}^{\downarrow\{A_6\}} | \mathbf{c}^{\downarrow\{A_5\}}). \end{aligned}$$

Utilizing these conditional independencies, the joint probability distribution ϕ written using the chain rule can be expressed in the simpler form:

$$\begin{aligned} \phi(\mathbf{c}) &= \phi(\mathbf{c}^{\downarrow\{A_1\}}) \cdot \phi(\mathbf{c}^{\downarrow\{A_2\}} | \mathbf{c}^{\downarrow\{A_1\}}) \cdot \phi(\mathbf{c}^{\downarrow\{A_3\}} | \mathbf{c}^{\downarrow\{A_1\}}) \\ &\quad \cdot \phi(\mathbf{c}^{\downarrow\{A_4\}} | \mathbf{c}^{\downarrow\{A_2, A_3\}}) \cdot \phi(\mathbf{c}^{\downarrow\{A_5\}} | \mathbf{c}^{\downarrow\{A_2, A_3\}}) \\ &\quad \cdot \phi(\mathbf{c}^{\downarrow\{A_6\}} | \mathbf{c}^{\downarrow\{A_5\}}). \end{aligned} \qquad (3)$$

A *directed acyclic graph* (DAG) is used to graphically encode the known conditional independencies. Thus, the DAG aids in the *acquisition* of the probabilistic knowledge.

To facilitate probabilistic *inference* in practice, however, it is useful to represent ϕ in terms of marginal distributions over an acyclic hypergraph. The representation of ϕ in terms of conditional probability tables can always be transformed [5, 8] into an expression in terms of marginal distributions defined over an acyclic hypergraph. For example, the representation of ϕ in equation (3) can be rewritten as:

$$\phi(\mathbf{c}) = \frac{\phi(\mathbf{c}^{\downarrow h_1}) \cdot \phi(\mathbf{c}^{\downarrow h_2}) \cdot \phi(\mathbf{c}^{\downarrow h_3}) \cdot \phi(\mathbf{c}^{\downarrow h_4})}{\phi(\mathbf{c}^{\downarrow h_1 \cap h_2}) \cdot \phi(\mathbf{c}^{\downarrow h_2 \cap h_3}) \cdot \phi(\mathbf{c}^{\downarrow h_3 \cap h_4})}, \qquad (4)$$

where $\mathcal{H} = \{h_1 = \{A_1, A_2, A_3\}, h_2 = \{A_2, A_3, A_4\}, h_3 = \{A_2, A_3, A_5\}, h_4 = \{A_5, A_6\}\}$ is an acyclic hypergraph.

A DAG coupled with the corresponding conditional probability tables is called a *Bayesian* network [5, 8]. On the other hand, an acyclic hypergraph coupled with the corresponding marginal distributions is called a *Markov* network [5]. A Markov network is also referred to as a *join* or *junction* tree [8].

A Bayesian network is more expressive than a Markov network. A Bayesian network utilizes *both* embedded and nonembedded conditional independencies. However, a Markov network only utilizes nonembedded conditional independencies. For example, the embedded conditional independency of $\{A_2\}$ and $\{A_3\}$ given $\{A_1\}$ under ϕ in equation (3) is *not* utilized in equation (4).

A_1	A_2	...	A_m	f_ϕ
$t_1[A_1]$	$t_1[A_2]$	...	$t_1[A_m]$	$t_1[f_\phi] = \phi(t_1[\mathcal{N}])$
$t_2[A_1]$	$t_2[A_2]$	...	$t_2[A_m]$	$t_2[f_\phi] = \phi(t_2[\mathcal{N}])$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_s[A_1]$	$t_s[A_2]$	...	$t_s[A_m]$	$t_s[f_\phi] = \phi(t_s[\mathcal{N}])$

Figure 2: A distribution ϕ expressed as a *relation* Φ.

3 The Generalized Relational Data Model

In this section, we present a *generalized relational data model* which subsumes the traditional relational data model. Probabilistic notions are then expressed in familiar relational terminology. At the same time, this model is *not* restricted to a probabilistic interpretation but in fact can be applied to a variety of problems [10, 12].

Let V be the frame of a finite set of discrete variables $\mathcal{N} = \{A_1, A_2, \ldots, A_m\}$, and ϕ be a joint probability distribution over V. We can express the joint probability distribution ϕ as a generalized relation Φ. The generalized relation Φ is defined by the set of attributes $\{A_1, A_2, \ldots, A_m, f_\phi\}$. Each row in Φ is defined by a tuple t_i in a standard relation $r[\mathcal{N}]$ as shown in Figure 2. We will say Φ is a relation over $\mathcal{N}$ since the attribute f_ϕ is understood by context.

Having defined a joint probability distribution as a generalized relation, probabilistic operations on distributions can be defined as generalized relational operations. Computing the projection of a relation and the natural join of two relations in relational database theory corresponds to computing a marginal distribution and the product of two distributions in probabilistic reasoning theory, respectively.

If $\Phi_\mathcal{N}$ is a relation and $X \subseteq \mathcal{N}$, then the marginalization of $\Phi_\mathcal{N}$ onto X is the *marginal* relation, denoted $\Phi_\mathcal{N}^{\downarrow X}$, with attributes $X \cup \{f_{\phi_\mathcal{N}^{\downarrow X}}\}$, defined by:

$$\begin{aligned}\Phi_\mathcal{N}^{\downarrow X} &= \{t[X \cup \{f_{\phi_\mathcal{N}^{\downarrow X}}\}] \mid t[X] \in \Phi_\mathcal{N}[X], \text{ and} \\ &\qquad t[f_{\phi_\mathcal{N}^{\downarrow X}}] = \phi_\mathcal{N}^{\downarrow X}(t[X]) = \sum_{t' \in \Phi_\mathcal{N}} \phi_\mathcal{N}(t'[\mathcal{N}])\},\end{aligned}$$

where $t'[X] = t[X]$.

Let ϕ_X and ϕ_Y be two distributions over X and Y, respectively. We can express the product $\phi_X \cdot \phi_Y$ as the *product join* $\Phi_X \times \Phi_Y$ of the corresponding relations Φ_X and Φ_Y. That is, $\Phi_X \times \Phi_Y$ is a relation on the set of attributes $XY \cup \{f_{\phi_X \cdot \phi_Y}\}$ defined as follows:

$$\begin{aligned}&\Phi_X \times \Phi_Y \\ = \;& \{t[XY \cup \{f_{\phi_X \cdot \phi_Y}\}] \mid t[XY] \in (\Phi_X[X] \bowtie \Phi_Y[Y]), \\ & \quad \text{and } t[f_{\phi_X \cdot \phi_Y}] = \phi_X(t[X]) \cdot \phi_Y(t[Y])\}.\end{aligned}$$

Let Φ be a relation. The inverse of Φ is the *inverse* relation, denoted Φ^{-1}, with attributes $\mathcal{N} \cup \{f_{\phi^{-1}}\}$, defined by:

$$\begin{aligned}\Phi^{-1} = \; & \{t[\mathcal{N} \cup \{f_{\phi^{-1}}\}] \mid t[\mathcal{N}] \in \Phi[\mathcal{N}] \text{ and} \\ & t[f_{\phi^{-1}}] = \begin{cases} 1/t[f_\phi] & \text{if } t[f_\phi] > 0, \\ t[f_\phi] & \text{otherwise} \end{cases} \}.\end{aligned}$$

Generalized relational data dependencies can now be introduced using the above generalized operators.

Let $X, Y, Z \subseteq \mathcal{N}$ such that $Y \cap Z \subseteq X$, and $\Phi_\mathcal{N}$ a relation over $\mathcal{N}$. We say that the *generalized multivalued dependency* (GMVD), written

$$X \Rightarrow\Rightarrow Y \mid Z,$$

is satisfied by the relation $\Phi_\mathcal{N}$ if and only if the marginal relation $\Phi_\mathcal{N}^{\downarrow XYZ}$ of $\Phi_\mathcal{N}$ can be factorized as follows:

$$\begin{aligned}\Phi_\mathcal{N}^{\downarrow XYZ} &= \Phi_\mathcal{N}^{\downarrow XY} \times \Phi_\mathcal{N}^{\downarrow XZ} \times (\Phi_\mathcal{N}^{\downarrow X})^{-1} \\ &\equiv \Phi_\mathcal{N}^{\downarrow XY} \otimes \Phi_\mathcal{N}^{\downarrow XZ}. \qquad (5)\end{aligned}$$

We refer to the binary operation $\otimes$ defined above as the *generalized join*. For example, it can be verified that the relation Φ over $N = \{A_1, A_2, A_3\}$ depicted in Figure 3 satisfies the GMVD $\{A_2\} \Rightarrow\Rightarrow \{A_1\} \mid \{A_3\}$, i.e., $\Phi = \Phi^{\downarrow\{A_1,A_2\}} \otimes \Phi^{\downarrow\{A_2,A_3\}}$.

A_1	A_2	A_3	f_ϕ
0	0	0	0.2
0	0	1	0.4
0	1	1	0.3
1	1	1	0.1

$= \quad \Phi^{\downarrow\{A_1,A_2\}} \quad \otimes \quad \Phi^{\downarrow\{A_2,A_3\}}$

Figure 3: The relation Φ over $\mathcal{N} = \{A_1, A_2, A_3\}$ satisfies the GMVD $\{A_2\} \Rightarrow\Rightarrow \{A_1\} \mid \{A_3\}$.

GMVD is a special case of a more general dependency called generalized acyclic join dependency. We say a *generalized acyclic join dependency* (GAJD) $\otimes\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ is satisfied by a relation $\Phi_\mathcal{N}$, if $\Phi_\mathcal{N}$ can be written as:

$$\Phi_\mathcal{N} = (\ldots((\Phi_\mathcal{N}^{\downarrow h_1} \otimes \Phi_\mathcal{N}^{\downarrow h_2}) \otimes \Phi_\mathcal{N}^{\downarrow h_3}) \ldots \otimes \Phi_\mathcal{N}^{\downarrow h_n}), \qquad (6)$$

where the sequence $h_1, h_2, \ldots, h_n$ is a *hypertree construction ordering* [2, 10, 12] for $\mathcal{H}$.

For example, by the definition of the generalized join operator $\otimes$, equation (4) can be expressed as:

$$\Phi = ((\Phi^{\downarrow\{A_1,A_2,A_3\}} \otimes \Phi^{\downarrow\{A_2,A_3,A_4\}}) \otimes \Phi^{\downarrow\{A_2,A_3,A_5\}}) \otimes \Phi^{\downarrow\{A_5,A_6\}}. \quad (7)$$

Thus, the GAJD $\otimes\mathcal{H} = \{h_1 = \{A_1, A_2, A_3\}, h_2 = \{A_2, A_3, A_4\}, h_3 = \{A_2, A_3, A_5\}, h_4 = \{A_5, A_6\}\}$ is satisfied by Φ. The important point to realize is that the Markov network in equation (4) can be stated as the GAJD in equation (7).

4 Comparison of Data and Probabilistic Dependencies

It should be clear that the definition of probabilistic conditional independence given in equation (1) is equivalent to stating that the generalized relation $\Phi_{\mathcal{N}}^{\downarrow YXZ}$ satisfies the GMVD $X \Rightarrow\Rightarrow Y \mid Z$ in equation (5), namely, $\Phi_{\mathcal{N}}^{\downarrow YXZ} = \Phi_{\mathcal{N}}^{\downarrow XY} \otimes \Phi_{\mathcal{N}}^{\downarrow XZ}$. Thereby, we may use the terms GMVD and probabilistic conditional independence interchangeably.

We first consider the special case where the joint probability distribution ϕ is *uniform* (constant), i.e., $\phi(\mathbf{c}) = k$ for all configurations $\mathbf{c}$ in the frame $V_{\mathcal{N}}$.

Theorem 1 [11] Let ϕ be a *uniform* distribution over the frame of a finite set of variables $\mathcal{N}$. Let Φ be a generalized relation representing ϕ. Let X, Y, and Z be disjoint subsets such that $\mathcal{N} = XYZ$. Then the generalized relation Φ satisfies the GMVD $X \Rightarrow\Rightarrow Y \mid Z$ if and only if the traditional relation $\Phi[\mathcal{N}]$ satisfies the MVD $X \rightarrow\rightarrow Y \mid Z$.

For a *uniform* distribution, Theorem 1 indicates that

$$\Phi = \Phi^{\downarrow XY} \otimes \Phi^{\downarrow XZ},$$

if and only if

$$\Phi[\mathcal{N}] = \Phi[XY] \bowtie \Phi[XZ],$$

where $\Phi[\mathcal{N}]$ denotes the projection of Φ onto the set of attributes $\mathcal{N}$.

Corollary 1 Let ϕ and Φ be as in Theorem 1. Then the generalized relation Φ satisfies the GAJD $\otimes\mathcal{H}$ if and only if the traditional relation $\Phi[\mathcal{N}]$ satisfies the AJD $\bowtie\mathcal{H}$.

We now turn our attention to the case where the joint probability distribution ϕ is *arbitrary*, i.e., not necessarily uniform.

For arbitrary distributions, MVD is a necessary but not sufficient condition for GMVD. In our generalized relational data model, we represent the frame of the set of variables $\mathcal{N}$ as a *fixed* relation. Fagin [4] has shown that a relation can be losslessly decomposed into two projections if and only if the corresponding MVD holds. It immediately follows that MVD is a *necessary* condition for GMVD to hold, namely, GMVD implies MVD. However, MVD does not imply GMVD. For example, consider the distribution Φ illustrated in Figure 4. It can be verified that the GMVD $\{A_2\} \Rightarrow\Rightarrow \{A_1\} \mid \{A_3\}$ is not satisfied by Φ, i.e., $\Phi \neq \Phi^{\downarrow\{A_1,A_2\}} \otimes \Phi^{\downarrow\{A_2,A_3\}}$. However, the MVD $\{A_2\} \rightarrow\rightarrow \{A_1\} \mid \{A_3\}$ is satisfied by the traditional relation $\Phi[\mathcal{N}]$, obtained by projecting Φ onto the set of attributes $\{A_1, A_2, A_3\}$, as shown in Figure 1. This example demonstrates that MVD is not a sufficient condition for GMVD.

A_1	A_2	A_3	f_ϕ
0	0	0	0.4
0	0	1	0.2
1	0	0	0.1
1	0	1	0.1
1	1	1	0.2

$\neq \quad \Phi^{\downarrow\{A_1,A_2\}} \quad \otimes \quad \Phi^{\downarrow\{A_2,A_3\}}$

Figure 4: The GMVD $\{A_2\} \Rightarrow\Rightarrow \{A_1\} \mid \{A_3\}$ is *not* satisfied by Φ. However, the MVD $\{A_2\} \rightarrow\rightarrow \{A_1\} \mid \{A_3\}$ is satisfied by the traditional relation $\Phi[\mathcal{N}]$ depicted in Figure 1.

These results clearly indicate that MVD is a necessary but not sufficient condition for probabilistic conditional independence. Since GMVD is a special case of GAJD, it follows that AJD is a necessary but not sufficient condition for GAJD.

On the other hand, we now show that functional dependency is a sufficient but not necessary condition for probabilistic conditional independence.

Theorem 2 Let ϕ be a joint probability distribution over the frame of a finite set of variables $\mathcal{N}$. Let Φ be a generalized relation representing ϕ. Let $X, Y \subseteq \mathcal{N}$ and $Z = \mathcal{N} - XY$. The *functional dependency* $X \rightarrow Y$ satisfied by the traditional relation $\Phi[\mathcal{N}]$ is a sufficient but not necessary condition for the GMVD $X \rightarrow\rightarrow Y \mid Z$ to be satisfied by the generalized relation Φ.

Proof: We first show that FD is sufficient for GMVD.

Suppose the FD $X \rightarrow Y$ is satisfied by the traditional relation $\Phi[\mathcal{N}]$. This means that every X-value is associated with precisely one Y-value. In other words, if $t_i[X] = t_j[X]$, then $t_i[Y] = t_j[Y]$, $t_i, t_j \in \Phi[\mathcal{N}]$. Since $\mathcal{N} = XYZ$, it immediately follows that there are no duplicate XZ-values in $\Phi[\mathcal{N}]$, namely, $t_i[XZ] \neq t_j[XZ]$ for all $t_i, t_j \in \Phi[\mathcal{N}]$. Consider any tuple $t = \{t[X] = x, t[Y] = y, t[Z] = z\}$ in $\Phi[\mathcal{N}]$, namely, $(x \circ y \circ z)$. The marginal distribution $\phi(x)$ of the X-value x is defined by

$$\phi(x) = \sum_{t'[Y], t'[Z]} \phi(x \circ t'[Y] \circ t'[Z]). \quad (8)$$

By the initial assumption, equation (8) can be rewritten as

$$\phi(x) = \sum_{t'[Z]} \phi(x \circ y \circ t'[Z]) = \phi(xy). \quad (9)$$

Using equation (9) we obtain

$$\frac{\phi(xy) \cdot \phi(xz)}{\phi(x)} = \phi(xz).$$

However, $\phi(xz) = \phi(xyz)$ since there are no duplicate XZ-values in $\Phi[\mathcal{N}]$. By definition, Y and Z are conditionally independent given X under ϕ. That is, the GMVD $X \Rightarrow\Rightarrow Y \mid Z$ is satisfied by Φ.

We now show that FD is not a necessary condition for GMVD. The GMVD $\{A_2\} \Rightarrow\Rightarrow \{A_1\} \mid \{A_3\}$ is satisfied by the generalized relation Φ over $\mathcal{N} = \{A_1, A_2, A_3\}$ as shown in Figure 3. However, no nontrivial FDs are satisfied by $\Phi[\mathcal{N}]$. □

Theorem 2 indicates that functional dependency is a sufficient but not necessary condition for probabilistic conditional independence.

5 Conclusion

In this paper we have shown that *functional dependency* is a sufficient but not necessary condition for probabilistic conditional independence. Studying the relationship between data and probabilistic dependencies is important in understanding the intriguing relationship between relational databases and Bayesian networks. It may then be possible to design one *unified* system for both data and uncertainty management. Furthermore, these results may also be useful in knowledge discovery from database algorithms which utilize probabilistic notions.

References

[1] A.V. Aho, C. Beeri and J.D. Ullman, "The Theory of Joins in Relational Databases", *ACM Transactions on Database Systems*, Vol. 4, No. 3, pp. 297-314, 1979.

[2] C. Beeri, R. Fagin, D. Maier and M. Yannakakis, "On the Desirability of Acyclic Database Schemes", *Journal of the ACM*, Vol. 30, No. 3, pp. 479-513, 1983.

[3] E.F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Communication of ACM*, Vol. 13, No. 6, pp. 377-387, 1970.

[4] R. Fagin, "Multivalued Dependencies and a New Normal Form for Relational Databases", *ACM Transactions on Database Systems*, Vol. 2, No. 3, pp. 262-278, 1977.

[5] P. Hajek, T. Havranek and R. Jirousek, *Uncertain Information Processing in Expert Systems*, CRC Press, 1992.

[6] D. Maier, *The Theory of Relational Databases*, Computer Science Press, 1983.

[7] R.E. Neapolitan, *Probabilistic Reasoning in Expert Systems*, Wiley, 1990.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[9] J. Rissanen, "Independent Components of Relations", *ACM Transactions on Database Systems*, Vol. 2, No. 4, pp. 317-325, 1977.

[10] G. Shafer, "An Axiomatic Study of Computation in Hypertrees", University of Kansas, School of Business Working Papers, No. 232, 1991.

[11] S.K.M. Wong, "An Extended Relational Data Model for Probabilistic Reasoning", *Journal of Intelligent Information Systems*, Vol. 9, pp. 181–202, 1997.

[12] S.K.M. Wong, C.J. Butz and Y. Xiang, "A Method for Implementing a Probabilistic Model as a Relational Database", Eleventh Conference on Uncertainty in Artificial Intelligence, pp. 556-564, 1995.