# A Bayesian Framework for Modifications of Probabilistic Relational Data 

Debabrata Dey<br>Louisiana State University and A\&M College<br>Sumit Sarkar<br>Louisiana State University and A\&M College

Follow this and additional works at: http://aisel.aisnet.org/icis 1996

## Recommended Citation

Dey, Debabrata and Sarkar, Sumit, "A Bayesian Framework for Modifications of Probabilistic Relational Data" (1996). ICIS 1996 Proceedings. 11.
http://aisel.aisnet.org/icis1996/11

# A BAYESIAN FRAMEWORK FOR MODIFICATIONS OF PROBABILISTIC RELATIONAL DATA 

Debabrata Dey<br>Sumit Sarkar<br>Louisiana State University and A\&M College


#### Abstract

The inherent uncertainty pervasive over the real world often forces business decisions to be made using uncertain data. The conventional relational model does not have the ability to handle uncertain data. In recent years, several approaches have been proposed in the literature for representing uncertain data by extending the relational model, primarily using probability theory. However, the aspect of database modification has been overlooked in these investigations. It is clear that any modification of existing probabilistic data, based on new information, amounts to the revision of one's belief about real world objects. In this paper, we examine the aspect of belief revision and develop a generalized algorithm that can be used for modification of existing data in a probabilistic relational database.


## 1. INTRODUCTION

Decision making in most modern businesses requires extensive use of large volumes of data; the quality of these decisions depends critically on the quality of the available data. The database approach to business information systems provides the decision maker with an easy to use interface, and yet maintains a high level of quality and integrity on large volumes of data. In recent years, database systems based on the relational model (Codd 1970) have become very popular for handling business data. Unfortunately, relational databases do not handle incomplete and uncertain data in a comprehensive manner. In many real world business applications, however, the available data are often uncertain. For instance, the decision to introduce a new product may be based on market survey results, but no survey data can determine the market behavior with certainty. Similarly, a securities database can provide useful insights into various investment opportunities, but can never predict how a particular stock will perform in the future. In both cases, the decision maker must rely on incomplete and uncertain data that is available. Uncertainty in data values can also arise from several other sources:

- The actual value of a data item may be unknown. For example, it may be known that a new employee has been hired, but the actual salary of that employee may be unknown. Of course, one can use other information such as rank, department, job description, and competitive market salary to form an opinion about the salary of the new employee.
- The data item may not be realized yet. For example, the instructor for the graduate level database course may not be assigned yet for the next semester. However, if there are only two faculty members who can teach this course, then it may be reasonable to assign a probability of 0.5 that either will offer that course.
- Uncertainty may also arise from consolidation or summarization of data. For example, the results of a market survey are often expressed in a consolidated manner in which the details of individual consumer preferences are summarized.
- Another source of data uncertainty is data heterogeneity (Motro 1990). When two heterogeneous databases show different values for the same real world data item, its actual value is not known with certainty.

Acknowledgments: The authors wish to thank Dr. Veda C. Storey and Dr. Amit Basu for their helpful comments.

Having no means to model them, the relational model ignores all uncertain data, and focuses primarily on data values that are known with certainty; uncertain data values are represented using "null" (or unknown) values (Codd 1979; Date 1986). Consequently, relational databases do not yield satisfactory results in many business situations. The decision making process requires a data model that (i) can organize uncertain information from the real world in a readily usable form, and (ii) can consistently update this information based on changes in the real world.

Recently, several extensions of the relational model have been proposed, primarily based on probability theory, ${ }^{1}$ that attempt to overcome the above problem. For example, Cavallo and Pittarelli (1987) extend the relational model to represent uncertainty using the popular probability measure. They assign a probability with every tuple in a relation; it indicates the joint probability of all the attribute values in that tuple. Barbará, Garcia-Molina, and Porter (1992) propose an extension of the relational model using probability theory. They adopt a non-1NF (first normal form) view of probabilistic relations, and redefine the project, select and join operations using semantics of probability theory. Dey and Sarkar (1996) propose a 1NF structure for probabilistic relations and the associated algebra. Their algebra is a consistent extension of the relational algebra, and reduces to the latter.

However, these extensions overlook one of the most important aspects of a data model, namely modification of existing data. Most databases, with the possible exception of historical databases, go through the normal process of addition, deletion and update of data items. In a probabilistic database, these modifications could result from two sources. First, the belief about the joint distribution of attributes may undergo modifications as the user obtains more and more reliable information. Second, some of the objects in the database may change their states. For example, a new employee may be hired, or an old one may be transferred to a new department. Clearly, the database must accommodate both by changing the existing belief about one or more objects. The purpose of this research is to examine the revision of belief in a probabilistic database. We adopt the point estimate interpretation of the probabilistic relational model by Dey and Sarkar and propose a belief revision algorithm for it. There are several reasons for our choice of the above model over other existing alternatives. First, the above model is in 1NF and is easy to implement. Second, their model and algebra are formally outlined in terms of the valid objects and associated operations. Finally, in this model, uncertainty in data has been modeled using probability theory which, due to its wide acceptance, is easy to interpret. Probability measures have a rich theoretical basis for representing uncertainty and they lend themselves to empirical testability and provide an easy-to-use semantics (Pearl 1986, 1989).

The rest of this paper is organized as follows. The probabilistic relational model of Dey and Sarkar is outlined in section 2. In section 3, we discuss the Bayesian framework for belief revision and its applicability to probabilistic databases. The belief revision scheme and the accompanying algorithm are proposed in section 4 . In section 5 , we provide examples to illustrate the scheme, and in section 6 , we conclude the paper and offers future research directions.

## 2. THE PROBABILISTIC RELATIONAL MODEL

### 2.1 Structure and Meaning of Probabilistic Relations

Before discussing how uncertain data can be represented using relations, let us first examine how the conventional relational model represents facts about the real world. To illustrate this, consider the EMPLOYEE relation shown in Table 1. In this relation, the first row means that the predicate EMPLOYEE(3025, Lyons, James, cashier, 20K, shoe) must be assigned a value of "TRUE" (represented by 1). Any predicate not listed in this relation is assigned a value of "FALSE" (o0); this is the usual closed world assumption. However, when there is uncertainty associated with real world objects, it is not always possiblect assign a value of 0 or 1 to these predicates. As aresult, the first-order predicate calculus-as well as a conventional relation which may be viewed as a collection of first-order predicates-cannot represent uncertain data.
${ }^{1}$ A noted exception is the model of Raju and Majumdar (1988). They generalize the basic relational concepts by using fuzzy relations. In their model, each tuple is assigned a possibility measure that represents the possibility of its membership in the relation. Their attributes can also take fuzzy subsets as their values.

Table 1. EMPLOYEE: A Conventional Relation

| EMP\# | LName | FName | rank | salary | dept |
| :---: | :--- | :--- | :--- | :---: | :---: |
| 3025 | Lyons | James | cashier | 20 K | shoe |
| 6723 | Kivari | Jack | clerk | 18 K | auto |
| 6879 | Peters | Julia | cashier | 25 K | toy |

This limitation can be addressed by using the probability calculus in the place of the first-order predicate calculus. Instead of a value of 0 or 1 , a predicate can be assigned a probability. We would write P PEMPLOYEE(3025, Lyons, James, cashier, 20K, shoe) $]=0.6$, to represent the fact that, with a probability of 0.6 , there exists an employee with the following attribute values: EMP\#=3025, LName=Lyons, FName=James, rank=cashier, salary=20K, and dept=shoe. This can be represented in the usual tabular format of a relation by appending a special column called the probability stamp or $p S$. Table 2 illustrates this representation, where only predicates with non-zero probabilities are explicitly written as tuples or rows. Thus, a relation captures the joint distribution over all of its non-probability attributes; the probability associated with a particular attribute value can be obtained by appropriate marginalization of this joint distribution. For example, from the first three rows in the EMPLOYEE relation in Table 2, we can infer that the marginal probability associated with EMP\# 3025 is one. This implies that this employee is known to exist with certainty, although some of the attribute values are not known precisely. The probability that the rank of this employee is cashier is only 0.8 (from the first and the third rows).

Table 2. EMPLOYEE: A Probabilistic Relation

| EMP\# | LName | FName | rank | salary | dept | $p S$ |
| :---: | :--- | :--- | :--- | :---: | :---: | :---: |
| 3025 | Lyons | James | cashier | 20 K | shoe | 0.6 |
| 3025 | Lyons | James | clerk | 15 K | toy | 0.2 |
| 3025 | Lyons | James | cashier | 15 K | auto | 0.2 |
| 6723 | Kivari | Jack | clerk | 18 K | toy | 0.4 |
| 6723 | Kivari | Jack | cashier | 20 K | auto | 0.4 |
| 6723 | Kivari | Jack | $*$ | $*$ | $*$ | 0.1 |
| 6879 | Peters | Julia | clerk | 25 K | toy | 0.3 |
| 6879 | Peters | Julia | clerk | $*$ | toy | 0.1 |
| 6879 | Peters | Julia | cashier | $*$ | $*$ | 0.6 |

It is also possible that the user may have a more complete probabilistic information about some attributes over others. In that case, null values $\left(^{*}\right)$ are used as values for stochastic attributes that are not completely specified. This is illustrated with the second and third tuples corresponding to EMP\# 6879; additional information about rank is captured in these two tuples. An important question, then, is the interpretation of the probability stamp when the conditional probability distribution is not fully specified. How we interpret the probability stamp has to do with the interpretation given to the portion of the total probability mass (associated with a key value) that is not specified, called the missing probability in Barbará, Garcia-Molina, and Porter. There are two possible interpretations that may be given to the missing probability. The first is that they could be distributed over the entire set of realizations of the attributes, including the ones that already appear in the relation. In that case, the uncertainty associated with the attribute values for tuples that appear in the relation are represented by probability intervals. The probability stamp associated with a tuple is then the lower bound of the probability interval for that tuple (as in Barbará, Garcia-Molina, and Porter 1992). Consider the example of employee with EMP\# 6723; this key value has a missing probability of 0.1 . Since this probability mass could be assigned to any value, including those that have already appeared, the probability that EMP\#=6723, rank="clerk," salary=18K, and dept="toy" lies in the interval [0.4, 0.5].

The second, and a more intuitive, interpretation for the missing probabilities is that the missing probability is associated with realizations of those values of attributes that are not already included in the relation. Thus, in Table 2, the missing probability of 0.1 for EMP\# 6723 could be distributed in any manner over those joint realizations for rank, salary and department that are not already included in the table. With this interpretation, the probability stamps for tuples that do appear in the relation are construed as point estimates of the probabilities for given values of the attributes. Therefore, the probability that EMP\#=6723, rank="clerk," salary $=18 \mathrm{~K}$ and dept="toy" is interpreted to be 0.4 . When the joint distribution is completely specified, it is clear that a probability interval reduces to a point. For this work, we consider only the second (point estimate) interpretation of the probability stamp and provide the revision scheme based on that. Future research will examine possible extension of the belief revision approach presented here to the first (interval-based) interpretation.

An issue of practical significance is how to obtain the probabilistic data. Several methods have been presented in the literature for this purpose. Pearl (1986) discusses how the probability values could be assigned based on the user's confidence. It is also possible to assign probability values based on sampling, where a portion of the population is sampled to estimate the distribution (Barbará, Garcia-Molina, and Porter 1992). Yet a third method, based on maximizing the entropy subject to a set of known constraints, is described in the classical work by Jaynes (1968).

A few restrictions are imposed on a probabilistic relation so that data integrity can be maintained. We do not allow two distinct tuples with the same values for all non-probability attributes-also called value-equivalent tuples-to be present in a relation. Value-equivalent tuples are similar to duplicates in the conventional relational model. Existence of value-equivalent tuples introduces ambiguity about the distribution and are not allowed.

The second restriction is about the primary key of a relation. In the relational model, every tuple in a relation represents a unique object (i.e., an entity or a relationship) from the real world; a primary key is a minimal set of attributes that uniquely identifies a tuple, and hence an object. In the probabilistic extension, associated with every object, there may be several tuples representing the complete joint distribution of its attributes. Hence, we retain the object surrogate interpretation of the primary key (i.e., unique identifier of real world objects) and discard the notion of the primary key as a unique identifier of tuples. Ideally then, probability stamps associated with a primary key value should add up to one. In that case, the existence of the object (identified by that key value) is certain and the joint probability distribution for all its attributes is completely specified. However, the complete distribution is not necessary in order to store probabilities about attributes. If the existence of the object itself is uncertain, then the probability stamps associated with the key value of that object could be less than one. The modified requirement, then, is that the probability stamps associated with any given key value must add up to no more than one. Furthermore, the primary key and the probability stamp are not allowed to have "null" values.

The third and final restriction is about the foreign keys in a relation (referential integrity constraint). Since the foreign key values in a relation refer to other objects (in the same or a different relation), the probability assignments for a foreign key value cannot be more than the probability of existence of the referred object.

### 2.1 The Relational Algebra

We can now express the relational structure and operations more formally. In these definitions, we ignore the existence of null values. However, the extension to incorporate null values is straightforward and has been discussed by Dey and Sarkar.

Let $N=\{1,2, \ldots, n\}$ be an arbitrary set of integers. A relation scheme $R$ is a set of attribute names $\left\{A_{1}, A_{2}, \ldots A_{n}\right\}$, one of which may be a probability stamp $p S$. Corresponding to each attribute name $A_{i}, i=0 N$, is a set $D_{i}$ called the domain of $A_{2}$. If $A_{i}$ $=p S$, then $D_{i}=(0,1]$. The multiset $\boldsymbol{D}=\left\{D_{1}, D_{2}, \ldots, D_{n}\right\}$ is called the domain of $R$. A tuple $\alpha$ over $R$ is a function from $R$ to $D$, such that $\alpha\left(A_{1}\right) \in D_{i}, i \in N$. In other words, a tuple $\alpha$ over $R$ can be viewed as a set of attribute name-value pairs:

$$
=\left\{A, v_{i}| | \forall i \in N\left(A \in R \wedge v_{i} \in D_{i}\right)\right\}
$$

Restriction of a tuple $\alpha$ over $S \subset R$, IS, written $\alpha(S)$, is the sub-tuple containing values for attribute names in $S$ only, i.e., $\alpha\left\langle\left\langle\mathrm{A}_{\mathrm{i}}, \mathrm{v}_{\mathrm{i}}\right\rangle \in \alpha \mid \mathrm{A} \in \mathrm{S}\right\rangle$.

A tuple $\alpha$ over $R$ represents our belief about attributes (in $R$ ) of a real world object. If $p S$ is in $R$, then a probability of $\alpha(p S)>0$ is assigned to the fact that an object has the values $\alpha(R-\{p S\})$ for the corresponding attributes. In other words, the attribute $p S$ represents the joint distribution of all the attributes taken together. If $p S$ is not in $R$, i.e., if the relation scheme $R$ is deterministic, then every tuple on $R$ is assigned a probability of one, and is not explicitly written. However, if $\alpha$ is a tuple on a deterministic scheme $R$, it is implicitly assumed that $\alpha(p S)=1$.

Two tuples $\alpha$ and $\beta$ on relation scheme $R$ are value-equivalent (written $\alpha \cong \beta$ ) if and only if, for all $A 0 R,(A$ ÖpS) $\Rightarrow \alpha(\mathrm{A})=$ $\beta(\mathrm{A})$. Value-equivalent tuples are not allowed in a relation; they must be coalesced. The coalescence operation (denoted by $\oplus$ ) on two value-equivalent tuples $\alpha$ and $\beta$ is defined as:

$$
=\oplus \Leftrightarrow(\cong) \wedge(\cong) \wedge((\mathrm{pS})=\min \{1, \quad(\mathrm{pS})+(\mathrm{pS})\})
$$

The idea of value-equivalent tuples and the coalescence operation can be extended to more than two tuples by treating them in a pairwise fashion.

We are now ready to define a relation. Let $R$ be a relation scheme. A relation $r$ on the scheme $R$ is a finite collection of tuples $x$ on $R$ such that no two tuples in $r$ are value-equivalent. The three most common relational operations, namely projection, selection and join, are redefined as:

1. Projection. Let $r$ be a relation on scheme $R$, and let $S \mathrm{~d} R$. The projection of $r$ onto $S$ is defined as:

$$
\prod_{\mathrm{S}}(\mathrm{r})=\left\{(\mathrm{S}) \mid=\bigoplus_{\substack{\in \mathrm{r} \\(\mathrm{~S})=}}(\mathrm{S})\right\}
$$

2. Selection. Let $r$ be a relation on scheme $R$. Let $\Theta$ be a set of comparators over domains of attribute names in $R$. Let $P$ be a predicate (called the selection predicate) formed by attributes in $R$, comparators in $\Theta$, constants in the domain of $A \in R$, and logical connectives. The selection on $r$ for $P$, written $\sigma_{P}(r)$, is the set $\{\alpha \in r \mid P(\alpha)\}$.
3. Natural Join. Let $r$ and $s$ be any two relations on schemes $R$ and $S$ respectively, and let $R^{\prime}=R-\{p S\}$ and $S^{\prime}=S-\{p S\}$. The natural join of $r$ and $s$ is defined as:
$\left.\mathrm{r} \triangleright \triangleleft \mathrm{s}=\left\{(\mathrm{R} \cup \mathrm{S}) \exists \in \mathrm{F} \exists \in \mathrm{s}\left(\left(\mathrm{R}^{\prime}\right)=\left(\mathrm{R}^{\prime}\right)\right) \wedge\left(\left(\mathrm{S}^{\prime}\right)=\left(\mathrm{S}^{\prime}\right)\right) \wedge((\mathrm{pS})=(\mathrm{pS})(\mathrm{pS}))\right)\right\}$

The attributes in $R$ and $S$ should be stochastically independent for the natural join operation to yield meaningful results.

## 3. THE BAYESIAN FRAMEWORK

Let us first examine the Bayesian framework for belief revision. Let $B_{i}, i=1,2, \ldots, n$, be a set of exhaustive and mutually exclusive propositions. The degree of belief in another proposition $C$, which is dependent on the propositions $B_{i}$ can be written, using Bayes' conditionalization formula, as:

$$
\operatorname{pr}[\mathrm{C}]=\sum_{\mathrm{i}=1}^{\mathrm{n}} \operatorname{pr}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}\right] \operatorname{pr}\left[\mathrm{B}_{\mathrm{i}}\right] .
$$

Jeffrey (1983) addresses the problem of updating the degree of belief in $C$ when some new evidence (or some passage of an experience) $e$ is obtained. It is assumed that $e$ does not affect the probability of $C$ directly, neither does it change the conditional degree of belief in $C$ given $B_{i}$. In other words, if this new degree of belief is denoted by PR, then

$$
\operatorname{PR}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}\right]=\operatorname{pr}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}\right] .
$$

However, if $e$ changes the degree of belief in $B_{i}$, it is clear that it should (indirectly) affect the degree of belief in $C$. The revised degree of belief in $C$ is then written as (Jeffrey's rule of probability kinematics):

$$
\operatorname{PR}[\mathrm{C}]=\sum_{\mathrm{i}=1}^{\mathrm{n}} \operatorname{pr}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}\right] \operatorname{PR}\left[\mathrm{B}_{\mathrm{i}}\right] .
$$

Pearl (1988) illustrates this using the following example. Assume that an agent is uncertain about the color of a cloth, and is considering the propositions that the cloth is green, blue or violet. These three propositions are denoted as $G, B$ and $V$ respectively. Let the agent's current degrees of belief in these propositions be given by:

$$
\operatorname{pr}[G]=0.3, \operatorname{pr}[B]=0.3, \operatorname{pr}[V]=0.4
$$

Now, consider a different proposition $C$ that the cloth will be sold the next day. If we make the additional assumption that the chances of selling the cloth depend on its color in the following manner:

$$
\operatorname{pr}[C \mid G]=0.4, \operatorname{pr}[C \mid B]=0.4, \operatorname{pr}[C \mid V]=0.8
$$

then the agent's belief of the cloth selling the next day should be given by:

$$
\operatorname{pr}[C]=(0.4)(0.3)+(0.4)(0.3)+(0.8)(0.4)=0.56
$$

Now consider the case that the agent gets to inspect the cloth by candlelight, and subsequently revises the belief about the color of the cloth as:

$$
\operatorname{PR}[G]=0.70, \operatorname{PR}[B]=0.25, \operatorname{PR}[V]=0.05
$$

Since the dependence of saleability of the cloth on its color ought not to have changed, the agent's revised belief of the cloth selling the next day is given by:

$$
\operatorname{PR}[C]=(0.4)(0.70)+(0.4)(0.25)+(0.8)(0.05)=0.42
$$

Pearl (1988) argues that, based on the new evidence $e$, the update should be based on the conditionalization formula:

$$
\operatorname{pr}[\mathrm{C} \mid \mathrm{e}]=\sum_{\mathrm{i}=1}^{\mathrm{n}} \operatorname{pr}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}, \mathrm{e}\right] \operatorname{pr}\left[\mathrm{B}_{\mathrm{i}} \mid \mathrm{e}\right]
$$

and, in the case where $\operatorname{pr}\left[C \mid B_{i}, \mathrm{e}\right]=\operatorname{pr}\left[C \mid B_{i}\right]$, simply as:

$$
\operatorname{pr}[\mathrm{C} \mid \mathrm{e}]=\sum_{\mathrm{i}=1}^{\mathrm{n}} \operatorname{pr}\left[\mathrm{C} \mid \mathrm{B}_{\mathrm{i}}\right] \operatorname{pr}\left[\mathrm{B}_{\mathrm{i}} \mid \mathrm{e}\right],
$$

It is clear, in that case, that $\operatorname{PR}\left[B_{i}\right]=\operatorname{pr}\left[B_{i} \mid e\right], i=1,2, \ldots, n$. It is possible to argue that the evidence $e$ may not always be expressible as a proposition ${ }^{2}$ and Jeffrey's rule is, in that sense, more general than the above conditionalization formula. However, we will not take part in this argument; for our purposes, either notation provides a satisfactory working formula.

Let us now examine how this can be used in updating belief about objects in a database. Consider the following relation:

| $\underline{K}$ | $X$ | $Y$ | $p S$ |
| :---: | :---: | :---: | :---: |
| $k_{1}$ | $x_{1}$ | $y_{1}$ | 0.5 |
| $k_{1}$ | $x_{1}$ | $y_{2}$ | 0.5 |
| $k_{2}$ | $x_{2}$ | $y_{1}$ | 0.3 |

It implies that

$$
\begin{aligned}
& \operatorname{pr}\left[\mathrm{K}=\mathrm{k}_{1}, \mathrm{X}=\mathrm{x}_{1}, \mathrm{Y}=\mathrm{y}_{1}\right]=0.5, \\
& \operatorname{pr}\left[\mathrm{~K}=\mathrm{k}_{1}, \mathrm{X}=\mathrm{x}_{1}, \mathrm{Y}=\mathrm{y}_{2}\right]=0.5, \\
& \operatorname{pr}\left[\mathrm{~K}=\mathrm{k}_{2}, \mathrm{X}=\mathrm{x}_{2}, \mathrm{Y}=\mathrm{y}_{1}\right]=0.3 .
\end{aligned}
$$

Now assume that some new information (passage of a new experience) makes us change our belief about the marginal probability of [ $K=k_{1}, X=x_{1}$ ] to 0.8 . In other words, we have

$$
\operatorname{PR}\left[K=k_{1}, X=x_{1}\right]=0.8 .
$$

How should this new information affect the first two tuples in the previous relation? If we assume that the conditional distribution of $Y$ given $K$ and $X$ has not changed, i.e., if

$$
\operatorname{PR}\left[Y=y_{1} \mid K=k_{1}, X=x_{1}\right]=\operatorname{pr}\left[Y=y_{1} \mid K=k_{1}, X=x_{1}\right],
$$

then, using Jeffrey's rule of probability kinematics, we have

$$
\operatorname{PR}\left[K=k_{1}, X=x_{1}, Y=y_{1}\right]=\operatorname{pr}\left[Y=y_{1} \mid K=k_{1}, X=x_{1}\right] \times \operatorname{PR}\left[K=k_{1}, X=x_{1}\right]=0.5 \times 0.8=0.4 .
$$

Similarly,

$$
\operatorname{PR}\left[K=k_{1}, X=x_{1}, Y=y_{2}\right]=0.4,
$$

and the resulting new relation would reflect the revised belief as:

| $\underline{K}$ | $X$ | $Y$ | $p S$ |
| :---: | :---: | :---: | :---: |
| $k_{1}$ | $x_{1}$ | $y_{1}$ | 0.4 |
| $k_{1}$ | $x_{1}$ | $y_{2}$ | 0.4 |
| $k_{2}$ | $x_{2}$ | $y_{1}$ | 0.3 |

${ }^{2}$ For instance, the agent's experience after the observation by candlelight in the above example cannot strictly be expressed in terms of a proposition that can be taken into evidence.

The question that arises is how justified are we in assuming that the conditional distribution of the unspecified attribute $Y$, with respect to the attributes $K$ and $X$, has not changed. We conclude that it is the only practical assumption to make in absence of further information. If the user were to know how the conditional distribution has changed, then the information would not be provided only in terms of $K$ and $X$; it should be specified as a joint distribution involving $K, X$ and $Y$. So we must assume that the user does not know about the nature of the changed conditional distribution. In that case, using any other value for the dependence of $Y$ on $K$ and $X$ clearly has no basis and one should persist with the old information about the dependence in updating the relation. Thus, the use of stale information in the absence of fresh information is only a practical course of action.

It must be emphasized that we use the notation $e$ in a more general sense than the conventional Bayesian framework (see Figure 1). In that framework, the belief about the dependent proposition $C$ at a specific point in time is computed based on new information about the propositions $B_{i}$. In the above example (observation by candlelight, as it has come to be called), we were only considering the salability of the cloth at a specific time; it is understood that it might change with time and the temporal behavior was not considered relevant.


Figure 1. Sources of Belief Revision for a Database

In database systems, however, the temporal nature of objects and their properties is important. A database stores probabilistic information about real world objects, the probabilities indicating the current best estimates of our belief in the properties of these objects. There are two ways in which these estimates may change. First, an event may have recently taken place in the real world that forces some objects to change their states. This implies that the database must reflect a change in our beliefs about those objects. For example, if the government imposes heavier taxes on tobacco products, we must revise our belief about profitability of R. J. Reynolds Tobacco Company or the risk associated with investment in that company. On the other hand, even though an object may persist in its original state, some new information may have become available which may change our belief about it. For example, a better financial analysis may reveal that the riskiness of a particular investment is actually lower than it was perceived earlier. We use $e$ as a joint notation for real world events as well as evidences. While the evidences typically do not change the conditional dependence of $Y$ on $X$, the events might do that. However, if fresh information to the contrary is not explicitly provided, we will assume that the conditional distribution has not changed. We formalize this idea in the next section by extending it to the case where $X$ and $Y$ are not just attributes, rather they are sets of attributes.

## 4. THE BELIEF REVISION SCHEME

Let $R$ be a probabilistic relation scheme. We can, in general, express $R$ as $R=K X Y \cup\{p S\}$, where $K, X, Y$ are sets of attributes, and $p S$ is the probability stamp. $K$ denotes the primary key of the relation scheme. Let $r$ be a relation on $R$. We consider arrival of new information about object $k$ in $r$ in the form:

$$
\operatorname{PR}\left[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}_{\mathrm{i}}\right]=\mathrm{p}_{\mathrm{i}}, \quad \mathrm{i}=1,2, \ldots, \mathrm{n} . .
$$

Clearly, for this information to be meaningful, $p_{\mathrm{i}}$ 's must add up to no more than one and $x_{\mathrm{i}}$ 's must be distinct; i.e., $x_{i}=x_{j}$ implies $i=j$, for all $i, j=1,2, \ldots, n$. Often, users may not be able to provide new information as a complete distribution for attributes $K$
and $X$. Therefore, we allow the new information to be an incomplete specification of the distribution. Irrespective of whether this new information is complete or otherwise, the database must be revised in a manner that (i) is consistent with the new information and (ii) results in assigning probabilities to unspecified realizations of the stochastic variables in a manner consistent with the existing data.

As in the previous section, we will use the following notation. All new information, as well as revised beliefs, will be denoted as "PR," whereas old beliefs will be represented as "pr."

In order to revise beliefs about attributes of an object, we need to consider the following three cases:

- Case 1: There exists tuple $\langle k, x, y, q\rangle$ for some $q \in(0,1]$, such that $x=x_{i}$, for some $i 0\{1,2, \ldots, n\}$. This is the case where an existing tuple matches a tuple in the new information on the attributes $K$ and $X$.
- Case 2: There exists $\langle k, x, y, q\rangle \in r$, for some $q \in(0,1]$, such that $x \ddot{O}_{i}$, for all $i 0\{1,2, \ldots, n\}$. In this case, a tuple in the existing relation has no matching tuple in the new information.
- Case 3: There exists some $i 0\{1,2, \ldots, n\}$, such that the tuple $\left\langle k, x_{j}\right\rangle \notin \Pi_{K X}$. This would be the case when a tuple in the new information does not match with any tuple in the existing relation.


### 4.1 Case 1

Consider the tuple $\langle k, x, y, q\rangle 0 r$, for some $q \in(0,1]$. Let $x=x_{i}$, for some $i 0\{1,2, \ldots, n\}$. It is now straightforward to calculate the new probability $Q$ associated with the above tuple:

$$
\begin{aligned}
\mathrm{Q} & =\operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}, \mathrm{Y}=\mathrm{y}] \\
& =\operatorname{pr}[\mathrm{Y}=\mathrm{y} \mid \mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \times \operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \\
& =\frac{\operatorname{pr}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}, \mathrm{Y}=\mathrm{y}]}{\operatorname{pr}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}]} \times \operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \\
& =\frac{\mathrm{qp}_{\mathrm{i}}}{\prod_{\mathrm{pS}}\left({ }_{\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}}(\mathrm{r})\right)} .
\end{aligned}
$$

Clearly, a new tuple $\langle k, x, y, Q\rangle$ should replace the corresponding old tuple in $r$.
The assumption made here is that the conditional probability of $Y=y$ given $K=k$ and $X=x$ is unchanged given the new information (as discussed in the previous section). If we consider the new information to be the result of taking some experience (or event) $e$ into evidence, then it is assumed that

$$
\operatorname{pr}[\mathrm{Y}=\mathrm{y} \mid \mathrm{X}=\mathrm{x}, \mathrm{~K}=\mathrm{k}, \mathrm{E}]=\operatorname{pr}[\mathrm{Y}=\mathrm{y} \mid \mathrm{X}=\mathrm{x}, \mathrm{~K}=\mathrm{k}],
$$

that is, $Y$ and $e$ are conditionally independent of each other with respect to $K$ and $X$. However, since $\operatorname{pr}[K=k, X=x \mid e]$ is different from $\operatorname{pr}[K=k, X=x]$, the revised value of $\operatorname{pr}[K=k, X=x, Y=y \mid e]$ is also different from $\operatorname{pr}[K=k, X=x, Y=y]$. The net outcome is that probability masses in the original relation that are assigned to different values of $Y$ for given $K$ and $X$ are pro-rated such that the marginal probability of $K=k$ and $X=x$ in the revised database matches the new probability $\operatorname{pr}[K=k, X=x \mid e]$.

### 4.2 Case 2

Again, consider a tuple $\langle k, x, y, q\rangle 0 r$, for some $q \in(0,1]$. In this case, no new information is available on tuples with $X=x$. The total probability currently assigned to object $k$ is

## Dey and Sarkar

$$
\operatorname{pr}[\mathrm{K}=\mathrm{k}]=\Pi_{\mathrm{pS}}\left({ }_{\mathrm{K}=\mathrm{k}}(\mathrm{r})\right)=\mathrm{P}_{\mathrm{k}} \text { (say). }
$$

The new information, on the other hand, assigns a probability of $\sum_{i=1}^{n} p_{i}$ to the existence of object $k$. We assume that the revised distribution assigns a value of $\max \left\{P_{k}, \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{i}}\right\}$ to object $k$. If $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{i}} \geq \mathrm{P}_{\mathrm{k}}$, then the entire probability mass assigned to object $k$ is specified in the new distribution. Therefore, $\operatorname{PR}[K=k, X=x]$ becomes zero, and the corresponding tuples should be just eliminated from $r$. However, if $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{i}}<\mathrm{P}_{\mathrm{k}}$, then $\mathrm{PR}[K=k]$ retains the value $P_{k}$. In that case, $\mathrm{PR}[K=k, X=x]$ cannot be computed directly, since part of the probability mass is unspecified. It is, however, clear that $\operatorname{PR}[K=k, X=x]$ will be a fraction of the unspecified probability mass, i.e., $\left(P_{k}-\sum_{i=1}^{n} p_{i}\right)$

We recommend that this unspecified mass be distributed in a way such that the new assignment is proportional to the older assignment of probabilities. We see that the missing probability mass of $\left(\mathrm{P}_{\mathrm{k}}-\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{i}}\right)$ has to be assigned to values other than $x_{i}, i=1,2, \ldots, n$. The corresponding probability mass that was assigned to these values in the older distribution is

$$
\left.\mathrm{P}_{\mathrm{k}}-\sum_{\mathrm{i}=1}^{\mathrm{n}} \Pi_{\mathrm{pS}}\left(\mathrm{~K}=\mathrm{k}, \mathrm{X}=\mathrm{x}_{\mathrm{i}} \mathrm{r}\right)\right)
$$

We also know that the old probability is given by:

$$
\operatorname{pr}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}]=\Pi_{\mathrm{pS}}(\underset{\mathrm{~K}=\mathrm{k}, \mathrm{X}=\mathrm{x}}{ }(\mathrm{r})) .
$$

The new probability can now be calculated by appropriate normalization as:

$$
\operatorname{PR}[K=k, X=x]=\Pi_{p S}(\quad K=k, X=x=(r)) \times \frac{P_{k}-\sum_{i=1}^{n} p_{i}}{P_{k}-\sum_{i=1}^{n} \prod_{p S}\left({ }_{K=k, X=x_{i}}(r)\right)} .
$$

The calculation of the new probability $Q$ associated with $\langle k, x, y\rangle$ is now easy:

$$
\begin{aligned}
& \mathrm{Q}=\operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}, \mathrm{Y}=\mathrm{y}] \\
& =\operatorname{pr}[\mathrm{Y}=\mathrm{y} \mid \mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \times \operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \\
& =\frac{\operatorname{pr}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}, \mathrm{Y}=\mathrm{y}]}{\operatorname{pr}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}]} \times \operatorname{PR}[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}] \\
& =\frac{q}{\Pi_{p S}\left({ }_{K=k, X=x}(r)\right)} \times \Pi_{p S}(\quad K=k, X=x(r)) \times \frac{P_{k}-\sum_{i=1}^{n} p_{i}}{P_{k}-\sum_{i=1}^{n} \Pi_{p S}(K=k, z} \\
& =\mathrm{q} \times \frac{\Pi_{\mathrm{pS}}\left({ }_{\mathrm{K}=\mathrm{k}}(\mathrm{r})\right)-\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}_{\mathrm{i}}}{\Pi_{\mathrm{pS}}(\quad \mathrm{~K}=\mathrm{k}(\mathrm{r}))-\sum_{\mathrm{i}=1}^{\mathrm{n}} \Pi_{\mathrm{pS}}\left({ }_{\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}_{\mathrm{i}}}(\mathrm{r})\right)}
\end{aligned}
$$

The old tuple $\langle K, x, y, q\rangle$ should then be deleted from the relation, and, if $Q>0$, a new tuple $\langle k, x, y, Q\rangle$ should be included in $r$.
The above scheme can also be explained in terms of an assumption of conditional independence. Let $Q>0$ be the set of values of attributes of $X$ that appear in the original relation for $K=k$, but are not specified as part of the new information. Then the assumption made here is that

$$
\operatorname{pr}\left[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}_{\mathrm{j}} \mid \widetilde{\mathrm{X}}, \mathrm{e}\right]=\operatorname{pr}\left[\mathrm{K}=\mathrm{k}, \mathrm{X}=\mathrm{x}_{\mathrm{j}} \mid \widetilde{\mathrm{X}}\right]
$$

where $\mathrm{x}_{\mathrm{j}} \in \widetilde{\mathrm{X}}$. Thus, for lack of any other information, we assume that the probability of any of the outcomes $x_{j}$ given $\widetilde{\mathrm{X}}$ is unchanged in the new distribution. Once the revised beliefs are obtained for $\mathrm{x}_{\mathrm{j}} \in \widetilde{\mathrm{X}}$, the beliefs for $\left[K=k, Y=y, X=x_{j}\right]$ are updated in a fashion similar to Case 1.

### 4.3 Case 3

If there is new information in the form $\operatorname{PR}\left[X=x_{j} \mid K=k\right]=p_{j}$, where $\left\langle\mathrm{k}, \mathrm{x}_{\mathrm{j}}\right\rangle \notin \Pi_{\mathrm{KX}}(\mathrm{r})$ for some $j 0\{1,2 \ldots n\}$, this new information must be incorporated by adding a new tuple $\left\langle k, x_{j}{ }^{*}, p_{j}\right\rangle$ to $r$.

The above belief revision scheme is simple and can be easily implemented. The algorithm in Figure 2 shows how the revision schemes can be embedded in a simple procedure that modifies an existing database given new information.

```
                    Algorithm: Belief Revision
input: relation }r\mathrm{ on scheme R; PR[K=k,X=x i}]=\mp@subsup{p}{i}{},i=1,2,\ldots,n
output: revised relation r.
BEGIN
for i:=1 to n do
    m
for all }\alpha\in\mp@subsup{\sigma}{\textrm{K}=\textrm{k}}{(}(r)\mathrm{ do
    begin
            found := false;
            r:= r-";
            for i:=1 to n do
                if }\alpha(X)=\mp@subsup{x}{i}{}\mathrm{ then
                begin
                    calculate Q from Case 1;
                    m
                            found := true;
                    end;
            if not(found) then
                calculate Q from Case 2;
            if Q>0 then
            r:=r\cup\langlek, " (X), " (Y),Q\rangle;
        end;
for j:=1 to n do
    if m
            r:=ru\k, " (X), " (Y), Q \;
END.
```

Figure 2. Complete Algorithm for Belief Revision in a Probabilistic Relation

## Dey and Sarkar

## 5. EXAMPLES

In this section, we show how the proposed algorithm works and demonstrate that it provides intuitively meaningful results. For the purposes of illustration, we will make use of the probabilistic employee relation shown in Table 3 (obtained by projecting the relevant attributes from the relation in Table 2). In this table, we have probabilistic information about three different employees. Let us consider the following examples and the associated belief revisions.

### 5.1 Example 1

Let the new information be:

$$
\text { PR[EMP\#=6723] }=0.45 \text {. }
$$

Table 3. A Probabilistic EMPLOYEE Relation

| EMP\# | rank | salary | dept | $p S$ |
| :---: | :--- | :---: | :---: | :---: |
| 3025 | cashier | 20 K | shoe | 0.6 |
| 3025 | clerk | 15 K | toy | 0.2 |
| 3025 | cashier | 15 K | auto | 0.2 |
| 6723 | clerk | 18 K | toy | 0.4 |
| 6723 | cashier | 20 K | auto | 0.4 |
| 6723 | $*$ | $*$ | $*$ | 0.1 |
| 6879 | clerk | 25 K | toy | 0.3 |
| 6879 | clerk | $*$ | toy | 0.1 |
| 6879 | cashier | $*$ | $*$ | 0.6 |

This information implies that the probability of existence of an employee with EMP\#=6723 is 0.45 . Note that $X=\varnothing$ in this case. The revision is straightforward; all tuples in the existing relation fall under Case 1 . The total probability of 0.45 should now be proportionately distributed across the existing tuples. The resulting relation is shown in Table 4.

Table 4. Revised EMPLOYEE Relation for EMP\#=6723

| EMP\# | rank | salary | dept | $p S$ |
| :---: | :---: | :---: | :---: | :---: |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 6723 | clerk | 18 K | toy | 0.2 |
| 6723 | cashier | 20 K | auto | 0.2 |
| 6723 | $*$ | $*$ | $*$ | 0.05 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

### 5.2 Example 2

Let the new information be:

$$
\text { PR[EMP\#=3025,rank=clerk] }=0.6 \text {. }
$$

This information will affect the tuples in the old relation in two ways. There are three tuples in this relation satisfying EMP\#=3025; the second tuple falls under Case 1, whereas the first and the third tuples fall under Case 2. It is clear that the second tuple should have a $p S$ value of 0.6 . For the other two tuples, note that the probability of rank="cashier" for this employee is no longer 0.8 . In absence of better information, we assume that the residual probability of 0.4 is assigned to it. Then, this probability mass 0.4 is proportionately distributed over the other two tuples (based on the old distribution). The resulting relation is shown in Table 5.

Table 5. Revised EMPLOYEE Relation for EMP\#=3025

| EMP\# | rank | salary | dept | $p S$ |
| :---: | :--- | :---: | :---: | :---: |
| 3025 | cashier | 20 K | shoe | 0.3 |
| 3025 | clerk | 15 K | toy | 0.6 |
| 3025 | cashier | 15 K | auto | 0.1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

### 5.3 Example 3

It is possible that the new information has attribute values that were not explicitly listed in the relation (Case 3). For example, consider the following new information:

$$
\begin{aligned}
& \text { PR[EMP\#=6723,rank=packer] }=0.45, \\
& \text { PR[EMP\#=6723,rank=clerk] }=0.2 .
\end{aligned}
$$

This information will affect the tuples in the old relation in three different ways. The first among the existing tuples for EMP\#=6723 falls under Case 1, while the second and the third fall under Case 2. Also, based on the new information, a new tuple has to be created according to Case 3. The resulting relation is shown in Table 6.

Table 6. Revised EMPLOYEE Relation for EMP\#=6723

| EMP\# | rank | salary | dept | $p S$ |
| :---: | :---: | :---: | :---: | :---: |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 6723 | clerk | 18 K | toy | 0.20 |
| 6723 | cashier | 20 K | auto | 0.20 |
| 6723 | $*$ | $*$ | $*$ | 0.05 |
| 6723 | packer | $*$ | $*$ | 0.45 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

### 5.4 Example 4

When information about the joint distribution on all attributes is available, the revision is straightforward. In this case, $Y$ is a null set of attributes. Again the same algorithm can be used for this special case. Consider, as an example, the case where the new information has the form:

$$
\begin{aligned}
& \text { PR[EMP\#=6879, rank=clerk, salary }=20 \mathrm{~K}, \text { dept=auto }]=0.3 \text {, } \\
& \text { PR[EMP\#=6879, rank=packer, salary }=28 \mathrm{~K}, \text { dept }=\text { toy }]=0.2 \text {. }
\end{aligned}
$$

These two new tuples, according to Case 3, must be appended to the existing relation. Moreover, this information will also change the probabilities of the tuples in the old relation; all existing tuples fall under Case 2 . The revised relation is shown in Table 7.

Table 7. Revised EMPLOYEE Relation for EMP\#=6879

| EMP\# | rank | salary | dept | $p S$ |
| :---: | :---: | :---: | :---: | :---: |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 6879 | clerk | 20 K | auto | 0.30 |
| 6879 | packer | 28 K | toy | 0.20 |
| 6879 | clerk | 25 K | toy | 0.15 |
| 6879 | clerk | $*$ | toy | 0.05 |
| 6879 | cashier | $*$ | $*$ | 0.30 |

## 6. CONCLUSIONS

Although relational databases enjoy a very widespread popularity in modern business information systems, they lack the power to model uncertainty in data items. Several extensions of the relational model have been proposed in the literature to this end. However, these extensions overlook a very important aspect of a data model, namely modification of data items. In a probabilistic database, where uncertain data items are stored along with the users' degree of belief in them, modification of data is equivalent to revision of belief about the data. In this paper, we present a simple scheme for revision of belief in a probabilistic database based on the Bayesian framework.

We discuss the basic assumptions in applying the Bayesian framework, and formally describe an algorithm that updates the existing relation based on the user's input. Modification of data in a probabilistic database could arise from two sources: (i) change of state of one or more data items due to some real world event and (ii) change of degree of belief about one or more data items due to new (better) information. Our revision scheme is general enough to accommodate both types of modifications.

Our revision scheme considers only point estimate interpretation of probabilistic relations. The ideas presented in this paper may be extended to other interpretations as well. The modification of a database must address the issue of data redundancy and normalization. Dependency and normalization theories are well understood for conventional databases; we are exploring how they can be generalized for probabilistic databases.

There are several directions for future research. For example, new information may affect the belief in objects and their attributes stored across several relations. In order to handle revision of belief that may be propagated across multiple relations, the notions of probabilistic dependency and integrity constraints will be examined. Furthermore, implementation issues such as physical storage of records and optimal access paths need to be addressed for this scheme.

## 7. REFERENCES

Barbará, D.; Garcia-Molina, H.; and Porter, D. "The Management of Probabilistic Data." IEEE Transactions on Knowledge and Data Engineering, Volume 4, Number 5, October 1992, pp. 487-502.

Cavallo, R., and Pittarelli, M. "The Theory of Probabilistic Databases." Proceedings of the Thirteenth Very Large Data Base Conference, Brighton, 1987, pp. 71-81.

Codd, E. F. "Extending the Database Relational Model to Capture More Meaning." ACM Transactions on Database Systems, Volume 4, Number 4, 1979, pp. 397-434.

Codd, E. F. "A Relational Model of Data for Large Shared Data Banks." Communication of the ACM, Volume 13, Number 6, June 1970, pp. 377-387.

Date, C. J. Relational Database: Selected Writings. Reading, Massachusetts: Addison-Wesley, 1986.
Dey, D., and Sarkar, S. "A Probabilistic Relational Model and Algebra." ACM Transactions on Database Systems, to appear in Volume 21, Number 3, September 1996.

Jaynes, E. T. "Prior Probabilities." IEEE Transactions on Systems Science and Cybernatics, Volume SSC-4, Number 3, September 1968, pp. 227-241.

Jeffrey, R. The Logic of Decision. Chicago: University of Chicago Press, 1983.
Motro, A. "Accommodating Imprecision in Database Systems: Issues and Solutions." Database Engineering, Volume 9, 1990, pp. 213-218.

Pearl, J. "Fusion, Propagation, and Structuring in Belief Networks." Artificial Intelligence, Volume 29, 1986, pp. 241-288.
Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. New York: Morgan Kaufmann, 1988.

Pearl, J. "Probabilistic Semantics for Nonmonotonic Reasoning: A Survey." In R. J. Brachman, N. J. Levesque, and R. Reiter, Editors, Proceedings of the First Conference on Principles of Knowledge Representation and Reasoning. New York: Morgan Kaufmann, 1989, pp. 505-516.

Raju, K. V. S. V. N., and Majumdar, A. K. "Fuzzy Functional Dependencies and Lossless Join Decomposition of Fuzzy Relational Database Systems." ACM Transactions on Database Systems, Volume 13, Number 2, June 1988, pp. 129-166.

