Integration of Probabilistic Information

Fereidoon Sadri(✉) and Gayatri Tallur

Department of Computer Science, University of North Carolina,
Greensboro, NC, USA
f_sadri@uncg.edu

Abstract. We study the problem of data integration from sources that contain probabilistic uncertain information. Data is modeled by possible-worlds with probability distribution, compactly represented in the probabilistic relation model. Integration is achieved efficiently using the extended probabilistic relation model. We study the problem of determining the probability distribution of the integration result. It has been shown that, in general, only probability ranges can be determined for the result of integration. We show that under intuitive and reasonable assumptions we can determine the exact probability distribution of the result of integration. Our methodologies are presented in possible-worlds as well as probabilistic-relation frameworks.

Keywords: Data integration · Probabilistic data · Uncertain data · Probabilistic relation model

1 Introduction

Information integration and modeling and management of uncertain information have been active research areas for decades, with both areas receiving significant renewed interest in recent years (*e.g.*, [8,11]). The importance of information integration *with uncertainty*, on the other hand, has been realized more recently (*e.g.*, [2,12]). In a world with ever increasing data generated by both humans and machines alike, the field of computer science has seen a transition from computation-intensive applications to data-intensive ones. Most of the data, in particular data discovered through data mining and knowledge discovery, is uncertain. Hence, integration of *uncertain* data has become a necessity for many modern applications. It has been observed that "While in traditional database management managing uncertainty and lineage seems like a nice feature, in data integration it becomes a necessity" [12].

The widely accepted conceptual model for uncertain data is the possible-worlds model [1]. For practical applications, a representation of choice is the probabilistic relation model [10], which provides a compact and efficient representation for uncertain data. It has been shown that integration of uncertain data represented in the probabilistic relation model can be achieved efficiently using the extended probabilistic relation model [6].

M.A. Cheema et al. (Eds.): ADC 2016, LNCS 9877, pp. 179–190, 2016.
DOI: 10.1007/978-3-319-46922-5_14

In this paper we concentrate on the integration of *probabilistic* uncertain data. Integration in the probabilistic relation framework is the most efficient approach but this approach faces challenges when probabilities are included. There is no clear way to associate probabilities with extended probabilistic relations (unlike probabilistic relations). Further, it has been shown that, even in the possible-worlds model, it is only possible to obtain probability *ranges* for the result of data integration [14]. In this paper we study the problem of determining the probability distribution of the integration result in the two main frameworks: The probabilistic possible-worlds model, and the probabilistic relation model. We show that, under intuitive and reasonable assumptions, we can determine the exact probability distribution of integration in either of the frameworks. Further, we show that the two approaches are equivalent while the probabilistic relation approach provides a significantly more efficient method in practice.

2 Preliminaries

Foundations of uncertain information integration were discussed in the seminal work of Agrawal *et al.* [2]. The goal of integration is to obtain the best possible uncertain database that contains all the information implied by sources, and nothing more. We presented an alternative integration approach in [14]. These approaches are based on the well-known *possible-worlds* model of uncertain information [1]. The possible-worlds model is widely accepted as the conceptual model for uncertain information, and is used as the theoretical basis for operations and algorithms on uncertain data. But it is not a suitable representation for the *implementation* of uncertain information systems due to lack of efficiency. Instead, compact representations, such as the *probabilistic relation model* [9,10], are more appropriate for the implementation. We also studied the problem of integration of information represented by probabilistic relations in [6], and presented efficient algorithms for the integration. In this section, we will review some of the observations and results from these works.

2.1 Integration Algorithm for Uncertain Data Represented in the Possible-Worlds Model

We begin with the following definition of *uncertain database* from [2].

Definition 1. (UNCERTAIN DATABASE) *An* uncertain database U *consists of a finite set of tuples* $T(U)$ *and a nonempty set of possible worlds* $PW(U) = \{D_1, \ldots, D_n\}$*, where each* $D_i \subseteq T(U)$ *is a certain database.*

We presented a logic-based approach to the representation and integration of uncertain data in the possible-world model in [14], and showed it was equivalent to the integration approach of [2]. Algorithm 1 below is an alternative integration algorithm. It is easy to show it is equivalent to the aforementioned algorithms. First, we need the following definition:

Definition 2. (Compatible possible-world relations). *Let S and S' be information sources with possible worlds $\{D_1, \ldots, D_n\}$ and $\{D'_1, \ldots, D'_{n'}\}$, respectively. Let T and T' be the tuple-sets of S and S'. A pair of possible-world relations D_i and D'_j are* compatible *if for each tuple $t \in T \cap T'$ either both D_i and D_j contain t, (*i.e., *$t \in D_i$ and $t \in D'_j$), or neither D_i nor D'_j contain t (*i.e., *$t \notin D_i$ and $t \notin D'_j$). Otherwise D_i and D'_j are not compatible.*

Given information sources S and S', the integration algorithm (Algorithm 1) considers all possible-world pairs from the two sources. If they are compatible, their union forms a possible-world of the integration.

Algorithm 1. Integration of uncertain data represented in the possible-worlds model

```
Given information sources $S$ and $S'$ with possible worlds $\{D_1, \ldots, D_n\}$ and
$\{D'_1, \ldots, D'_{n'}\}$ and tuple sets $T$ and $T'$
For every pair of possible-world relations $D_i \in S, D'_j \in S'$
    if $D_i$ and $D'_j$ are compatible then let $Q_{ij} = D_i \cup D'_j$
The possible-worlds model of the result of integrating $S$ and $S'$ has the set of possible-
world relations $Q_{ij}$ for every compatible pair $D_i$ and $D'_j$, and the tuple set $T \cup T'$.
```

3 Integration of Probabilistic Uncertain Data

3.1 Probabilistic Possible-Worlds Model

Definition 3. *A* probabilistic uncertain database *U consists of a finite set of tuples $T(U)$ and a nonempty set of possible worlds $PW(U) = \{D_1, \ldots, D_n\}$, where each $D_i \subseteq T(U)$ is a certain database. Each possible world D_i has a probability $0 < P(D_i) \leq 1$ associated with it, such that $\sum_{i=1}^{n} P(D_i) = 1$.*

Our goal is to integrate information from sources containing probabilistic uncertain data, and to compute the probability distribution of the possible-worlds of the result of the integration. It has been shown that, in general, exact probabilities of the result of integration can not be obtained [14]. Rather, only a *range* of probabilities can be computed for each possible world in the integration. In this paper, we show that, under intuitive and reasonable assumptions, it is possible to obtain exact probabilities for the result of integration.

3.2 Integration in the Probabilistic Possible-Worlds Framework

Let S and S' be sources with possible worlds $\{D_1, \ldots, D_n\}$ and $\{D'_1, \ldots, D'_{n'}\}$, respectively. Consider the bi-partite graph G defined by the relation (D_i, D'_j): D_i and D'_j are compatible (See Definition 2). The graph G is called the *compatibility graph* for sources S and S': There is an edge between D_i and D'_j if they are compatible. It has been shown that [14]

- Each connected component of G is a complete bipartite graph.
- Let H be a connected component of G. Then $\sum_{D_i \in H} P(D_i) = \sum_{D'_j \in H} P(D'_j)$. These conditions have been called *probabilistic constraints* in [14].

Probabilistic constraints are imposed by the semantics of probabilistic integration. But it is unlikely that they hold in practice. We regard these constraints as important means to adjust (or revise) the original probabilities of the sources when the constraints are violated [15]. Henthforth we assume the probabilities have been adjusted and probabilistic constraints hold.

Example 1. Consider the possible worlds of information sources S and S' shown in Figs. 1 and 2.

The compatibility bipartite graph G for the possible-world relations of these sources is shown in Fig. 3. Note that we have $P(D_1) + P(D_2) = P(D'_1) + P(D'_2)$ and $P(D_3) = P(D'_3) + P(D'_4)$ by the probabilistic constraints. □

student	course
Bob	CS100

D1

student	course
Bob	CS100
Bob	CS101

D2

student	course
Bob	CS101

D3

Fig. 1. Possible Worlds of source S

student	course
Bob	CS100

D'1

student	course
Bob	CS100
Bob	CS201

D'2

student	course
Bob	CS201

D'3

student	course
Bob	CS201
Bob	CS202

D'4

Fig. 2. Possible Worlds of source S'

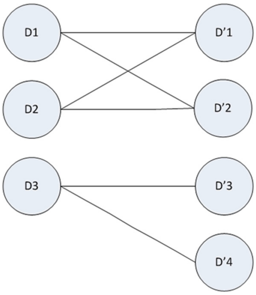

Fig. 3. Compatibility Graph for Example 1

There are 6 possible-world relations in the result of integration, corresponding to the two connected components of the compatibility graph. These 6 relations are shown in Fig. 6. Let us concentrate on the top connected component portion of the compatibility bipartite graph G. This connected component gives rise to 4 possible-world relations corresponding to $D_1 \wedge D'_1$, $D_1 \wedge D'_2$, $D_2 \wedge D'_1$, and $D_2 \wedge D'_2$. We want to compute the probabilities of these possible-world relations, $P(D_1 \wedge D'_1)$, $P(D_1 \wedge D'_2)$, $P(D_2 \wedge D'_1)$, and $P(D_2 \wedge D'_2)$, given the probability distribution of the possible worlds of the sources, $P(D_1), P(D_2), P(D'_1), p(D'_2)$.

We have four unknowns. We can write the following four equations:

$$\begin{aligned}
&P(D_1 \wedge D'_1) + P(D_1 \wedge D'_2) = P(D_1),\\
&P(D_2 \wedge D'_1) + P(D_2 \wedge D'_2) = P(D_2),\\
&P(D_1 \wedge D'_1) + P(D_2 \wedge D'_1) = P(D'_1),\\
&P(D_1 \wedge D'_2) + P(D_2 \wedge D'_2) = P(D'_2).
\end{aligned}$$

But, unfortunately, these equations are not independent. Note that the probabilistic constraint requires that $P(D_1) + P(D_2) = P(D'_1) + P(D'_2)$. Hence, any one of the 4 equations can be obtained from the other 3 using the probabilistic constraint. Hence we can only compute a probability range for each of these four possible-world relation.

So, how can we obtain exact probabilities for the possible-world relations of an integration? We make the following *partial independence assumption*.

Partial Independence Assumption: *The only dependencies among the probabilities of possible-world relations are those induced by probabilistic constraints.*

Armed with this intuitive and reasonable assumption, we are able to compute exact probabilities for the result of an integration. We use our example to explain the approach, then present the solution for the general case.

Example 2. Consider again the top connected component in the compatibility graph of Example 1. The structure of the graph tells us that if we have the evidence that the correct database of the first source S is D_1, then we know the correct database of the second source S' is either D'_1 or D'_2. Similarly, if we have the evidence that the correct database of the first source S is D_2, then we know the correct database of the second source S' is either D'_1 or D'_2. But, by the partial independence assumption, the knowledge of D_1 or D_2 does not influence the probability of D'_1. In other words, $P(D'_1 \mid D_1)$ is equal to $P(D'_1 \mid D_2)$. Since $P(D'_1 \wedge D_1) = P(D'_1 \mid D_1)P(D_1)$ and $P(D'_1 \wedge D_2) = P(D'_1 \mid D_2)P(D_2)$ we get

$$\frac{P(D_1 \wedge D'_1)}{P(D_2 \wedge D'_1)} = \frac{P(D'_1 \wedge D_1)}{P(D'_1 \wedge D_2)} = \frac{P(D_1)}{P(D_2)}$$

This serves as an additional equation that enables us to solve for the 4 unknowns. We get:

$$\begin{aligned}
&P(D_1 \wedge D'_1) = P(D_1)P(D'_1)/(P(D_1) + P(D_2))\\
&P(D_2 \wedge D'_1) = P(D_2)P(D'_1)/(P(D_1) + P(D_2))\\
&P(D_1 \wedge D'_2) = P(D_1)P(D'_2)/(P(D_1) + P(D_2))\\
&P(D_2 \wedge D'_2) = P(D_2)P(D'_2)/(P(D_1) + P(D_2))
\end{aligned}$$

The observations of the above example can be generalized. Let S_1 and S_2 contain information in probabilistic possible-worlds model. Consider a connected component G_1 of the compatibility bipartite graph G of S_1 and S_2. Let

$D_1, \ldots, D_m$ and $D'_1, \ldots, D'_{m'}$ be the nodes of G_1 corresponding to possible worlds of S_1 and S_2, respectively. We can write the following $m + m'$ equations:

$$\sum_{j=1}^{m'} P(D_i \wedge D'_j) = P(D_i), i = 1, \ldots, m; \text{ and } \sum_{i=1}^{m} P(D_i \wedge D'_j) = P(D'_j), j = 1, \ldots, m'$$

But $m + m' - 1$ of these equations are independent. Any one can be obtained from the rest using the probabilistic constraint $\sum_{i=1}^{m} P(D_i) = \sum_{j=1}^{m'} P(D'_j)$. On the other hand, we have $m \times m'$ unknowns $P(D_i \wedge D'_j), i = 1, \ldots, m, j = 1, \ldots, m'$. Additional equations are obtained from the independence assumption $\frac{P(D_1 \wedge D'_j)}{P(D_i \wedge D'_j)} = \frac{P(D_1)}{P(D_i)}$. It can be shown that $(m - 1) \times (m' - 1)$ of these equations are independent. Together with the $m + m' - 1$ equations of the first group we have the needed $m \times m'$ equations to solve for the unknowns. The solutions are,

$$P(D_i \wedge D'_j) = \frac{P(D_i)P(D'_j)}{P}$$

where P is the probabilistic constraint constant $P = \sum_{i=1}^{m} P(D_i) = \sum_{j=1}^{m'} P(D'_j)$.

4 Integration in the Probabilistic Relation Framework

A number of models have been proposed for the representation of uncertain information such as the "maybe" tuples model [7,13], set of alternatives or block-independent disjoint model (BID) [4,5], the probabilistic relation model [9,10], and the U-relational database model [3]. The probabilistic relation model has been widely accepted for compact representation of uncertain and probabilistic data. It is a *complete* model: Any uncertain data in the (probabilistic) possible-worlds model can be represented in an equivalent probabilistic relation [10]. Intuitively, this representation is based on the relational model where each tuple t is associated with a propositional logic formula $f(t)$ (called an *event* in [9].) The Boolean variables in the formulas are called *event variables*. A probabilistic relation r represents the set of possible-world relations corresponding to truth assignments to the set of event variables. A truth assignment μ defines a possible-world relation $r_\mu = \{t \mid t \in r \text{ and } f(t) = \mathit{true} \text{ under } \mu\}$.

In the previous section we presented an approach for the integration of probabilistic uncertain data in the probabilistic possible-worlds framework. As mentioned earlier, the possible-worlds framework is not suitable for practical applications. The size of the input, namely the possible-worlds relations, can be exponential in the size of the equivalent representation in the probabilistic relation framework. Further, we have a very efficient integration algorithm in the probabilistic relation framework. In this section we concentrate on the problem of determining the probability distribution for the integration result in the probabilistic relation framework.

4.1 Probabilistic Data

A probabilistic relation can represent *probabilistic* possible-worlds data by associating probabilities with event variables. Let r be a probabilistic relation. We can compute the probabilities associated with possible-world relations represented by r as follows. Let $V = \{a_1, a_2, \ldots, a_k\}$ be the set of event variables of r. Let μ be a truth assignment to event variables. μ defines a relation instance $r_\mu = \{t \mid t \in r \text{ and } f(t) = \mathit{true} \text{ under } \mu\}$. The probability associated with r_μ is $\prod_{\mu(a_j)=true} P(a_j) \prod_{\mu(a_j)=false}(1-P(a_j))$. Note that this formula is based on the assumption that event variables are independent of each other. A possible-world relation r_i of r can result from multiple truth assignments to event variables, in which case the probability of r_i, $P(r_i)$ is the sum of probabilities of r_μ for all truth assignments μ that generate r_i.

Example 3. Consider the possible worlds of information sources S and S' from Example 1, shown in Figs. 1 and 2. Assume the probability distributions are $P(D_1) = 0.3$, $P(D_2) = 0.5$, $P(D_3) = 0.2$, $P(D'_1) = 0.35$, $P(D'_2) = 0.45$, $P(D'_3) = 0.05$, and $P(D'_4) = 0.15$. Algorithms for producing probabilistic relations for uncertain probabilistic databases have been presented in [6,10]. We have used the algorithm of [6] to obtain the probabilistic relations r_1 and r_2 of Fig. 4 for the uncertain probabilistic databases of Figs. 1 and 2. In Fig. 4, b_1, b_2, b_3, c_1 and c_2 are event variables, and column E records the event formulas associated with each tuple. Probabilities of the event variables are also computed by the algorithm and are: $P(b_1) = 0.35$, $P(b_2) = \frac{9}{13}$, $P(b_3) = 0.25$, $P(c_1) = 0.2$, and $P(c_2) = 0.625$. □

student	course	E
Bob	CS100	$\neg c_1$
Bob	CS101	$c_1 \vee c_2$

pr-relation r

student	course	E
Bob	CS100	$b_1 \vee b_2$
Bob	CS201	$\neg b_1$
Bob	CS202	$\neg b_1 \wedge \neg b_2 \wedge \neg b_3$

pr-relation r'

Fig. 4. Probabilistic relations for sources S and S' of Example 1

4.2 Integration of Uncertain Data Represented in the Probabilistic Relation Model

As mentioned earlier, for efficiency reasons a compact representation of uncertain data is utilized in practice. We will summarize an algorithm for the integration of uncertain data represented in the probabilistic relation model from [6]. First we need the following definition.

Definition 4. *An* extended probabilistic relation *(epr-relation, for short) is a probabilistic relation with a set of* event constraints*. Each event constraint is a propositional formula in event variables.*

Semantics of an extended probabilistic relation is similar to that of probabilistic relation, with the exception that only truth assignments that satisfy event constraints are considered. More specifically, A truth assignment μ to event variables is *valid* if it satisfies all event constraints. A valid truth assignment μ defines a relation instance $r_\mu = \{t \mid t \in r \text{ and } f(t) = \mathit{true}\text{under } \mu\}$, where $f(t)$ is the event formula associated with tuple t in r. The extended probabilistic relation r represents the set of relations, called its possible-world set, defined by the set of all valid truth assignments to the event variables. We will use the abbreviation *epr-relation* for extended probabilistic relation henceforth.

Given information sources S and S', let r and r' be the probabilistic relations that represent the data in S and S', respectively. We represent a tuple in a probabilistic relation as $t@f$, where t is the pure tuple, and f is the propositional event formula associated with t. Let $r = \{t_1@f_1, \ldots, t_n@f_n\}$, where f_i is the event formula associated with the tuple t_i. Similarly, let $r' = \{u_1@g_1 \ldots, u_m@g_m\}$. We assume the set of event variables of r (*i.e.*, event variables appearing in formulas $f_1, \ldots, f_n$) and those of r' (*i.e.*, event variables appearing in formulas $g_1, \ldots, g_m$) to be disjoint. If not, a simple renaming can be used to make the two sets disjoint. r and r' can have zero or more common tuples. Assume, without loss of generality, that r and r' have p tuples in common, $0 \leq p \leq min(n, m)$, $t_1 = u_1, \ldots, t_p = u_p$. The integration algorithm is represented in Algorithm 2. In Algorithm 2, $f_i \equiv g_i$ is equivalent to the logical formula $(f_i \rightarrow g_i) \wedge (g_i \rightarrow f_i)$. We will use the notation $q = r \uplus r'$ to mean that q is the epr-relation that is the result of integration of probabilistic relations r and r'.

Algorithm 2. Integration of uncertain data represented by probabilistic relations

Given information sources S and S', let r and r' be the probabilistic relations that represent the data in S and S'. The result of integration of S and S' is represented by an epr-relation q obtained as follows:

- Copy to q the tuples in r that are not in common with r'
- Copy to q the tuples in r' that are not in common with r For each of the common tuples, copy to q the tuple either from r or from r'.
- For each common tuple t_i, add a constraint $f_i \equiv g_i$, to the set of event constraints of q, where f_i and g_i are the event formulas associated with t_i in r and r', respectively.

It has been shown in [6] that Algorithm 2 is correct. That is, when $q = r \uplus r'$ is obtained by this algorithm, then the possible-worlds of q coincide with the possible-worlds obtained by integrating possible-worlds of r and r' by Algorithm 1.

The complexity of Algorithm 2 is $O(n \log n)$, where n is the size of input (pr-relations of the sources). While the complexity of the possible-worlds integration algorithm (Algorithm 1) is quadratic in the size of its input (possible-world relations of the sources) which itself can be exponential in the size of the input of Algorithm 2.

student	course	E
Bob	CS100	$\neg c_1$
Bob	CS101	$c_1 \vee c_2$
Bob	CS201	$\neg b_1$
Bob	CS202	$\neg b_1 \wedge \neg b_2 \wedge \neg b_3$
$\neg c_1 \equiv b_1 \vee b_2$		

epr-relation $q = r \uplus r'$

Fig. 5. Extended Probabilistic relation for the integration of sources S and S'

Example 4. The result of integration of probabilistic relations of Example 3 (which themselves are equivalent to the possible-worlds relations of Example 1) is the epr-relation of Fig. 5, obtained using Algorithm 2. ∎

4.3 Determining Probabilities for Extended Probabilistic Relations

While probability computation is straightforward for probabilistic relations as discussed in Sect. 4.1, we do not have a general approach for probability computation for extended probabilistic relations. The reason is that we can no longer assume event variables are independent. Event constraints impose certain dependencies among event variables. Indeed, if we assume event variables are independent, the sum of the probabilities calculated for the possible-worlds of an epr-relation is not equal to 1. This is due to the fact that only a subset of all possible-world relations, those that correspond to valid truth assignments to event variables, are taken into account. We need an approach for probability computation for extended probabilistic relations. If not, we are forced to use the highly inefficient probabilistic possible-worlds approach for the integration of probabilistic data. Further, our probability computation approach for epr-relations must be *equivalent* to the possible-worlds approach. In other words, we have two conceptually equivalent methodologies for probabilistic data: the possible-worlds (highly inefficient) and probabilistic relation (efficient). But whatever we achieve in the probabilistic relation domain must coincide with the possible-worlds domain.

We show that under an intuitive and reasonable assumption regarding the correlation of event variables of epr-relations we are able to compute the probabilities of the result of integration. Further, this assumption is closely related to the *Partial Independence Assumption* in the possible-worlds domain (discussed in Sect. 3.1).

Partial Independence Assumption for Extended Probabilistic Relations: *Event variables are independent except for the relationships induced by the event constraints.*

The following example shows how this assumption allows us to compute probabilities for the extended probabilistic relation framework. We will discuss the correctness of this approach in the next section (Sect. 5).

student	course
Bob	CS100

(D1,D'1)

student	course
Bob	CS100
Bob	CS201

(D1,D'2)

student	course
Bob	CS100
Bob	CS101

(D2,D'1)

student	course
Bob	CS100
Bob	CS101
Bob	CS201

(D2,D'2)

student	course
Bob	CS101
Bob	CS201

(D3,D'3)

student	course
Bob	CS101
Bob	CS201
Bob	CS202

(D3,D'4)

Fig. 6. Possible-world relations of the result of integration of sources S and S'

Example 5. Let's go back to Example 4. The result of integration is the epr-relation shown in Fig. 5. The possible-worlds relations of this epr-relation are shown in Fig. 6.

How can we calculate the probability distribution of the result of integration (possible-world relations of Fig. 6)? The event-variable formulas for the 6 possible-world relations of the integration in this case are $\neg c_1 \wedge \neg c_2 \wedge b_1$, $\neg c_1 \wedge \neg c_2 \wedge \neg b_1 \wedge b_2$, $\neg c_1 \wedge c_2 \wedge b_1$, $\neg c_1 \wedge c_2 \wedge \neg b_1 \wedge b_2$, $c_1 \wedge \neg b_1 \wedge \neg b_2 \wedge b_3$, and $c_1 \wedge \neg b_1 \wedge \neg b_2 \wedge b_3$.

By the partial independence assumption event variables are independent except for the relationships induced by the event constraints. The constraint $\neg c_1 \equiv b_1 \vee b_2$ induces a relationship between c_1 on one hand, and b_1 and b_2 on the other. The rest are still independent. So, for example, c_1 and c_2 are independent, and so are c_2 and b_1; etc. In particular, b_1 and b_2 are also independent. To compute the probability associated with an event-variable formula, we rewrite the formula so that it only contains mutually independent event variables. For example, $\neg c_1 \wedge \neg c_2 \wedge b_1$ is simplified to $\neg c_2 \wedge b_1$ using the equivalence $\neg c_1 \equiv b_1 \vee b_2$. Then we are able to compute the probabilities. In this example, we obtain the following probabilities for the 6 possible-world relations: 0.13125, 0.16875, 0.21875, 0.28125, 0.05, and 0.15.

Let us compare this approach with the integration in the probabilistic possible-worlds framework (Sect. 3.1). It is easy to verify that the probabilistic distribution of the result of the integration computed by the formula $P(D_i \wedge D'_j) = P(D_i)P(D_j)/P$ is exactly the same as the distribution obtained above. For example, the probability of the possible world corresponding to (D_1, D'_1) is $0.3 \times 0.35/(0.3 + 0.5) = 0.13125$. □

5 Correctness of Probability Computation

In this section we present an overview of theoretical issues that form the basis of probability computation algorithms presented in previous sections. Interested readers are referred to the full paper which contains detailed discussions and proofs [16].

An important issue that we must address is the following:

Let a source S contain uncertain data in the probabilistic possible-worlds model, and r be an equivalent representation in the probabilistic relation model for data contained in S. It is well known that r is not unique. There can be many probabilistic relations representing the data in S.

This means that the result of integration of two sources S and S' in the possible-worlds domain is unique, but it is not unique in the probabilistic relation domain. For example, let r_1 and r_2 be alternative probabilistic relation representations for S, and r_1' and r_2' be alternative representations for S'. The result of integration of S and S' in the possible-worlds domain is unique. But in the probabilistic relation domain we can get (extended probabilistic relations) $r_1 \uplus r_1'$, $r_1 \uplus r_2'$, $r_2 \uplus r_1'$, and $r_2 \uplus r_2'$, where $\uplus$ is the integration operator (*e.g.*, using Algorithm 2). We must prove that, by our probability computation algorithm for epr-relations, all these epr-representations are indeed equivalent in the sense that they correspond to the probabilistic possible-worlds representation of the integration of S and S' (e.g., using Algorithm 1) where the probability distribution is obtained according to the approach of Sect. 3.2.

To address the above issues, we have shown the following [16].

- Given a probabilistic (or extended probabilistic) relation r, we associate a logical formula in terms of the event variables of r with each possible-world relation represented by r.
- We considered a subclass of extended probabilistic relations, namely, those that can be obtained through integration of sources represented by probabilistic relations.
- We prove that when the result of integration can be obtained by multiple epr-relations, these relations are equivalent in the following sense:
  - All epr-relations have exactly the same set of possible-worlds relations.
  - The logical formulas associated with each possible-world relation in different epr-relations are (logically) equivalent when event constraints are taken into account.
- Since the probability computation for epr-relations uses the logical formulas plus event constraints, different epr-relations for the integration of the same two sources have exactly the same probability distribution.
- Further, we also show that in the possible-worlds integration approach, the probabilities computed for the possible-world relations are exactly the same as those computed in the equivalent probabilistic relations framework.

Interested readers are referred to [16] for detailed discussions.

6 Conclusion

We focused on data integration from sources containing probabilistic uncertain information, in particular, on computing the probability distribution of the result of integration. We presented integration algorithms for data represented in two frameworks: The probabilistic possible-worlds model and the probabilistic relation model. In the latter case the result of integration is represented by an extended probabilistic relation. We showed that under intuitive and reasonable assumptions the exact probability distribution of the result of integration can be computed in the two frameworks. Alternative approaches to the computation of the probability distribution were presented in the two frameworks.

References

1. Abiteboul, S., Kanellakis, P.C., Grahne, G.: On the representation and querying of sets of possible worlds. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 34–48 (1987)
2. Agrawal, P., Sarma, A.D., Ullman, J.D., Widom, J.: Foundations of uncertain-data integration. Proc. VLDB Endowment **3**(1), 1080–1090 (2010)
3. Antova, L., Jansen, T., Koch, C., Olteanu, D.: Fast and simple relational processing of uncertain data. In: Proceedings of IEEE International Conference on Data Engineering, pp. 983–992 (2008)
4. Barbará, D., Garcia-Molina, H., Porter, D.: The management of probabilistic data. IEEE Trans. Knowl. Data Eng. **4**(5), 487–502 (1992)
5. Benjelloun, O., Sarma, A.D., Halevy, A.Y., Theobald, M., Widom, J.: Databases with uncertainty and lineage. VLDB J. **17**(2), 243–264 (2008)
6. Dayyan Borhanian, A., Sadri, F.: A compact representation for efficient uncertain-information integration. In: Proceedings of International Database Engineering and Applications IDEAS, pp. 122–131 (2013)
7. Codd, E.F.: Extending the database relational model to capture more meaning. ACM Trans. Database Syst. **4**(4), 397–434 (1979)
8. Dalvi, N.N., Ré, C., Suciu, D.: Probabilistic databases: diamonds in the dirt. Commun. ACM **52**(7), 86–94 (2009)
9. Dalvi, N.N., Suciu, D.: Efficient query evaluation on probabilistic databases. In: Proceedings of International Conference on Very Large Databases, pp. 864–875 (2004)
10. Dalvi, N.N., Suciu, D.: Efficient query evaluation on probabilistic databases. VLDB J. **16**(4), 523–544 (2007)
11. Haas, L.: Beauty and the beast: the theory and practice of information integration. In: Schwentick, T., Suciu, D. (eds.) ICDT 2007. LNCS, vol. 4353, pp. 28–43. Springer, Heidelberg (2006). doi:10.1007/11965893_3
12. Halevy, A.Y., Rajaraman, A., Ordille, J.J.: Data integration: The teenage years. In: Proceedings of International Conference on Very Large Databases, pp. 9–16 (2006)
13. Liu, K.C., Sunderraman, R.: On representing indefinite and maybe information in relational databases. In: Proceedings of IEEE International Conference on Data Engineering, pp. 250–257 (1988)
14. Sadri, F.: On the foundations of probabilistic information integration. In: Proceedings of International Conference on Information and Knowledge Management, pp. 882–891 (2012)
15. Sadri, F.: Belief revision in uncertain data integration. In: Sharaf, M.A., Cheema, M.A., Qi, J. (eds.) ADC 2015. LNCS, vol. 9093, pp. 78–90. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19548-3_7
16. Sadri, F., Tallur, G.: Integration of probabilistic uncertain information (2016). CoRR, abs/1607.05702