Belief Revision in Uncertain Data Integration

Fereidoon Sadri(✉)

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

Abstract. This paper studies the problem of integrating probabilistic uncertain information. Certain constraints are imposed by the semantics of integration, but there is no guarantee that they are satisfied in practical situations. We present a Bayesian-based approach to revise the probability distribution of the information in the sources in a systematic way to remedy this difficulty. The revision step is similar in spirit to tasks like data cleaning and record linkage and should be carried out before integration can be achieved for probabilistic uncertain data.

Keywords: Information integration · Uncertain data · Probabilistic data · Belief revision

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 [3–6,13,15]. The importance of information integration *with uncertainty*, on the other hand, has been realized more recently. [8,9,11,13–15,17–19,21,24–26]. It has been observed that [15]:

> While in traditional database management managing uncertainty and lineage seems like a nice feature, in data integration it becomes a necessity.

In this paper we study the problem of integrating probabilistic uncertain information. Certain constraints are imposed by the semantics of integration, but there is no guarantee that they are satisfied in practical situations. We present a Bayesian-based approach to revise the probability distribution of the information in the sources in a systematic way to remedy this difficulty. The revision step is similar in spirit to tasks like data cleaning and record linkage and should be carried out before integration can be achieved for probabilistic uncertain data.

This paper is organized as follows: We present the theory of uncertain information integration in Section 2. Probabilistic constraints are discussed in Section 3 and our proposed Bayesian-based approach to revise information sources probabilities is presented in Section 4. Conclusions are presented in Section 5.

M.A. Sharaf et al. (Eds.): ADC 2015, LNCS 9093, pp. 78–90, 2015.
DOI: 10.1007/978-3-319-19548-3_7

2 Preliminaries – Foundations of Information Integration

Foundations of information integration with uncertainty have been discussed in [2,22]. We present a brief summary here. We begin with an example from [22].

Example 1. John and Jane are talking about fellow student Bob. John says "I am taking CS100 and CS101, and Bob is in one of them, but not in both." Jane says "I am taking CS101 and CS102 and Bob is in one of them, but not in both."

Intuitively, if we integrate the information from these two sources (John and Jane), we should infer that Bob is either taking CS101, or he is taking both CS100 and CS102. We present an algorithm for the integration of uncertain information in Section 2.1. ■

The model used in [2,22] for the representation of uncertain information is the well-known *possible-worlds* model [1]. We should emphasize that the possible-worlds model is used in the *formalization* of information integration. It is not, in general, efficient for implementation. In Example 1, the information presented by the two sources (John and Jane) is represented by the possible-worlds shown in Figures 1 and 2.

student	course
Bob	CS100

D1

student	course
Bob	CS101

D2

Fig. 1. Possible Worlds of source S1

student	course
Bob	CS101

D3

student	course
Bob	CS102

D4

Fig. 2. Possible Worlds of source S2

We will summarize the integration approach from [22] which uses a simple logic-based technique in Section 2.1). This approach has been shown to be equivalent to the integration approach of [2] which is based on the concept of superset-containment. Interested readers are referred to [22] for details.

First, we should mention that the pure possible world model is not adequate for integration applications. We need additional information, namely, the set of all tuples. The following example demonstrates the possible-worlds with tuple sets model.

Example 2. Andy and Jane are talking about fellow student Bob. Andy says "I am taking CS100, CS101, and CS102 and Bob is in either CS100 or CS101 but not in both." Jane says "I am taking CS101 and CS102 and Bob is in one of them, but not in both."

Intuitively, if we integrate the information from these two sources, we should infer that Bob is taking CS101. The second possibility from Example 1, namely

Bob taking CS100 and CS102, is not valid anymore since Andy's statement rules out the possibility that Bob is taking 102.

However, the possible-worlds representations of these sources (Andy and Jane) are exactly the same as those of Example 1 (Figures 1 and 2). Only when we add the tuple-set to possible worlds of Andy, namely {(Bob, CS100), (Bob, CS101), (Bob, CS102)}, It becomes explicit that Andy's statement eliminates the possibility that Bob is taking CS102. ■

Hence, we will use the following definition from [2] for uncertain databases that adds tuple sets to the possible-worlds model. To simplify presentation, we assume that possible worlds are sets of tuples in a single relation.

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

This definition adds tuple-set $T(U)$ to the traditional possible-worlds model. In fact, as shown in Example 2, there may be tuples in the tuple set, $t \in T(U)$, that do not appear in any possible world of the uncertain database U. If $T(U)$ is not provided explicitly, then we use the set of all tuples in the possible worlds, *i.e.*, $T(U) = D_1 \cup \cdots \cup D_n$. It is interesting to notice that this model exhibits both closed-world and open-world properties: If a tuple $t \in T(U)$ does not appear in a possible world D_i, then it is assumed to be *false* for D_i (hence, closed-world assumption). In other words, D_i explicitly rules out t. The justification is that the source providing the uncertain information represented by U is aware of (the information represented by) all $t \in T(U)$. If some $t \in T(U)$ is absent from D_i, then the source explicitly rules out t from D_i. On the other hand, all other tuples $t \notin T(U)$ are assumed possible (*unknown*) for possible-worlds D_i (hence, open-world assumption). This distinction is important for integration: Consider integrating D_i, where $t \notin D_i$, with a possible-world D'_j from another source, where $t \in D'j$. For the first case ($t \in T(U)$), D_i and D'_j are not compatible and can not be integrated. This is because D_i explicitly rules out t while D'_j explicitly includes it. On the other hand, for the second case ($t \notin T(U)$), D_i and D'_j can be integrated since D_i can accept t as a valid tuple.

2.1 Integration Using Logical Representation

In this section we review some results from [22]. Given an uncertain database U, we assign a propositional variable x_i to each tuple $t_i \in T(U)$. We define the formula f_j corresponding to a possible world D_j, and the formula f corresponding to the uncertain database U as follows:

Definition 2. (Logical Representation of an Uncertain Database). *Let* D_j *be a database in the possible worlds of uncertain Database* U. *Construct a formula as the conjunction of all variables* x_i *where the corresponding tuple* t_i

is in D_j, and the conjunction of $\neg x_i$ where the corresponding tuple t_i is not in D_j. That is,

$$f_j = \bigwedge_{t_i \in D_j} x_i \bigwedge_{t_i \notin D_j} \neg x_i \tag{1}$$

The formula corresponding to the uncertain database U *is the disjunction of the formulas corresponding to the possible worlds of U. That is,*

$$f = \bigvee_{D_j \in PW(U)} f_j \tag{2}$$

Now we can integrate uncertain databases using their logical representations as follows:

Let $S_1, \ldots, S_n$ be sources containing (uncertain) databases $U_1, \ldots, U_n$. Let the propositional formulas corresponding to $U_1, \ldots, U_n$ be $f_1, \ldots, f_n$. We obtain the formula f corresponding to the uncertain database resulting from integrating $U_1, \ldots, U_n$ by conjuncting the formulas of the databases: $f = f_1 \wedge \cdots \wedge f_n$.

Example 3. (Integration Using Logical Representation) Consider Example 1. The uncertain database corresponding to John's statement is represented by $(x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$, where x_1, and x_2 correspond to the tuples (Bob, CS100) and (Bob, CS101), respectively. The uncertain database corresponding to Jane's statement is represented by $(x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3)$, where x_2 is as above and x_3 corresponds to the tuple (Bob, CS102). The integration in this case is obtained as

$$\begin{aligned}&((x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)) \wedge ((x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3))\\ &\quad = (x_1 \wedge \neg x_2 \wedge x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)\end{aligned}$$

which corresponds to the possible worlds of Figure 3. The result is consistent with our intuition: Based on statements by John and Jane, Bob is taking either CS101 or both CS100 and CS102.

student	course
Bob	CS101

student	course
Bob	CS100
Bob	CS102

Fig. 3. Possible Worlds of the Integration for Example 1

Now consider Example 2. The uncertain database corresponding to Andy's statement is represented by $(x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)$, where x_1, x_2, and x_3 represent (Bob, CS100), (Bob, CS101), and (Bob, CS102), respectively. The uncertain database corresponding to Jane's statement is the same as above $(x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3)$. The integration in this case is obtained as

$$\begin{aligned}&((x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)) \wedge ((x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3))\\ &\quad = (\neg x_1 \wedge x_2 \wedge \neg x_3)\end{aligned}$$

corresponding to the (in this case, definite) relation consisting only of the tuple (Bob, CS101). Again, this result is consistent with our intuition: Based on statements by Andy and Jane, Bob is taking CS101. ■

2.2 Probabilistic Uncertain Information

The conceptual model for probabilistic uncertain information is the possible-worlds with tuple-set model with a probability distribution over the set of possible worlds. More formally,

Definition 3. (Probabilistic Uncertain Database). *A* probabilistic uncertain database U *consists of a finite set of tuples* $T(U)$ *and a nonempty set of possible worlds* $PW(U) = \{D_1, \ldots, D_m\}$. *Each* $D_i \subseteq T(U)$ *is associated with a probability* $P(D_i)$ *in the [0,1] range, where* $\sum_{i=1}^{m} P(D_i) = 1$. ■

The integration technique of Section 2.1 can be applied to the probabilistic case to obtain the possible-worlds of the result. We have shown in [22] that very interesting constraints are imposed on the probabilistic structure of information sources in the integration of probabilistic data. We discuss these constraints in Section 3 below. First, we need a few definitions and observations.

Definition 4. (Compatible Possible Worlds). *Let* S *and* S' *be sources containing probabilistic uncertain information* $\{D_1, \ldots, D_m\}$ *and* $\{D'_1, \ldots, D'_{m'}\}$, *respectively. Let* T *and* T' *be the tuple-sets of* S *and* S'. *A pair of possible-worlds* (D_i, D'_j) *from* S *and* S' *are said to be* compatible *if (1) For all tuples* $t \in D_i - D'_j$, $t \notin T'$, *and (2) For all tuples* $t \in D'_j - D_i$, $t \notin T$. ■

It is easy to verify that, Given two information sources, only compatible pairs of possible worlds from the two sources can be integrated (combined). Each compatible pair produces a possible world in the answer.

We use a *compatibility* graph G to capture the compatibility relationship defined above. Let S and S' be sources containing probabilistic uncertain information $\{D_1, \ldots, D_m\}$ and $\{D'_1, \ldots, D'_{m'}\}$, respectively. The compatibility graph G for S and S' is a bipartite graph. Nodes of G have a one-to-one correspondence with possible worlds of S and S'. That is, G has nodes $\{N_1, \ldots, N_m, N'_1, \ldots, N'_{m'}\}$, where node N_i, $i = 1, \ldots, m$, corresponds to the world D_i of S, and node N'_j, $j = 1, \ldots, m'$, corresponds to the world D'_j of S'. There is an edge between N_i and N'_j if the pair of possible worlds (D_i, D'_j) are compatible. We sometimes overload the notation and use $\{D_1, \ldots, D_m, D'_1, \ldots, D'_{m'}\}$ for possible-worlds as well as for nodes of the compatibility graph of the two sources.

The following result was proven in [22].

Theorem 1. *Let* G *be the compatibility graph of sources* S *and* S'. *Each connected component of* G *is a complete bi-partite graph.* ■

So, if we consider a connected component of G, it has a set of nodes from S (*e.g.*, $\{D_{i1}, \ldots, D_{ik}\} \subseteq \{D_1, \ldots, D_m\}$) and another set of nodes from S' (*e.g.*, $\{D'_{j1}, \ldots, D'_{jk'}\} \subseteq \{D'_1, \ldots, D'_{m'}\}$). Then, by Theorem 1, every node in the first group is connected to every node in the second group. Further, these nodes are not connected to any other nodes. (Note that we are using a symbol D to refer both to a possible-world D and to the node representing D in the compatibility graph.)

3 Probabilistic Constraints

When integrating sources containing probabilistic uncertain information, certain constraints are imposed on the probabilistic distributions of the possible worlds of the sources. The following theorem is from [22]:

Theorem 2. *Let S and S' be sources containing probabilistic uncertain information $\{D_1, \ldots, D_m\}$ and $\{D'_1, \ldots, D'_{m'}\}$, respectively. Let G be their (bipartite) compatibility graph. Let G_1 be a connected component of G, with the set of nodes $\{D_{i1}, \ldots, D_{ik}\} \subseteq \{D_1, \ldots, D_m\}$ and $\{D'_{j1}, \ldots, D'_{jk'}\} \subseteq \{D'_1, \ldots, D'_{m'}\}$. Then the following constraint between the probabilities of the possible-worlds represented by the nodes of the connected component G_1 must hold:*

$$\sum_{D \in \{D_{i1}, \ldots, D_{ik}\}} P(D) = \sum_{D' \in \{D'_{j1}, \ldots, D'_{jk'}\}} P(D') \tag{3}$$

In other words, each connected component G_1 of the bipartite compatibility graph G of S and S' enforces a constraint that the sum of probabilities of possible-worlds associated with S in the connected component should be equal to the sum of probabilities of possible-worlds associated with S' in the same connected component.

Example 4. The compatibility graph G for the sources of Example 1 (John and Jane) is simple: G has four nodes corresponding to the possible worlds of source 1, D_1 and D_2, and the possible worlds of source 2, D_3 and D_4 (See Figures 1 and 2). The edges of G are (D_1, D_4) and (D_2, D_3). Hence, G has two connected components: $\{D_1, D_4\}$ and $\{D_2, D_3\}$. The probabilistic constraints for this case are $P(D_1) = P(D_4)$ and $P(D_2) = P(D_3)$.

For another example, consider information sources B_1 and B_2 about books and their authors whose possible-worlds are shown in Figures 4 and 5.

book1	Dan
book2	Jen

E1

book1	Dan
book3	Bob

E2

book2	Amy
book3	Bob

E3

Fig. 4. Possible Worlds of source B_1

book1	Dan
book4	Jan

E'1

book1	Dan
book4	Pam

E'2

book4	Jan
book4	Pam

E'3

Fig. 5. Possible Worlds of source B_2

The compatibility graph in this case has two connected components with edges $\{(E_1, E'_1), (E_1, E'_2), (E_2, E'_1), (E_2, E'_2)\}$, and $\{(E_3, E'_3)\}$, respectively (See Figure 6). There are two probabilistic constraints corresponding to the two connected components: $P(E_1) + P(E_2) = P(E'_1) + P(E'_2)$, and $P(E_3) = P(E'_3)$.

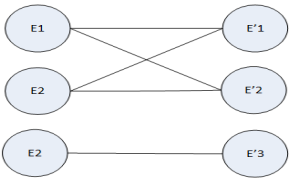

Fig. 6. Compatibility graph for Sources B1 and B2 (Example 4)

These constraints may appear counterintuitive in the first sight. Basically they state that, in general, sources containing probabilistic uncertain information are not independent. Rather, every pair of sources may be correlated. Recent research on *data fusion* (for example, [7,10,12,20,28]) confirms this fact. It has been shown that by taking into account the correlations among sources, significantly better fusion (integration) results can be obtained. Our framework is different from that of data fusion. Nevertheless, the correlation between information sources remains valid.

We will summarize the proof of Theorem 2 below to shed more light on the correlation between sources. Let S and S' be sources containing probabilistic uncertain information $\{D_1, \ldots, D_m\}$ and $\{D'_1, \ldots, D'_{m'}\}$, respectively. Let $\{P(D_1), \ldots, P(D_m)\}$ and $\{P(D'_1), \ldots, P(D'_{m'})\}$ be the probability distributions of the possible-worlds of S and S'. Intuitively, $P(D)$ is the probability of the event that the real world database is D. Note that the probability distribution $\{P(D_1), \ldots, P(D_m)\}$ consists of events that are mutually exclusive and exhaustive. That is, (1) $P(D_i \wedge D_j) = 0$ for $D_i \neq D_j$, in other words, the real world can not be D_i and D_j at the same time, and (2) $\sum_{i=1}^{m} P(D_i) = 1$. Similarly, the probability distribution $\{P(D'_1), \ldots, P(D'_{m'})\}$ is also mutually exclusive and exhaustive. So, we can write

$$P(D_i) = \sum_{j=1}^{m'} P(D_i \wedge D'_j), i = 1, \ldots, m \tag{4}$$

and

$$P(D'_j) = \sum_{i=1}^{m} P(D_i \wedge D'_j), j = 1, \ldots, m' \tag{5}$$

Given a pair of possible-worlds (D_i, D'_j), if D_i and D'_j are not compatible, they contain contradictory information and can not be combined. That is, the events "the real world is D_i" and "the real world is D'_j" are contradictory. Hence, $P(D_i \wedge D'_j) = 0$.

Now consider a connected component G_1 with the set of nodes $\{D_{i1}, \ldots, D_{ik}\} \subseteq \{D_1, \ldots, D_m\}$ and $\{D'_{j1}, \ldots, D'_{jk'}\} \subseteq \{D'_1, \ldots, D'_{m'}\}$. Each possible-world in the first set is compatible with every possible-world in the second set, and vice-versa. Further, these possible worlds are not compatible with any other possible worlds. It follows that $P(D_{iq}) = \sum_{r=1}^{k'} P(D_{iq} \wedge D'_{jr}), q = 1, \ldots, k$, and

$P(D'_{jr}) = \sum_{q=1}^{k} P(D_{iq} \wedge D'_{jr}), r = 1, \ldots, k'$. Then

$$\sum_{q=1}^{k} P(D_{iq}) = \sum_{q=1}^{k} \sum_{r=1}^{k'} P(D_{iq} \wedge D'_{jr})$$

and

$$\sum_{r=1}^{k'} P(D'_{jr}) = \sum_{r=1}^{k'} \sum_{q=1}^{k} P(D_{iq} \wedge D'_{jr})$$

Hence

$$\sum_{q=1}^{k} P(D_{iq}) = \sum_{r=1}^{k'} P(D'_{jr})$$

which is the same as Equation 3. ■

We have presented algorithms for the calculation of the probabilities of the possible-worlds of the result of integrating probabilistic uncertain information in the case where probabilistic constraints are satisfied [22,23]. But, in practice, the probabilistic distribution of information sources are provided by the sources themselves or through certain data mining or analytic processing. There is no guarantee that probabilistic constraints are indeed satisfied in practice. The rest of this paper provides approaches for these cases when the probabilistic constraints are not satisfied.

4 Revising Probability Distribution of Sources

In this section we concentrate on the case where we have sources S and S' containing probabilistic uncertain information and one or more of the probabilistic constraints are not satisfied. We use a Bayesian-based approach to revise the probabilistic distributions of the sources such that the revised distributions do satisfy all constraints.

Let the possible-worlds of S and S' be $\{D_1, \ldots, D_m\}$ and $\{D'_1, \ldots, D'_{m'}\}$, respectively. Let us begin by treating S as the original set of events, and S' as the new *evidence* by which the probabilities of the original events, $P(D_i)$'s, are revised. In other words, we want to compute the conditional probabilities

$$P(D_i \mid \text{The evidence provided by } S')$$

which we will simply denote by $P(D_i \mid S')$ henceforth. We will use Q for the *posterior* probability distributions. So, $Q(D_i) = P(D_i \mid \text{The evidence provided by } S')$ is the revised (or posterior) probability of D_i.

This is a case where the evidence itself is probabilistic. Hence, we will use Richard Jeffrey's rule of conditioning [16,27] which is an extension of Bayes' rule to probabilistic evidence.

$$Q(D_i) = \sum_{j=1}^{m'} Q(D'_j) P(D_i \mid D'_j), \quad i = 1, \ldots, m \tag{6}$$

but

$$P(D_i \mid D'_j) = \frac{P(D'_j \mid D_i)P(D_i)}{P(D'_j)} = \frac{P(D'_j \mid D_i)P(D_i)}{\sum_{k=1}^{m} P(D'_j \mid D_k)P(D_k)}$$

Hence, we obtain the following alternative formulation:

$$Q(D_i) = \sum_{j=1}^{m'} Q(D'_j) \left(\frac{P(D'_j \mid D_i)P(D_i)}{\sum_{k=1}^{m} P(D'_j \mid D_k)P(D_k)} \right), \quad i = 1, \ldots, m \tag{7}$$

As mentioned earlier, Given sources S and S' containing probabilistic uncertain information, probability constraints (Equation 3) may not hold in practice. Next, we prove that the *revised* probability distributions $Q(D_i), i = 1, \ldots, m$ and $Q(D'_j), j = 1, \ldots, m'$, as obtained by Equation 6 (or Equation 7), satisfy all probabilistic constraints.

Theorem 3. *Consider sources S and S' containing probabilistic uncertain information $\{D_1, \ldots, D_m\}$ and $\{D'_1, \ldots, D'_{m'}\}$, respectively. Let G be the compatibility graph of S and S', and $\{P(D_1), \ldots, P(D_m)\}$ and $\{P(D'_1), \ldots, P(D'_{m'})\}$ be their probability distributions. Consider a connected component G_1 of G. Then*

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} Q(D'_j)$$

where $Q(D_i)$ and $Q(D'_j)$ are revised probability distributions according to Equation 6.

Proof. Consider a node $D_i \in G_1$. Note that $P(D_i \mid D') = 0$ for all nodes $D' \notin G_1$ (D_i and D' are not compatible if they do not belong to the same connected component.) So, we can write by Equation 6,

$$Q(D_i) = \sum_{j=1}^{m'} Q(D'_j)P(D_i \mid D'_j) = \sum_{D'_j \in G_1} Q(D'_j)P(D_i \mid D'_j)$$

Then,

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D_i \in G_1} \sum_{D'_j \in G_1} Q(D'_j)P(D_i \mid D'_j) = \sum_{D'_j \in G_1} Q(D'_j) \sum_{D_i \in G_1} P(D_i \mid D'_j)$$

and (again, since $P(D \mid D'_j) = 0$ for all nodes D that are not in the same connected component as D'_j – which is G_1):

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} Q(D'_j) \sum_{i=1}^{m} P(D_i \mid D'_j)$$

But $P(D_i \mid D'_j) = \frac{P(D_i \wedge D'_j)}{P(D'_j)}$. Hence,

$$\sum_{i=1}^{m} P(D_i \mid D'_j) = \sum_{i=1}^{m} \frac{P(D_i \wedge D'_j)}{P(D'_j)} = \frac{\sum_{i=1}^{m} P(D_i \wedge D'_j)}{P(D'_j)} = \frac{P(D'_j)}{P(D'_j)} = 1$$

It follows that

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} Q(D'_j)$$

Equations 6 and 7 contain two sets of unknowns: Both posterior probability distributions, $Q(D_i), i = 1, \ldots, m$ and $Q(D'_j), j = 1, \ldots, m'$, are unknown. Choosing the values of one set impacts those of the other set. So, our task is to compute these two sets of unknowns. We will discuss below how to use our confidence in the sources to compute these parameters.

4.1 Total Confidence in the Evidence

In some applications, we may have complete confidence in the evidence (*i.e.*, information provided by source S' in our case), and want to revise the probability distribution of the original set of events (information provided by source S) with respect to the evidence. In this case the probability distribution of S' remains unchanged. In other words, we have $Q(D'_j) = P(D'_j), j = 1, \ldots, m$.

We know by Theorem 3 that for every connected component G_1 of the compatibility graph G:

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} Q(D'_j)$$

So, if we have total confidence in the evidence S':

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} P(D'_j)$$

In other words, the probability distribution of S is revised in a way that the sum of (revised) probabilities on the S side of a connected component G_1 equals the sum of (original) probabilities on the S' side of G_1.

The "dual" of this situation is when we have total confidence in S, in which case the probability distribution of S' will be revised such that

$$\sum_{D_i \in G_1} P(D_i) = \sum_{D'_j \in G_1} Q(D'_j)$$

4.2 General Case

In general, we will not have total confidence in either of sources. Rather, we may have subjective or analytic confidence measures for the sources. We formalize this situation by introducing *confidence measures* α_i for each source S_i, such that $\sum_{i=1}^{n} \alpha_i = 1$, where n is the number of sources. If there are two sources S

and S', their confidence measures can be denoted by α and $1-\alpha$. Our approach is to revise the probability distributions of both sources to obtain a weighted probability sum for each connected component G_1 as follows:

$$\sum_{D_i \in G_1} Q(D_i) = \sum_{D'_j \in G_1} Q(D'_j) = \alpha \sum_{D_i \in G_1} P(D_i) + (1-\alpha) \sum_{D'_j \in G_1} P(D'_j) \quad (8)$$

The cases for total confidence in S and in S' correspond to $\alpha = 1$ and $\alpha = 0$, respectively.

5 Conclusion

We have studied the problem of integrating probabilistic uncertain information. Certain constraints are imposed by the semantics of integration, but there is no guarantee that they are satisfied in practical situations. We presented a Bayesian-based approach to revise the probability distribution of the information in the sources in a systematic way to remedy this difficulty. The revision step is similar in spirit to tasks like data cleaning and record linkage and should be carried out before integration can be achieved for probabilistic uncertain data.

There is a close relationship between uncertain-data integration and *data fusion*, which refers to the integration of massive amounts of mined data. The process of mining data from sources such as web pages, social media and email messgaes generate large amounts of data with differing degrees of correctness confidence, which can be conveniently modeled by probabilistic uncertain data. In the future, we intend to study the application of our Bayesian probability revision approach to data fusion.

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 Managementof Data, pp. 34–48 (1987)
2. Agrawal, P., Sarma, A.D., Ullman, J.D., Widom, J.: Foundations of uncertain-data integration. Proceedings of the 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. Antova, L., Koch, C., Olteanu, D.: 10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information. In: Proceedingsof IEEE International Conference on Data Engineering, pp. 606–615 (2007)
5. Chen, D., Chirkova, R., Sadri, F., Salo, T.J.: Query optimization in information integration. Acta Informatica **50**(4), 257–287 (2013)
6. Dalvi, N.N., Ré, C., Suciu, D.: Probabilistic databases: diamonds in the dirt. Communications of the ACM **52**(7), 86–94 (2009)
7. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: The role of source dependence. PVLDB **2**(1), 550–561 (2009)

8. Dong, X.L., Halevy, A., Yu, C.: Data integration with uncertainty. In: Proceedings of International Conference on Very Large Databases, pp. 687–698 (2007)
9. Dong, X.L., Halevy, A.Y., Yu, C.: Data integration with uncertainty. The VLDB Journal **18**(2), 469–500 (2009)
10. Dong, X.L., Saha, B., Srivastava, D.: Less is more: Selecting sources wisely for integration. Proceedings of the VLDB Endowment **6**(2), 37–48 (2012)
11. Eshmawi, A.A., Sadri, F.: Information integration with uncertainty. In: Proceedings of International Database Engineering and Applications, IDEAS, pp. 284–291 (2009)
12. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: Proceedings of ACM InternationalConference on Web Search and Data Mining, pp. 131–140 (2010)
13. 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)
14. Halevy, A.Y., Ashish, N., Bitton, D., Carey, M.J., Draper, D., Pollock, J., Rosenthal, A., Sikka, V.: Enterprise information integration: successes, challenges and controversies. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 778–787 (2005)
15. 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)
16. Jeffrey, R.: The Logic of Decision. McGraw-Hill (1965)
17. Magnani, M., Montesi, D.: Uncertainty in data integration: current approaches and open problems. In: Proceedings of VLDB Workshop on Managementof Uncertain Data, pp. 18–32 (2007)
18. Magnani, M., Montesi, D.: A survey on uncertainty management in data integration. ACM Journal of Data and Information Quality 2(1) (2010)
19. Olteanu, D., Huang, J., Koch, C.: SPROUT: Lazy vs. eager query plans for tuple-independent probabilistic databases. In: Proceedings of IEEE International Conference on Data Engineering, pp. 640–651 (2009)
20. Pochampally, R., Sarma, A.D., Dong, X.L., Meliou, A., Srivastava, D.:. Fusing data with correlations. In: Proceedings of ACM SIGMODInternational Conference on Management of Data, pp. 433–444 (2014)
21. Re, C., Dalvi, N.N., Suciu, D.: Efficient top-k query evaluation on probabilistic data. In: Proceedings of IEEE International Conference on DataEngineering, pp. 886–895 (2007)
22. Sadri, F.: On the foundations of probabilistic information integration. In: Proceedings of International Conference on Information and Knowledge Management, pp. 882–891 (2012)
23. Sadri, F., Tallur, G: Integration of probabilistic uncertain information (2014) (manuscript)
24. Sarma, A.D., Benjelloun, O., Halevy, A.Y., Nabar, S.U., Widom, J.: Representing uncertain data: models, properties, and algorithms. The VLDB Journal **18**(5), 989–1019 (2009)
25. Sarma, A.D., Benjelloun, O., Halevy, A.Y., Widom, J.: Working models for uncertain data. In: Proceedings of IEEE International Conferenceon Data Engineering, p. 7 (2006)

26. Sen, P., Deshpande, A.: Representing and querying correlated tuples in probabilistic databases. In: Proceedings of IEEE International Conference onData Engineering, pp. 596–605 (2007)
27. Shafer, G.: Jeffrey's rule of conditioning. Philosophy of Science **48**(3), 337–362 (1981)
28. Zhao, B., Rubinstein, B.I.P., Gemmell, J., Han, J.: A bayesian approach to discovering truth from conflicting sources for data integration. Proceedings of the VLDB Endowment **5**(6), 550–561 (2012)