On the Foundations of Probabilistic Information Integration

Fereidoon Sadri
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, NC 27402, USA
f_sadri@uncg.edu

ABSTRACT

Information integration has been a subject of research for several decades and still remains a very active research area. Many new applications depend or benefit from large scale integration. Examples include large research projects in life sciences, need for data sharing among government agencies, reliance of corporations on business intelligence (which requires data integration from many heterogeneous sources), and integration of information on the web. The importance of information integration *with uncertainty* has been observed in recent years. Frequently, information from multiple sources are uncertain and possibly inconsistent. Further the process of integration often depends on approximate schema mappings, another source of uncertainty. An integration system is useful only to the extent that the information it produces can be trusted. Hence, providing a measure of certainty for integrated information is of crucial importance in many important applications.

In this paper we study the problem of integration of uncertain information. We present a simple and intuitive approach to the representation and integration of uncertain information from multiple sources, and show that our integration approach coincides with a recent formalism for uncertain information integration. We extend the model to probabilistic possible-worlds, and show certain unintuitive constraints are imposed upon probabilities of possible-worlds of sources. In particular, we show the probabilities of possible worlds of a source are not independent, rather, they are dependent on probabilities of other sources. We study the problem of determining the probabilities for the result of integration. Finally, we present a practical approach to relaxing probabilistic constraints in integration.

Categories and Subject Descriptors

H.2 [**Database Management**]: [Miscellaneous]

General Terms

Theory

Keywords

Information integration, uncertain data, probabilistic data.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Information integration has been a subject of research for several decades, and despite significant progress, in particular in recent years, it still remains a very active research area [24, 28]. Many new applications depend or benefit from large scale integration. Examples include large research projects in life sciences (such as biological and astronomy research), need for data sharing among government agencies, reliance of corporations on business intelligence (which requires data integration from many heterogeneous sources), and integration of information on the web [25]. The importance of information integration *with uncertainty* has been observed in recent surveys on information integration [24, 28, 34, 35]. The following quote is from [28]:

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

Frequently, information from multiple sources are uncertain and possibly inconsistent. Further the process of integration often depends on approximate schema mappings, another source of uncertainty. An integration system is useful only to the extent that the information it produces can be trusted. Hence, providing a measure of certainty for integrated information is of crucial importance in many important applications.

In this paper we study the problem of integration of uncertain information. Our contributions include:

- We study the problem of uncertain information integration using the possible-worlds model and present a simple and intuitive approach to representation and integration of uncertain information from multiple sources.
- We show that our integration approach coincides with a recent formalism for uncertain information integration [2].
- We extend the model to probabilistic possible-worlds, and show certain unintuitive constraints are imposed upon probabilities of possible-worlds of sources. In particular, we show the probabilities of possible worlds of a source are not independent, rather, they are dependent on probabilities of other sources. We study the problem of determining the probabilities for the result of integration. Finally, we present a practical approach to relaxing probabilistic constraints in integration.

This paper is organized as follows. We start with a motivating example in Section 2, and discuss issues involved in uncertain information integration. Then we revisit the well-known possible-worlds framework for uncertain information in Section 3. We introduce a simple and intuitive logical representation of possible worlds, and present an algorithm for information integration using this representation (Sections 3.2 and 3.3), and show the correspondence of integration using logical representation with integration based on the concept of *superset containment* of [2] (Section 3.4). In Section 4 we extend the integration framework to probabilistic possible worlds model, and show that certain restrictive constraints are imposed on the probabilities assigned by sources to their possible worlds. Then we study the problem of determining the probabilities for the result of integration, and present a practical approach to relaxing probabilistic constraints in integration (Section4.2). We review related work in Section 5 and present conclusions and future work in Section 6.

2. A MOTIVATING EXAMPLE

EXAMPLE 1. *Consider information sources S1 and S2 containing data about student registrations. Source S1 states that Bob is taking CS100 or CS101 (but not both). Source S2 states that Bob is taking CS101 or CS 102 (but not both). How do we integrate the information from S1 and S2?*

This is a very simple example, yet, as we will see, certain questions arise when we attempt the integration. First, let us look at how the information in sources S1 and S2 are represented. The *possible worlds* model [1] has been widely accepted as the conceptual model of uncertain information. In this model, the information represented by each source is (conceptually) considered as a set of database instances, each instance being a *possible* state of the real world. In our example, each database instance for source S1 consists of a single relation. The possible worlds of S1 are shown in Figure 1. Possible worlds of S2 are similar and are shown in Figure 2.

student	course
Bob	CS100

D1

student	course
Bob	CS101

D2

Figure 1: Possible Worlds of source S1

student	course
Bob	CS101

D3

student	course
Bob	CS102

D4

Figure 2: Possible Worlds of source S2

In practice, a compact representation, such as probabilistic databases [11, 12] or U-Relational Model [3], is chosen that avoids enumerating all possible worlds, and, hopefully, permits efficient information integration and query processing. We will not address the issue of compact representation in this paper. Rather, we concentrate our discussion on the fundamental problems of information integration. Interested readers are referred to [44, 45] for a detailed discussion regarding representations of uncertain data and their properties.

2.1 How to Integrate?

When integrating information from multiple sources with definite data, the usual approach is to form the union of corresponding relations from the sources. For example, if S3 stated that Bob is (definitely) taking CS100 and S4 stated that Bob is (definitely) taking CS102, then the integration would state that Bob is taking both CS100 and CS102. Let us apply the union approach to uncertain case. A pairwise union of possible worlds of S1 and S2 in our first example yields the following four possible worlds:

student	course
Bob	CS100
Bob	CS101

student	course
Bob	CS101

student	course
Bob	CS100
Bob	CS102

student	course
Bob	CS101
Bob	CS102

Figure 3: A Possible Result of Integrating Sources S1 and S2

Is this the correct answer to integrating S1 and S2? As we will see, in some cases it is although there are many cases where it is not the correct answer.

2.2 Two Scenarios

Scenario 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 I am not sure which one." Note that John's statement coincides with source S1. Namely, Bob is taking CS100 or CS 101 (but not both).

Jane says "I am taking CS101 and CSC102 and Bob is in one of them, but I am not sure which one." Jane's statement coincides with source S2.

How do we integrate John's and Jane's statements? Intuitively, the combination of John and Jane's statements indicate that Bob is either taking CS101, or he is taking both CS100 and CS102. If we compare this with the four possible worlds of Figure 3 which were obtained by pairwise union operation, we note that the first and fourth possible worlds are ruled out. This is due to the fact that John's statement implies Bob is not in both CS100 and CS101, ruling out the first possible world. Similarly, Jane's statement implies Bob is not in both CS101 and CS102, ruling out the fourth possible world. Note that these observations, that John's statement implies that Bob is not in both CS100 and CS101 and Jane's statement implies that Bob is not in both CS101 and CS102 are evident from the possible worlds representations (Figures 1 and 2).

Scenario 2. Andy and Amy 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." Note that Andy's statement coincides with source S1. Additionally, his statement implies that Bob is not taking CS102. This fact is *not* evident from the possible worlds representation (Figure 1). Amy's statement is the same as Jane's statement. It coincides with source S2.

In this case, the integration is that Bob is in CS101. The first and last possible worlds in the pairwise union are ruled out as in the previous scenario. Additionally, the second possible world in the pairwise union is ruled out by Andy's statement as well.

3. POSSIBLE-WORLDS MODEL REVISITED

The examples of previous section, in particular Scenario 2, demonstrate that the "pure" possible worlds model is not sufficient for information integration applications. In this section we explore an extension to the possible worlds model that makes it better suited for integration.

3.1 Possible Worlds plus Tuple Set

The model used in [2] adds the set of all tuples to the pure possible worlds model. The following definition is taken from [2]. Note that, for simplicity, each database is assumed to contain a single relation scheme. We adopt the same assumption. Extension to the usual case is straightforward.

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.* ■

The addition of the tuple-set makes it possible to represent some negative information that is not possible to represent in the pure possible-worlds model. For example, consider Scenario 2 in Section 2.2. Recall that Andy's statement implies that Bob is not taking CS102. While this information can not be represented in the pure possible-worlds model (Figure 1), adding that corresponding tuple-set contains (Bob, CS102) makes this information explicit (that Bob is not taking CS 102). We explain this further below.

3.2 Propositional Logic Representation

In this section we present a logical representation of possible-worlds with tuple-set model that helps clarify the semantics and expressive power of this model.

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. (FORMULA CORRESPONDING TO 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,*

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

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

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

EXAMPLE 2. (FORMULA CORRESPONDING TO AN UNCERTAIN DATABASE). *Consider our Scenario 1 in Section 2.2. John's statement was captured in the possible-worlds representation of Figure 1. In this case, the tuple set for source S1 (John) consists of two tuples,* $\{$*(Bob, CS100), (Bob, CS101)*$\}$ *Let variable* x_1 *and* x_2 *correspond to each of these tuples, respectively. Then the formula for the first possible world, second possible world, and the database are, respectively,* $x_1 \wedge \neg x_2$, $\neg x_1 \wedge x_2$, *and* $(x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$.

Now consider Scenario 2. The difference in this case is Andy's knowledge regarding CS102. Here, the tuple set for Andy's information consists of three *tuples,* $\{$*(Bob, CS100), (Bob, CS101), (Bob, CS102)*$\}$. *Let* x_1, x_2, x_3 *correspond to these three tuples. Then the formula corresponding to the uncertain database representing Andy's statement is* $(x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)$. *This is despite the fact that the pure possible worlds representation of Andy's statement is the same as the representation of John's statement in Scenario 1 (Figure 1).*

3.3 Integration Using Logical Representation

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 Scenario 1 of Section 2.2. Uncertain database corresponding to John's statement is represented by formula* $(x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$ *(see Example 2). Similarly, the uncertain database corresponding to Jane's statement is represented by formula* $(x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3)$, *where* x_3 *corresponds to tuple (Bob, CS102). The integration is obtained as*

$$((x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)) \wedge ((x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3)) = (\neg x_1 \wedge x_2 \wedge \neg x_3) \vee (x_1 \wedge \neg x_2 \wedge x_3)$$

which corresponds to possible world databases of Figure 4. Note that the result is consistent with our intuition in Scenario 1: Based on statements by John and Jane, Bob is either taking CS101, or he is taking both CS100 and CS102.

student	course
Bob	CS101

student	course
Bob	CS100
Bob	CS102

Figure 4: Possible Worlds of Scenario 1

In contrast, consider Scenario 2. The uncertain database corresponding to Andy's statement is represented by $(x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)$ *(Example 2). The integration in this case is obtained as*

$$((x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (\neg x_1 \wedge x_2 \wedge \neg x_3)) \wedge ((x_2 \wedge \neg x_3) \vee (\neg x_2 \wedge x_3)) = (\neg x_1 \wedge x_2 \wedge \neg x_3)$$

corresponding to the (in this case, definite) database consisting only of the first database in Figure 4. Again, the result is consistent with our intuition in Scenario 12: Based on statement by Andy and Amy, Bob is taking CS101. ■

3.4 Equivalence of Integration Using Logical Representation and Superset-Containment Integration

In this section we show that our integration approach coincides with a recent formalism for uncertain information integration [2]. First, we review some definitions from [2].

DEFINITION 3. (SUPERSET-CONTAINMENT) *Consider uncertain databases* U_1 *and* U_2. *We say that* U_2 supersetcontains U_1, *denoted* $U_1 \sqsubseteq_S U_2$, *if and only if:*

- $T(U1) \subseteq T(U_2)$, *and*
- $PW(U_1) \supseteq \{W \cap T(U1) \mid W \in PW(U_2)\}$

The following definition is a simplified version of a definition in [2]. We have restricted the definition to identity views and queries, and have concentrated on superset containment.

DEFINITION 4. (CONSISTENT MEDIATED DATABASE) *The set of sources $S = \{S_1, \ldots, S_m\}$ with (uncertain) databases $U_1, \ldots, U_n$ is* consistent *if and only if there exists an uncertain database M such that:*

- $PW(M) \neq \phi$
- $U_i \sqsubseteq_S M$ *for all* $i = 1, \ldots, m$

M is called a consistent mediated database *for S (under superset containment).* ■

Intuitively, a consistent mediated database is a possible integration of the sources. Note that we are assuming each source database consists of a single uncertain relation, and the mediated database is also a single uncertain relation.

EXAMPLE 4. (CONSISTENT MEDIATED DATABASE) *Consider sources S1 and S2 of Figures 1 and 2. It is easy to verify that the database of Figure 4 is indeed a mediated database of S1 and S2 under scenario 1. Note that tuple sets for the sources and for the mediated database are, respectively, $T_1 = \{$(Bob, CS100), (Bob, CS101)$\}$, $T_2 = \{$(Bob, CS101), (Bob, CS102)$\}$, and $T_M = \{$(Bob, CS100), (Bob, CS101), (Bob, CS102)$\}$.* ■

A set of uncertain sources can have many consistent mediated databases. We are interested in a mediated database (which we call the integrated database) that contains all the information implied by sources and nothing more.

DEFINITION 5. (INTEGRATED DATABASE) *Given a set of sources $S = \{S_1, \ldots, S_m\}$, their* integrated database *is a consistent mediated database M for S such that for every consistent mediated database M' for S, $M \sqsubseteq_S M'$.*

Now we will prove that the notions of integration using logical formulas and that of integration using superset containment (Definition 5) coincide.

In the following, let $S_1, \ldots, S_m$ be sources with (uncertain) databases $U_1, \ldots, U_m$. Let the propositional formulas corresponding to $U_1, \ldots, U_m$ be $f_1, \ldots, f_m$. The logical integration results in the uncertain database M with the formula $f_M = f_1 \wedge \cdots \wedge f_m$ (Section 3.3). There is a special case where f_M obtained as above is *false*. In this case the sources are inconsistent and no consistent mediated database exists. Henceforth, we concentrate on the case where $f_M \neq$ *false*.

LEMMA 1. *Let $S = \{S_1, \ldots, S_m\}$ be a set of uncertain sources, and M their integration using logical formulas. Then M is a consistent mediated database for S.*

Proof. *We need to show that $PW(M) \neq \phi$ and $U_i \sqsubseteq_S M$ for all $i = 1, \ldots, m$. The first condition is satisfied for consistent sources since f_M represents at least one world. For the second condition, we should show*

- $T(U_i) \subseteq T(M)$ *for all* $i = 1, \ldots, m$, *and*
- $PW(U_i) \supseteq_S \{W \cap T(U_i) \mid W \in PW(M)\}$ *for all* $i = 1, \ldots, m$.

Consider the form of $f_M = f_1 \wedge \cdots \wedge f_m$. Each f_k has the form

$$f_k = \bigvee_{D_j \in PW(U_k)} \Big(\bigwedge_{t_i \in D_j} x_i \bigwedge_{t_i \notin D_j} \neg x_i \Big)$$

It follows that f_M, after expansion, is the disjunction of conjunctive terms that have the form

$$\Big(\bigwedge_{t_i \in D_j^1} x_i \bigwedge_{t_i \notin D_j^1} \neg x_i \Big) \wedge \ldots \wedge \Big(\bigwedge_{t_i \in D_j^m} x_i \bigwedge_{t_i \notin D_j^m} \neg x_i \Big)$$

Where $D_j^1, \ldots, D_j^m$ are databases in the possible worlds of $U_1, \ldots, U_m$, respectively. Each of these conjunctive terms will be false if for one variable x, both x and $\neg x$ appear in the term. Otherwise, it will have variables corresponding to all tuples in $T(U_1) \cup \cdots \cup T(U_m)$, and it represents one possible world in the logical integration. Since M is nonempty, it follows that $T(M) = T(U_1) \cup \cdots \cup T(U_m)$. Hence, $T(U_i) \subseteq T(M)$ for all $i = 1, \ldots, m$.

Further, if a conjunctive term as above is not false, then it is easy to verify that the possible world corresponding to the conjunction, when restricted to tuples in $T(U_k)$, is represented exactly by the formula

$$\bigwedge_{t_i \in D_j^k} x_i \bigwedge_{t_i \notin D_j^k} \neg x_i$$

and hence is equal to $D_j \in PW(U_k)$. It follows that $PW(U_k) \supseteq_S \{W \cap T(U_k) \mid W \in PW(M)\}$ for all $k = 1, \ldots, m$. ■

THEOREM 1. *Let $S = \{S_1, \ldots, S_m\}$ be a set of uncertain sources, and M their integration using logical formulas. Then M is the integrated database for S.*

Proof. *By Lemma 1 M is a consistent mediated schema for S. We need to prove for every consistent mediated database M' for S, $M \sqsubseteq_S M'$.*

Since M' is a consistent mediated database, we have $U_i \subseteq_S M'$ for all $i = 1, \ldots, m$. Then $T(U_i) \subseteq T(M')$ for all $i = 1, \ldots, m$, and $T(U_1) \cup \cdots \cup T(U_m) \subseteq T(M')$. But in Lemma 1 we showed that $T(M) = T(U_1) \cup \cdots \cup T(U_m)$. It follows that $T(M) \subseteq T(M')$.

It remains to show that $PW(M) \supseteq \{W \cap T(M) \mid W \in PW(M')\}$. Assume the contrary. That is, there exists $W \in PW(M')$ such that $W \cap T(M) \notin PW(M)$. On the other hand, since M' is a consistent mediate database, every uncertain database U_i should have a (regular) relation D_{j_i} such that $D_{j_i} = W \cap T(U_i)$. It is easy to see that the logical construction of the integrated database, when applied to the logical formulas of D_{j_i}, $i = 1, \ldots, m$, generates the logical formula for $W \cap T(M)$. Hence $W \cap T(M) \in PW(M)$, contradiction. ■

4. BEYOND POSSIBLE WORLDS: ADDING PROBABILITIES

We extend the possible-worlds model by associating a probability distribution with each uncertain database:

DEFINITION 6 (PROBABILISTIC UNCERTAIN DATABASE). *A probabilistic uncertain database U consists of a finite set of tuples $T(U)$, a nonempty set of possible worlds $PW(U) = \{D_1, \ldots, D_m\}$, where each $D_i \subseteq T(U)$ is a certain database with a probability p_i, $0 \leq p_i \leq 1$. Further, $\sum_{i=1}^{m} p_i = 1$.* ■

Given a probabilistic uncertain database U with $PW(U) = \{D_1, \ldots, D_m\}$, it is convenient to associate a probabilistic *event* e_i with each possible world D_i. Intuitively, e_i represents the event where the value of the uncertain database U

is equal to D_i. Then, the probability of e_i, $P(e_i) = p_i$. We can make the following observations:

- Each world in the set of possible worlds for a source is exclusive of other worlds for the same source. Hence, we have $P(e_i \mid e_j) = 0$ for all $j \neq i$, where $P(e \mid e')$ represents the conditional probability of event e given event e'.
- Given sources $S_1, \ldots, S_n$, with probabilistic uncertain databases $U_1, \ldots, U_n$, let $PW(U_i) = \{D_{i_1}, \ldots, D_{i_{m_i}}\}$. Each possible world of the integration of $U_1, \ldots, U_n$ corresponds to the conjunction of possible worlds, one from each source. That is, if e_{i_j} is the event associated with possible world D_{i_j} of U_i, then a possible world in the integration corresponds to the event

  $$e_{1_i} \wedge e_{2_j} \wedge \cdots \wedge e_{n_l}$$

  for some $1 \leq i \leq m_1$, $1 \leq j \leq m_2$, ..., $1 \leq l \leq m_n$.
- In the integration, depending on the contents of the sources, two worlds from different sources may be inconsistent. This can be detected using the formulas corresponding to the two worlds, when the conjunction of the two formulas is false. An example was discussed in Example 3. Let the events corresponding to two inconsistent worlds be e_{i_j} and e_{l_k}. Then we have $P(e_{i_j} \mid e_{l_k}) = P(e_{l_k} \mid e_{i_j}) = 0$.
- Note that if two of the worlds forming the integration corresponding to $e_{1_i} \wedge e_{2_k} \wedge \cdots \wedge e_{n_l}$ are inconsistent, then $P(e_{1_i} \wedge e_{2_k} \wedge \cdots \wedge e_{n_l}) = 0$. This is because $P(e_{i_j} \wedge e_{l_k}) = P(e_{i_j} \mid e_{l_k}) \times P(e_{l_k}) = 0$ for any two inconsistent events e_{i_j} and e_{l_k}.

We present an example to demonstrate that probabilities associated with possible worlds of different sources need to satisfy certain constraints. We will generalize this observation in Section 4.1 further below.

EXAMPLE 5. *Consider Scenario 1 of Section 2.2. Possible worlds of the sources are shown again for convenience (Figures 5 and 6). Possible worlds of the integration were shown in Figure 4 (see Example 3).*

student	course
Bob	CS100

D1

student	course
Bob	CS101

D2

Figure 5: Possible Worlds of source S1

student	course
Bob	CS101

D3

student	course
Bob	CS102

D4

Figure 6: Possible Worlds of source S2

Let the events associated with possible worlds D1, ..., D4 be $e_1, \ldots, e_4$, respectively. We have the following equations:
$P(e_1) + P(e_2) = 1$
$P(e_3) + P(e_4) = 1$
$P(e_1 \mid e_2) = P(e_2 \mid e_1) = 0$
$P(e_3 \mid e_4) = P(e_4 \mid e_3) = 0$

Further, by Example 3, we know e_1 and e_3 are inconsistent, and so are e_2 and e_4. Hence
$P(e_1 \mid e_3) = P(e_3 \mid e_1) = 0$
$P(e_2 \mid e_4) = P(e_4 \mid e_2) = 0$
Our goal is to compute $P(e_1 \wedge e_3)$, $P(e_1 \wedge e_4)$, $P(e_2 \wedge e_3)$, and $P(e_2 \wedge e_4)$ in terms of $P(e_1)$, $P(e_2)$, $P(e_3)$, and $P(e_4)$.
It is easy to verify that $P(e_1 \wedge e_3) = P(e_1 \mid e_3) \times P(e_3) = 0$. Similarly, $P(e_2 \wedge e_4) = P(e_2 \mid e_4) \times P(e_4) = 0$. The only possible worlds in the integration with nonzero probabilities correspond to $e_1 \wedge e_4$ and $e_2 \wedge e_3$. These possible worlds were shown in Figure 4.

Some Observations

First, we note that e_1 and e_2 are complementary events since they are mutually exclusive and $P(e_1)+P(e_2) = 1$. So, we can write $e_1 = \neg e_2$ and vice-versa. Similarly, e_3 and e_4 are complementary. Hence, we have $P(e_3) = P(e_3 \wedge e_1) + P(e_3 \wedge e_2)$. But $P(e_3 \wedge e_1) = 0$ as discussed above. So, we get $P(e_3) = P(e_3 \wedge e_2) = P(e_2 \wedge e_3) = P(e_2 \mid e_3) \times P(e_3)$. It follows that $P(e_2 \mid e_3) = 1$. Similarly, we can show $P(e_3 \mid e_2) = 1$, $P(e_1 \mid e_4) = 1$, and $P(e_4 \mid e_1) = 1$.

Finally, since $P(e_3) = P(e_3 \wedge e_2) = P(e_3 \mid e_2) \times P(e_1)$ and $P(e_3 \mid e_2) = 1$, we obtain $P(e_2) = P(e_3)$. We can also obtain $P(e_1) = P(e_4)$ in a similar fashion.

This is a very interesting result. It shows that probabilities associated with possible worlds D_1 and D_2 of source 1 have to be the same as probabilities associated with worlds D_4 and D_3 of source 2. Let's recall our Scenario 1: The information that Bob is taking CS100 or (xor) CSC101 was given by John, and the information that Bob is taking CS101 or (xor) CSC102 was given by Jane. The restriction that probabilities associated with possible worlds D_1 and D_2 of source 1 have to be the same as probabilities associated with worlds D_3 and D_4 of source 2 means that If John assigns probabilities, say, 40% and 60%, to his statements, then Jane also has to assign exactly the same probabilities to her two statements (40% that Bob is taking CS102, and 60% that Bob is taking CS101). This is highly counter intuitive. We will generalize this observation, which we will call probabilistic consistency constraint*, in the next section.*

Let's assume for now that indeed $P(e_1) = P(e_4)$ and $P(e_2) = P(e_3)$. Then it becomes easy to obtain the probabilities of the possible worlds in the integration:
$P(e_1 \wedge e_4) = P(e_1 \mid e_4) \times P(e_4) = P(e_4)$, *and*
$P(e_2 \wedge e_3) = P(e_2 \mid e_3) \times P(e_3) = P(e_3)$
To summarize, For Scenario 1: $P(e_1 \wedge e_3) = 0$, $P(e_1 \wedge e_4) = P(e_1) = P(e_4)$, $P(e_2 \wedge e_3) = P(e_2) = P(e_3)$, and $P(e_2 \wedge e_4) = 0$. ■

4.1 Probabilistic Consistency Constraints

Given the possible world sets of two information sources S and S', we use a bi-partite graph G to represent the consistency relation between the worlds of the sources. Let the possible world set of S be $\{D_1, \ldots, D_k\}$, and the possible world set of S' be $\{D'_1, \ldots, D'_{k'}\}$. Graph G has $k + k'$ nodes corresponding to possible world sets $\{D_1, \ldots, D_k\}$ and $\{D'_1, \ldots, D'_{k'}\}$. There is an edge between D_i and D'_j if the formulas $f(D_i)$ and $f(D'_j)$ corresponding to these worlds are mutually satisfiable. That is, $f(D_i) \wedge f(D'_j) \neq$ *false*. (See Section 3.2 regarding logical representation of possible worlds.) We call G the *possible-worlds consistency graph* (or *consistency graph* for short) of S and S'.

In the following, we overload possible world symbol D (or D') to also represent the event corresponding to D (or D')

(Section 4) as well as the node corresponding to D (or D') in the consistency graph G. The actual meaning will be clear from context.

We will prove the following:

- If a node D (or D') has no edges connected to it, then $P(D)$ (or $P(D')$) must be zero.
- Connected components of G are *complete* bi-partite graphs.
- Let G_1 be a connected component of G. Let N and N' denote the nodes of G_1 corresponding to the possible worlds of sources S and S', respectively. Then

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

- As a special case, if D and D' are connected in G, and not connected to any other node, then $P(D) = P(D')$. We had this case in Example 5.

THEOREM 2. *Let G be the consistency graph of sources S and S'. Let G_1 be a connected component of G. Then G_1 is a complete bi-partite graph.*

Proof: *We will show that if there is a path between nodes D and D' in G, where D and D' correspond to possible worlds of S and S', respectively, then there must be an edge between D and D'.*

Assume there is a path between D and D') but there is no edge (D, D'). Since there is no edge (D, D'), then $f(D)$ and $f(D')$ are not mutually satisfiable. That means there is a variable t such that one formula has t and the other has $\neg t$ in their conjuncts. Without loss of generality, assume $f(D)$ has t. Hence, (1) there is a tuple t in D (recall that a variable represents a tuple), (2) t is not in D', and (3) t is in the tuple set of S'.

The path between D and D' consists of a sequence of alternating nodes D_is and D'_js such that $D = D_1$, $D'_k = D'$, and the following edges are in G:

(D_i, D'_i) *for* $i = 1, \ldots, k$, *and*

(D'_i, D_{i+1}) *for* $i = 1, \ldots, k-1$.

Since t is in $D = D_1$ and edge (D_1, D'_1) exists, then t must also be in D'_1. Similarly, since edge (D'_1, D_2) exists, then t must also be in D_2. In a similar way we can show t must be in all possible worlds along this path, including D'. This provides a contradiction to the assumption that $\neg t$ is in $f(D')$. ■

THEOREM 3. *Let G_1 be a connected component of G. Let N and N' denote the nodes of G_1 corresponding to the worlds of sources S and S', respectively. Then*

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

Proof: *Let k and k' be the number of possible worlds for S and S', respectively. We have*

$$P(D_i) = \sum_{j=1}^{k'} P(D_i \wedge D'_j)$$

If $D_i \in N$, then $P(D_i \mid D_j)$ and $P(D_i \wedge D_j)$ are both zero for $D_j \notin N'$. So, for a node $D_i \in N$, we have

$$P(D_i) = \sum_{D'_j \in N'} P(D_i \wedge D'_j)$$

It follows that

$$\sum_{D_i \in N} P(D_i) = \sum_{D_i \in N} \sum_{D'_j \in N'} P(D_i \wedge D'_j)$$

Similarly, we can show, for a node $D'_j \in N'$,

$$P(D_j) = \sum_{D_i \in N} P(D_i \wedge D'_j)$$

and hence,

$$\sum_{D'_j \in N'} P(D'_j) = \sum_{D_i \in N} \sum_{D'_j \in N'} P(D_i \wedge D'_j)$$

Theorem follows. ■

EXAMPLE 6. *Consider the possible worlds of two sources shown in Figures 7 and 8.*

D1

student	course
Bob	CS100

D2

student	course
Bob	CS100
Bob	CS101

D3

student	course
Bob	CS101

Figure 7: Possible Worlds of source One

D'1

student	course
Bob	CS100

D'2

student	course
Bob	CS100
Bob	CS201

D'3

student	course
Bob	CS201

D'4

student	course
Bob	CS201
Bob	CS202

Figure 8: Possible Worlds of source Two

The consistency graph G for these sources is shown in Figure 9. Note that G has two connected components: G_1 consists of nodes D_1, D_2, D'_1, D'_2 and G_2 consists of nodes D_3, D'_3, D'_4. Note that these connected components are complete bipartite graphs. According to Theorem 3, we must have $P(D_1) + P(D_2) = P(D'_1) + P(D'_2)$, and $P(D_3) = P(D'_3) + P(D'_4)$.

The result of integration will consist of 6 possible-worlds relations, corresponding to compatible pairs of possible worlds from the two sources, namely, (D_1, D'_1), (D_1, D'_2), (D_2, D'_1), (D_2, D'_2), (D_3, D'_3), and (D_3, D'_4). It is shown in Figure 10. We will discuss the problem of determining probabilities of possible-worlds relations of the integration in the next section. ■

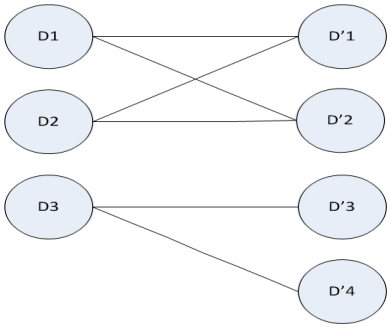

Figure 9: Consistency Graph for Example 6

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)

Figure 10: Result of integration of sources One and Two

4.2 Determining the Probabilities for the Result of Integration

In the integration of probabilistic uncertain databases, the goal is to obtain the probabilities of the possible worlds of the integration in terms of the probabilities of individual worlds (p_{i_j}'s, or, equivalently, $P(e_{i_j})$'s). In other words, we would like to obtain

$$P(e_{1_i} \wedge e_{2_k} \wedge \cdots \wedge e_{n_l})$$

in terms of $P(e_{i_j})$'s for all possible values of i, k, ..., and l. We discussed a simple case in Example 5. If worlds D and D' are connected in the consistency graph, and not connected to any other nodes, then (1) $P(D) = P(D')$ and (2) $P(D \wedge D') = P(D) = P(D')$ (note that we are overloading the symbol D to mean a possible world, the node in the consistency graph corresponding to the possible world D, and the event corresponding to the possible world D.) This case corresponds to simple connected components in the consistency graph, those with only two nodes.

We face two difficulties in determining the probabilities of possible worlds in the integration in the general case:

- How to determine the probabilities of possible worlds resulting from the integration of pairs of worlds in a connected component with more than two nodes?
- How to deal with violations of probabilistic consistency constraints?

In this section we discuss approaches to address these difficulties. First, we concentrate on the case where probabilistic consistency constraints are satisfied. The following lemma follows the proof of Theorem 3.

LEMMA 2. *Let G be the consistency graph of sources S and S', and G_1 be a connected component of G. Let the nodes of G_1 consist of nodes N corresponding to source S and N' corresponding to source S'. Then*

$$\sum_{D_i \in N} \sum_{D'_j \in N'} P(D_i \wedge D'_j) = \sum_{D_i \in N} P(D_i) = \sum_{D_j \in N'} P(D_j)$$

Note that each $D_i \wedge D'_j$ corresponds to a possible world in the integration. In other words, we know the sum *of the probabilities of possible worlds resulting from the integration corresponding to each connected component of the consistency graph.* ■

It is easy to extend the observation from Example 5 regarding connected components with only two nodes to the case where N or N' (in Lemma 2) is a single node. For example, if $N = \{D\}$ is a single node, then $P(D \wedge D'_j) = P(D'_j)$ for all $D'_j \in N'$. However, if neither N nor N' is a single node, then, in the absence of additional information, there are multiple (infinite) possibilities for the probabilities of the possible worlds in the integration (see Example 7). We can use the following approaches to compute the probabilities of possible worlds in the integration.

- Additional information can be used to pinpoint the probabilities of integrated possible worlds. One such information is the conditional probabilities of the form $P(D \mid D')$ for possible worlds in the same connected component of the consistency graph of the two sources.
- In the absence of any other information we adapt some heuristics to distribute the known sum of the probabilities over individual possible worlds in the integration (see Example 7).

EXAMPLE 7. *Consider Example 6. Let's assume the probabilities of possible worlds of sources S and S' 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 $P(D'_4) = 0.15$. Note that these probabilities satisfy the probabilistic consistency constraints $P(D_1) + P(D_2) = P(D'_1) + P(D'_2)$, and $P(D_3) = P(D'_3) + P(D'_4)$.*

The consistency graph for S and S' was shown in Figure 9. It has two connected components. The second component satisfies the special case where N is a single node. So, we can easily determine the probabilities of two of the possible worlds in the integration: $P(D_3 \wedge D'_3) = P(D'_3) = 0.05$ and $P(D_3 \wedge D'_4) = P(D'_4) = 0.15$.

The first connected component of the consistency graph gives rise to possible worlds corresponding to $D_1 \wedge D'_1$, $D_1 \wedge D'_2$, $D_2 \wedge D'_1$, and $D_2 \wedge D'_2$. We have four unknowns. We have four equations too but they are not independent:

$$P(D_1 \wedge D'_1) + P(D_1 \wedge D'_2) = P(D_1),$$
$$P(D_2 \wedge D'_1) + P(D_2 \wedge D'_2) = P(D_2),$$
$$P(D_1 \wedge D'_1) + P(D_2 \wedge D'_1) = P(D'_1),$$
$$P(D_1 \wedge D'_2) + P(D_2 \wedge D'_2) = P(D'_2).$$

Given one more equation, we will be able to determine the four probabilities. An attractive possibility is the knowledge of a conditional probability. For example, let $P(D_1 \mid D'_1) = 0.6$. *Then we get* $P(D_1 \wedge D'_1) = 0.21$, $P(D_1 \wedge D'_2) = 0.09$, $P(D_2 \wedge D'_1) = 0.14$, *and* $P(D_2 \wedge D'_2) = 0.36$.

In the absence of any additional knowledge, we can incorporate some heuristics to provide the probabilities of possible worlds in the integration. One possibility is to distribute the sum (0.8 in our example) according to the pairwise product of probabilities of underlying possible worlds (0.105, 0.135, 0.175, and 0.225 in our example). We obtain the following values for our example: $P(D_1 \wedge D'_1) = 0.13125$, $P(D_1 \wedge D'_2) = 0.16875$, $P(D_2 \wedge D'_1) = 0.21875$, *and* $P(D_2 \wedge D'_2) = 0.28125$. *Note that this distribution corresponds to the assumption*

$$P(D1 \mid D'1) = \frac{P(D_1)}{P(D_1) + P(D_2)} \qquad \blacksquare$$

Next, we consider the second difficulty: How to cope with violations of probabilistic consistency constraints? A strict approach would be to declare the sources as inconsistent and deem integration impossible. But in many applications, probabilities associated with possible worlds are approximate, and may represent some *degree of belief* rather than a strict mathematical probability. Rather than dismissing the sources' assessment completely when their probabilities violate some probabilistic constraint, we should try to provide approximate probabilities for possible worlds in the integration. The following example demonstrates a promising approach.

EXAMPLE 8. *Consider Example 5 again. The consistency graph corresponding to this example has two connected components:* (D_1, D_4) *and* (D_2, D_3), *where* D_1 *and* D_2 *are possible worlds of source 1 and* D_3 *and* D_4 *are possible worlds of source 2.*

Assume probabilities 40% and 60% for possible worlds of source 1, and 70% and 30% for the possible worlds of source 2. That is $P(D_1) = 0.4$, $P(D_2) = 0.6$, $P(D_3) = 0.7$, $P(D_4) = 0.3$. *Note that probabilistic consistency constraints are violated. That is,* $P(D_1) \neq P(D_4)$ *and* $P(D_2) \neq P(D_3)$. *We showed that* $P(D_1 \wedge D_4)$ *must be equal to* $P(D_1)$ *and* $P(D_4)$ *(and also* $P(D_1) = P(D_4)$*). But that is not satisfied here. So,* intuitively, *what is the best probability to associate with* $P(D_1 \wedge D_4)$*? Source 1 says it must be 40%, while source 2 says it is 30%. One possible choice is the average of the two values, 35%. This corresponds to the first possible world of the integration in Figure 4. Similarly, we use the average of 60% and 70%, namely 65%, for the probability of the second possible world of the integration in Figure 4.* ■

The following is the generalization of our heuristics:

Consider each connected component of the consistency graph independently. The probabilistic consistency constraint is that the sum of probabilities of the possible worlds corresponding to the first source is equal to the sum of probabilities of the possible worlds corresponding to the second source. If they are not equal, we use the average of the two values. Then we will distribute this value into probabilities for each "edge" of the connected component (note that an edge represents the integrated possible world that is the union of the two possible worlds represented by the nodes of that edge).

5. RELATED WORK

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 (for example, some recent publications in these areas include [2, 3, 4, 7, 10, 12, 15, 16, 17, 19, 20, 23, 24, 26, 28, 31, 39, 41, 42, 44, 45, 46, 47] Research on information integration with uncertainty, on the other hand, is quite recent [2, 18, 19, 20, 21, 34, 35]. In the following we review some of related research in these areas.

We start with a fundamental work on information integration with uncertainty [2]. The focus of this work is on the foundations for local-as-view information integration when the sources contain uncertain data. To our knowledge, this is the first work that enhances the well-known possible-worlds model with tuple sets, a concept that proves critical for information integration applications. They introduce fundamental concepts of equality and superset containment, consistency, and correct and strongest correct answer to queries. They prove that for superset containment, an uncertain database instance exists that gives the strongest correct answer to any query. We have chosen this instance as the formalism for the integration of uncertain databases, and provide a conceptually simpler algorithm to compute it. [2] also contains studies of complexities of checking consistency and evaluating queries in this model.

Most of the other research on information integration with uncertainty deal with the issue of uncertain schema mapping. Schema matching/mapping is a key step in information integration. Significant research has been spent on automated schema mapping (for example, [5, 6, 13, 14, 27, 29, 32, 33, 36, 37, 40, 48]). Automated schema mapping generates uncertain mappings and hence the study of uncertain mappings is important for information integration.

In [18, 19] authors argue that data integration systems should handle uncertainty at three levels: Semantic mappings, translation of keyword queries to structured queries, and data. They introduce the concepts of by-table and by-tuple probabilistic schema mappings. In by-table semantics, all tuples of a given table use the same mapping to the mediated schema, whereas in the by-tuple semantics, each tuple may use a different mapping. [9, 18, 19] present query complexity and algorithms for query processing over inexact schema mappings, as well as algorithms for efficiently computing top-k answers to queries.

A study and survey of uncertainty in data integration is presented in [34, 35]. They introduce a generic data integration process in terms of *wrapping*, *matching* and *merging* operations, and identify critical points for uncertain information integration. They demonstrate the importance of handling uncertainties produced by the matching process, and introduce approaches to improve the matching phase. They also present a survey of a number of data integration methods that explicitly deal with uncertainty.

The authors of [21] concentrate on another dimension of using probabilistic information in information integration. Their goal is "to improve the performance of a mediator system by obtaining answers to queries as fast as possible." They consider three types of probabilistic information: (1) degree of overlap between collections in the mediated schema, (2) degree of coverage of each information source,

and (3) degree of overlap between information sources (such as when a source is a *view* of another source.) They propose algorithms to order access to information sources using probabilistic information and study their performances.

Management of uncertainty in XML schema matching and efficient query processing in this context has been studied in [22]. Authors present approaches to improve the efficiency of generating, storing, and querying possible mappings.

Information integration from uncertain XML sources is studied in [20]. The underlying model for uncertain XML is the ProTDB probabilistic XML model of [38], and the integration system uses the *Semantic Model* approach of [7, 8, 30, 43]. It introduces the *lineage-encoding* relational model for the mediated schema which makes it possible to capture the uncertainties in the ProTDB-style probabilistic XML sources, and presents query processing and probability calculation algorithms for this framework.

6. CONCLUSIONS

We studied the problem of information integration from multiple sources with uncertain data. We reviewed the possible-worlds approach to the modeling of uncertain information, presented an intuitive logical representation of possible-worlds information, and presented an algorithm for uncertain information integration using the logical representation. We also showed that our integration approach coincides with a recent formalism for uncertain information integration. Then we extended the approach to probabilistic possible-worlds, and showed certain probabilistic consistency constraints are imposed upon possible-worlds of information sources in the integration. We studied the problem of determining the probabilities for the result of integration, and presented a practical approach to probabilistic information integration based on relaxation of probabilistic constraints.

Our work has established a formal model for the integration of information from sources containing uncertain data, including probabilistic information about sources contents. To be practical, uncertain information should be presented using a compact representation rather than an enumeration of the possible-worlds. Future work includes selecting an efficient representation (such as the probabilistic relation of [12]) and designing efficient algorithms for integration of uncertain information in the selected representation.

7. REFERENCES

[1] Serge Abiteboul, Paris C. Kanellakis, and Gösta Grahne. On the representation and querying of sets of possible worlds. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 34–48, 1987.

[2] Parag Agrawal, Anish Das Sarma, Jeffrey D. Ullman, and Jennifer Widom. Foundations of uncertain-data integration. *Proceedings of International Conference on Very Large Databases*, 3(1):1080–1090, 2010.

[3] Lyublena Antova, Thomas Jansen, Christoph Koch, and Dan Olteanu. Fast and simple relational processing of uncertain data. In *Proceedings of IEEE International Conference on Data Engineering*, pages 983–992, 2008.

[4] Lyublena Antova, Christoph Koch, and Dan Olteanu. 10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information. In *Proceedings of IEEE International Conference on Data Engineering*, pages 606–615, 2007.

[5] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors. *Schema Matching and Mapping*. Springer, 2011.

[6] Philip A. Bernstein and Sergey Melnik. Model management 2.0: manipulating richer mappings. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 1–12, 2007.

[7] Dongfeng Chen, Rada Chirkova, Maxim Kormilitsin, Fereidoon Sadri, and Timo J. Salo. Query optimization in XML-based information integration. In *Proceedings of International Conference on Information and Knowledge Management*, pages 1405–1406, 2008.

[8] Dongfeng Chen, Rada Chirkova, Fereidoon Sadri, and Timo J. Salo. Query optimization in information integration, December 2010. Submitted for publication.

[9] Reynold Cheng, Jian Gong, David W. Cheung, and Jiefeng Cheng. Evaluating probabilistic queries over uncertain matching. In *Proceedings of IEEE International Conference on Data Engineering*, pages 1096–1107, 2012.

[10] Nilesh N. Dalvi, Christopher Ré, and Dan Suciu. Probabilistic databases: diamonds in the dirt. *Communications of the ACM*, 52(7):86–94, 2009.

[11] Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *Proceedings of International Conference on Very Large Databases*, pages 864–875, 2004.

[12] Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal*, 16(4):523–544, 2007.

[13] Robin Dhamankar, Yoonkyong Lee, AnHai Doan, Alon Y. Halevy, and Pedro Domingos. iMAP: Discovering complex mappings between database schemas. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 383–394, 2004.

[14] Hong Hai Do and Erhard Rahm:. COMA - a system for flexible combination of schema matching approaches. In *Proceedings of International Conference on Very Large Databases*, pages 610–621, 2002.

[15] AnHai Doan and Alon Y. Halevy. Semantic integration research in the database community: A brief survey. *AI Magazine*, 26(1):83–94, 2005.

[16] AnHai Doan, Natalya F. Noy, and Alon Y. Halevy. Introduction to the special issue on semantic integration. *SIGMOD Record*, 33(4):11–13, 2004.

[17] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[18] Xin Luna Dong, Alon Halevy, and Cong Yu. Data integration with uncertainty. In *Proceedings of International Conference on Very Large Databases*, pages 687–698, 2007.

[19] Xin Luna Dong, Alon Y. Halevy, and Cong Yu. Data integration with uncertainty. *The VLDB Journal*, 18(2):469–500, 2009.

[20] Ala A. Eshmawi and Fereidoon Sadri. Information integration with uncertainty. In *Proceedings of*

International Database Engineering and Applications, IDEAS, pages 284–291, 2009.

[21] Daniela Florescu, Daphne Koller, and Alon Y. Levy. Using probabilistic information in data integration. In *Proceedings of International Conference on Very Large Databases*, pages 216–225, 1997.

[22] Jian Gong, Reynold Cheng, and David W. Cheung. Efficient management of uncertainty in xml schema matching. *The VLDB Journal*, 21(3):385–409, 2012.

[23] Todd J. Green, Gregory Karvounarakis, Nicholas E. Taylor, Olivier Biton, Zachary G. Ives, and Val Tannen. ORCHESTRA: facilitating collaborative data sharing. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 1131–1133, 2007.

[24] Laura M. Haas. Beauty and the beast: The theory and practice of information integration. In *Proceedings of International Conference on Database Theory*, pages 28–43, 2007.

[25] Alon Halevy and Chen Li. Information integration research: Summary of NSF IDM workshop breakout session. `http://www2.cs.washington.edu/nsf2003/final-reports/information integration breakout summary.pdf`, September 2003.

[26] Alon Y. Halevy, Naveen Ashish, Dina Bitton, Michael J. Carey, Denise Draper, Jeff Pollock, Arnon Rosenthal, and Vishal Sikka. Enterprise information integration: successes, challenges and controversies. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 778–787, 2005.

[27] Alon Y. Halevy, Zachary G. Ives, Dan Suciu, and Igor Tatarinov. Schema mediation for large-scale semantic data sharing. *The VLDB Journal*, 14(1):68–83, 2005.

[28] Alon Y. Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. *Proceedings of International Conference on Very Large Databases*, pages 9–16, 2006.

[29] Mauricio A. Hernández, Paolo Papotti, and Wang Chiew Tan. Data exchange with data-metadata translations. *Proceedings of the VLDB Endowment*, 1(1):260–273, 2008.

[30] Laks V. S. Lakshmanan and Fereidoon Sadri. Interoperability on XML data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 146–163, 2003.

[31] Chen Li, Jia Li, and Qi Zhong. Raccoon: A peer-based system for data integration and sharing. In *Proceedings of IEEE International Conference on Data Engineering*, page 852, 2004.

[32] Jayant Madhavan, Philip A. Bernstein, AnHai Doan, and Alon Y. Halevy. Corpus-based schema matching. In *Proceedings of IEEE International Conference on Data Engineering*, 2005.

[33] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. Generic schema matching with Cupid. In *Proceedings of International Conference on Very Large Databases*, pages 49–58, 2001.

[34] Matteo Magnani and Danilo Montesi. Uncertainty in data integration: current approaches and open problems. In *Proceedings of VLDB Workshop on Management of Uncertain Data*, pages 18–32, 2007.

[35] Matteo Magnani and Danilo Montesi. A survey on uncertainty management in data integration. *ACM Journal of Data and Information Quality*, 2(1), 2010.

[36] Sergey Melnik, Atul Adya, and Philip A. Bernstein. Compiling mappings to bridge applications and databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 461–472, 2007.

[37] Sergey Melnik, Philip A. Bernstein, Alon Y. Halevy, and Erhard Rahm. Supporting executable mappings in model management. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2005.

[38] Andrew Nierman and H. V. Jagadish. ProTDB: Probabilistic data in XML. In *Proceedings of International Conference on Very Large Databases*, pages 646–657, 2002.

[39] Dan Olteanu, Jiewen Huang, and Christoph Koch. SPROUT: Lazy vs. eager query plans for tuple-independent probabilistic databases. In *Proceedings of IEEE International Conference on Data Engineering*, pages 640–651, 2009.

[40] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[41] Christopher Re, Nilesh N. Dalvi, and Dan Suciu. Efficient top-k query evaluation on probabilistic data. In *Proceedings of IEEE International Conference on Data Engineering*, pages 886–895, 2007.

[42] Patricia Rodríguez-Gianolli, Maddalena Garzetti, Lei Jiang, Anastasios Kementsietsidis, Iluju Kiringa, Mehedi Masud, Renée J. Miller, and John Mylopoulos. Data sharing in the Hyperion peer database system. In *Proceedings of International Conference on Very Large Databases*, pages 1291–1294, 2005.

[43] Fereidoon Sadri. Local as view information integration in the semantic model approach. In *Proceedings of IEEE International Conference on Information Reuse and Integration (IRI)*, pages 148–153, 2011.

[44] Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, Shubha U. Nabar, and Jennifer Widom. Representing uncertain data: models, properties, and algorithms. *The VLDB Journal*, 18(5):989–1019, 2009.

[45] Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, and Jennifer Widom. Working models for uncertain data. In *Proceedings of IEEE International Conference on Data Engineering*, page 7, 2006.

[46] Prithviraj Sen and Amol Deshpande. Representing and querying correlated tuples in probabilistic databases. In *Proceedings of IEEE International Conference on Data Engineering*, 2007.

[47] Nicholas E. Taylor and Zachary G. Ives. Reliable storage and querying for collaborative data sharing systems. In *Proceedings of IEEE International Conference on Data Engineering*, pages 40–51, 2010.

[48] Ling-Ling Yan, Renée J. Miller, Laura M. Haas, and Ronald Fagin. Data-driven understanding and refinement of schema mappings. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 485–496, 2001.