ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

A Relational Database Model with Interval Probability Valued Attributes for Uncertain and Imprecise Information

Hoa Nguyen[1] and Duy Nhat Le[2]

ABSTRACT

Although the conventional relational database model (CRDB) is beneficial to model, design, and implement large-scale systems, it is limited to express and deal with uncertain and imprecise information. In this paper, we introduce a new relational database model as an extension of CRDB where relational attributes may take a value associated with a probability interval, named IPRDB, for representing and handling uncertain and imprecise information in practice. To build IPRDB, we employ three key methods: (1) Probabilistic values of data types are proposed for expressing uncertain and imprecise valued attributes; (2) the probabilistic interpretations of binary relations on sets and operators on probability intervals are used for computing the uncertain degree of functional dependencies, keys, and relations on value domains of attributes; and (3) the combination strategies of probabilistic values are defined for developing new relational algebraic operations. Then, fundamental concepts of the model, such as schemas, probabilistic relations, and probabilistic relational databases, are extended coherently and consistently with those of the conventional relational database model. A set of the properties of the basic probabilistic relational algebraic operations is also formulated and proven. The built IPRDB model can represent and manipulate effectively uncertain and imprecise information in real-world applications.

Article information:

Keywords: Probability Interval, Probabilistic Interpretation, Probabilistic Value, Probabilistic Relation, Probabilistic Relational Algebraic Operations, Probabilistic Relational Database

Article history:
Received: February 13, 2024
Revised: May 16, 2024
Accepted: May 30, 2024
Published: June 15, 2024 (Online)

DOI: 10.37936/ecti-cit.2024183.255697

1. INTRODUCTION

As shown in [1], [2], and [3], the conventional relational database model (CRDB) is beneficial to model, design, and implement large-scale systems. Still, it is limited to represent and handle uncertain and imprecise information in practice. Currently, there have been many non-conventional database models, including probabilistic relational database models (PRDB), studied and built to overcome the limitations of CRDB. For instance, in [4], authors proposed a PRDB model to compute the uncertain membership degree of each tuple in a relation, and in [5], authors introduced another PRDB model that can compute the uncertain degree of attribute values of each tuple in a relation. Probabilistic database models also have been used in many real applications, such as the works in [6], [7], and [8]. Notably, in [6], probabilistic queries were employed to express and handle uncertain multidimensional data. In [7], probabilistic databases were applied for detecting faulty sensors. And in [8], queries over the relational cross model were processed by using uncertain databases.

However, no model would be so universal that it could include all measures and tackle all aspects of uncertainty and imprecision of information in the real world.

Probabilistic relational database models are developed and built as extensions of CRDB based on the probability theory. There are two main classes of PRDB models extended from the CRDB model. The first one defines a probabilistic relation as a set of tuples such that each tuple is associated with a probability to represent its uncertainty degree in the relation. The second one defines a probabilistic relation as a set of tuples such that each tuple attribute is associated with a probability to express the uncertainty

[1] The author is with the Information Technology Faculty, Saigon University, Vietnam. E-mail: nguyenhoa@sgu.edu.vn
[1,2] The authors are with the Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam. E-mail: nguyenhoa@sgu.edu.vn and DuyLN@iuh.edu.vn
[1] Corresponding author: nguyenhoa@sgu.edu.vn

degree of the attribute value.

For the first PRDB model class, as the works in [9-14], each tuple of a relation was associated with a probability in the interval [0, 1] to represent the uncertainty membership degree of that tuple for the relation and the uncertainty degree of the attribute values of a tuple inferred from the uncertainty membership degree of that tuple. However, in many natural situations, we do not know precisely the probability as a number in the interval [0, 1]. We can only estimate it as an approximate number in a subinterval of [0, 1]. The extended models in [15-17] overcame the shortcoming by associating each tuple with a probability interval.

For the second PRDB model class, as in [18] and [19], each value of an attribute was assigned to a probability in the interval [0, 1] to represent the uncertain level for that attribute taking the value. More flexibly and generally, in [20], each attribute was associated with a probability distribution function on a set of values to express the possibility that the attribute might take one of the values of the set with a distributed probability. However, in many real cases, we cannot define precisely the probability distribution function for each value in a set. We can only estimate it to be an approximate number in a subinterval of [0, 1]. The model in [21] overcame the restriction by using a pair of lower and upper-bound probability distribution functions to represent the possibility of an attribute taking a value in a set with a computed probability interval from the distribution function pair. The models in [22] and [23] extended the model in [21] for uncertain multivalued attributes. However, when the probabilistic relations have many attributes, the number of generated probability distribution functions is too large to lead to low performance in manipulating data of the models.

In this paper, we propose a new probabilistic relational database model, abbreviated to IPRDB, as an extension of CRDB with interval probability valued attributes for uncertain and imprecise information to overcome the limitations of models in [20], [22], and [23]. The proposed IPRDB model is consistent with the CRDB model and more flexible than the models in [20], [22], and [23] by using probability intervals instead of probability single values and distribution functions.

To build IPRDB, we adapt probabilistic values in [24] with data types for representing uncertain and imprecise valued attributes of relations, employ probabilistic interpretations of binary relations on sets in [25] and operators on probability intervals in [23], and propose new combination strategies of probabilistic values to define the probabilistic relational algebraic operations for computing and querying uncertain and imprecise information on IPRDB relations. The built IPRDB model can represent and manipulate effectively uncertain and imprecise information and it can be applied to solve the real problems like [26].

The mathematical basis to build IPRDB is presented in Section 2. The IPRDB data model, including the schema, relation, database, probabilistic functional dependency, and the relational schema key is introduced in Section 3. Section 4 introduces probabilistic relational algebraic operations on IPRDB and their properties. Section 5 presents the achieved results and discussions of the IPRDB model. Finally, Section 6 concludes the paper and outlines further research directions.

2. PROBABILITY DEFINITIONS AND NOTIONS

This section presents probability definitions and notions as the mathematical bases to build IPRDB for representing and handling uncertain and imprecise information.

2.1 Probabilistic Values

The probabilistic value in [24] is adapted with data types to express uncertain and imprecise valued attributes in IPRDB as the following definition.

Definition 2.1 Let τ be a data type and D be the domain of τ, a *probabilistic value* on the domain of τ is a finite set of pairs $\{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$, where $v_i \in D$ and $0 \leq l_i \leq u_i \leq 1$, for every $i = 1, 2, \ldots, m$.

Informally, a probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ says that pv's value is one member of the set $\{v_1, \ldots, v_m\}$ and the probability that pv's value is v_i lies in the interval $[l_i, u_i]$. Thus, a probabilistic value represents both the uncertainty of its value and the imprecision of the probability of that value. A probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ corresponds with a probability distribution function p over $V = \{v_1, \ldots, v_m\}$ such that $p(v_i) \in [l_i, u_i]$ and $\sum_{v_i \in V} p(v_i) \leq 1$.

Example 2.1 Suppose a patient's disease is diagnosed as hepatitis with a probability between 0.5 and 0.7 or cholecystitis with a probability between 0.3 and 0.5. Then, this information may be represented by the probabilistic value {(hepatitis, [0.5, 0.7]), (cholecystitis, [0.3, 0.5])}.

We note that a probabilistic value can be denoted by $pv = \{(v, I) | v \in D, I = [l, u] \subseteq [0, 1]\}$.

2.2 Probabilistic Interpretation of Binary Relations on Sets

To compute the uncertain degree of relations on attribute values in IPRDB, we use the probabilistic interpretation of binary relations on sets in [25] as below.

Table 1: *Definitions of probabilistic combination strategies.*

Strategy	Operators
Ignorance	$([l_1,u_1] \otimes_{ig} [l_2,u_2]) = [\max(0, l_1+l_2–1), \min(u_1,u_2)]$ $([l_1,u_1] \oplus_{ig} [l_2,u_2]) = [\max(l_1,l_2), \min(1, u_1+u_2)]$ $([l_1,u_1] \ominus_{ig} [l_2,u_2]) = [\max(0, l_1–u_2), \min(u_1, 1–l_2)]$
Independence	$([l_1,u_1] \otimes_{in} [l_2,u_2]) = [l_1.l_2, u_1.u_2]$ $([l_1,u_1] \oplus_{in} [l_2,u_2]) = [l_1+l_2–(l_1.l_2), u_1+u_2–(u_1.u_2)]$ $([l_1,u_1] \ominus_{in} [l_2,u_2]) = [l_1.(1–u_2), u_1.(1–l_2)]$
Positive correlation (when e_1 implies e_2, or e_2 implies e_1)	$([l_1,u_1] \otimes_{pc} [l_2,u_2]) = [\min(l_1,l_2), \min(u_1,u_2)]$ $([l_1,u_1] \oplus_{pc} [l_2,u_2]) = [\max(l_1,l_2), \max(u_1,u_2)]$ $([l_1,u_1] \ominus_{pc} [l_2,u_2]) = [\max(0, l_1–u_2), \max(0, u_1–l_2)]$
Mutual exclusion (when e_1 and e_2 are mutually exclusive)	$([l_1,u_1] \otimes_{me} [l_2,u_2]) = [0,0]$ $([l_1,u_1] \oplus_{me} [l_2,u_2]) = [\min(1, l_1+l_2), \min(1, u_1+u_2)]$ $([l_1,u_1] \ominus_{me} [l_2,u_2]) = [l_1, \min(u_1, 1–l_2)]$

Definition 2.2 Let A and B be sets, U and V be value domains, and θ be a binary relation from $\{=, \neq, \leq, \geq, <, >, \Rightarrow\}$. The *probabilistic interpretation* of the relation $A\theta B$, denoted by $Pr(A\theta B)$, is a value in [0,1] that is defined by:

1. $Pr(A\theta B) = p(u\theta v | u \in A, v \in B)$, where A is a subset of U, B is a subset of V and $\theta \in \{=, \neq, \leq, <, \geq, >\}$ assumed to be valid on $(U \times V)$, $p(u\theta v | u \in A, v \in B)$ is the conditional probability of $u\theta v$ given $u \in A$ and $v \in B$.
2. $Pr(A \Rightarrow B) = p(u \in B | u \in A)$, where A and B are two subsets of $U, p(u \in B | u \in A)$ is the conditional probability for $u \in B$ given $u \in A$.

Intuitively, given propositions "$x \in A$" and "$y \in B$," $Pr(A\theta B)$ is the probability for $x\theta y$ being true. Meanwhile, $Pr(A \Rightarrow B)$ is that, given a proposition "$x \in A$" being true, $Pr(A \Rightarrow B)$ is the probability for "$x \in B$" being true.

We note that the probabilistic interpretation of binary relations in this definition can be applied to elements of U and V because each element in a set also considered a subset of that set.

Example 2.2 Some probabilistic interpretations of the set relations on the domain consisting of natural numbers are computed as follows.

$Pr(\{3,4\} = \{4,5\}) = p(u = v | u \in \{3,4\}, v \in \{4,5\}) = 0.25$.

$Pr(\{3,4\} \Rightarrow \{4,5\}) = p(u \in \{4,5\} | u \in \{3,4\}) = 0.5$.

$Pr(3 = 5) = p(3 = 5 | 3 \in \{3\}, 5 \in \{5\}) = 0.0$.

$Pr(5 \Rightarrow \{5,6\}) = p(u \in \{5,6\} | u \in \{5\}) = 1.0$.

2.3 Probabilistic Combination Strategies

In many real situations, the probability of an event may not be defined or computed exactly [27-28]. Then, a probability interval can use instead of a precise single probability value. Let two events e_1 and e_2 have probabilities in the intervals $[l_1, u_1]$ and $[l_2, u_2]$, respectively. Then, the probability intervals of the conjunction event $e_1 \wedge e_2$, disjunction event $e_1 \vee e_2$, and difference event $e_1 \wedge \neg e_2$ can be computed by alternative strategies. In this work, we employ the conjunction, disjunction, and difference strategies given in [24] as presented in Table 1, where $\otimes$, $\oplus$, and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively.

In the following sections, the notation $[l_1, u_1] \subseteq [l_2, u_2]$ is used to denote $l_2 \leq l_1$ and $u_1 \leq u_2$. Also, a single probability value p can be treated as the probability interval $[p, p]$ and the operation $p.[l, u]$ computed as $[p.l, p.u]$.

2.4 Conjunction, Disjunction, and Difference of Probabilistic Values

For building probabilistic relational algebraic operations in IPRDB, such as the projection, join, intersection, union, and difference, we propose operations of the conjunction, disjunction, and difference of probabilistic values as the basis for combining the probability of uncertain and imprecise values of attributes in outcome relations of these algebraic operations. First, the conjunction of probabilistic values is defined as follows.

Definition 2.3 Let pv_1 and pv_2 be two probabilistic values and $\otimes$ be a probabilistic conjunction strategy. The *conjunction* of pv_1 and pv_2 under $\otimes$, denoted by $pv_1 \otimes pv_2$, is the probabilistic value pv defined by $pv = \{(v, I_1 \otimes I_2) | (v, I_1) \in pv_1, (v, I_2) \in pv_2\}$.

Example 2.3 Let $pv_1 = \{(hepatitis, [0.4, 0.6]), (cholecystitis, [0.4, 0.6])\}$ and $pv_2 = \{(hepatitis, [1.0, 1.0])\}$ be probabilistic values, then $pv_1 \otimes_{in} pv_2$ under the independence probabilistic conjunction strategy is the probabilistic value $pv = \{(hepatitis, [0.4, 0.6])\}$.

Next, the disjunction and difference of probabilistic values are defined as follows.

Definition 2.4 Let pv_1 and pv_2 be two probabilistic values and $\oplus$ be a probabilistic disjunction strategy. The *disjunction* of pv_1 and pv_2 under $\oplus$, denoted by $pv_1 \oplus pv_2$, is the probabilistic value pv defined by $pv = \{(v, I_1) | (v, I_1) \in pv_1$ and $\neg \exists I_2, (v, I_2) \in pv_2\} \cup \{(v, I_2) | (v, I_2) \in pv_2$ and $\neg \exists I_1, (v, I_1) \in pv_1\} \cup \{(v, I_1 \oplus I_2) | (v, I_1) \in pv_1$ and $(v, I_2) \in pv_2\}$.

Example 2.4 Let $pv_1 = \{(hepatitis, [0.2, 0.5]), (cholecystitis, [0.3, 0.6])\}$ and $pv_2 = \{(hepatitis, [0.3, 0.5]), (pancreatitis, [0.2, 0.6])\}$ be probabilistic val-

Table 2: *Relation PATIENT.*

P_ID	P_NAME	P_AGE	P_DISEASE	D_COST
P104	John	{(65, [1, 1])}	{(lung cancer, [0.5, 0.5]), (tuberculosis, [0.5, 0.5])}	{($30, [0.3, 0.6]), ($35, [0.4, 0.7])}
P218	Paul	{(43, [0.5, 0.5]), (44, [0.5, 0.5])}	{(hepatitis, [0.3, 0.5]), (cirrhosis, [0.5, 0.7])}	{($6, [0.4, 0.6]), ($7, [0.4, 0.6])}
P325	Helen	{(36, [1, 1])}	{(duodenitis, [0.5, 0.5]), (gastritis, [0.5, 0.5])}	{($8, [0.5, 0.5]), ($9, [0.5, 0.5])}
P412	Anne	{(15, [1, 1])}	{(bronchitis, [1, 1])}	{($7, [1, 1)}
P426	George	{(36, [1, 1])}	{(duodenitis, [0.4, 0.5]), (gastritis, [0.5, 0.6])}	{($8, [0.3, 0.5]), ($9, [0.5, 0.7])}

ues, then $pv_1 \oplus_{in} pv_2$ under the independence probabilistic disjunction strategy is the probabilistic value $pv = \{(cholecystitis, [0.3, 0.6]), (pancreatitis, [0.2, 0.6]), (hepatitis, [0.44, 0.75])\}$.

Definition 2.5 Let pv_1 and pv_2 be two probabilistic values and $\ominus$ be a probabilistic difference strategy. The *difference* of pv_1 and pv_2 under $\ominus$, denoted by $pt_1 \ominus pt_2$, is the probabilistic value pv defined by $pv = \{(v, I_1) | (v, I_1) \in pv_1 \text{ and } \neg\exists I_2, (v, I_2) \in pv_2\} \cup \{(v, I_1 \ominus I_2) | (v, I_1) \in pv_1 \text{ and } (v, I_2) \in pv_2\}$.

3. IPRDB DATA MODEL

IPRDB data model is a structure with fundamental concepts such as the schema, probabilistic relation, and database to represent data and relationships between them.

3.1 IPRDB Schemas and Relations

An IPRDB schema consists of a set of relational attributes respectively associated with domains that define the probabilistic values of those attributes. The IPRDB schema is an extension of the CRDB schema with uncertain and imprecise valued attributes. The IPRDB schema is defined as follows.

Definition 3.1 *An IPRDB schema* is a pair $R = (\boldsymbol{U}, \wp)$, where

1. $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ is a set of pairwise different attributes.
2. $\wp$ is a function that maps each attribute $A \in \boldsymbol{U}$ to the set of all probabilistic values on the domain of A.

For simplicity, we can use the notation $R(\boldsymbol{U}, \wp)$ and R to denote $R = (\boldsymbol{U}, \wp)$. The domain of A is denoted by $dom(A)$.

An IPRDB relation or a probabilistic relation of IPRDB is an instance of an IPRDB schema, where each relational attribute is associated with a probabilistic value to represent an uncertain and imprecise value that the attribute may take, as defined below.

Definition 3.2 Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes. *An IPRDB relation* r over the schema $R(\boldsymbol{U}, \wp)$ is a finite set of elements $\{t_1, t_2, \ldots, t_n\}$, where each $t_i = (pv_{i1}, pv_{i2}, \ldots, pv_{ik})$ is a list of k probabilistic values $pv_{ij} = \{(v_{ij}, [l_{ij}, u_{ij}]) | v_{ij} \in dom(A_j), [l_{ij}, u_{ij}] \subseteq [0, 1]\}, j = 1, 2, \ldots, k$ such that $pv_{ij} \in \wp(A_j)$ for every $i = 1, 2, \ldots, n$.

Each element t_i in the relation r over $R(\boldsymbol{U}, \wp)$ is called a tuple on $\boldsymbol{U}$. The probabilistic value pv_{ij} represents the uncertain and imprecise value of the attribute A_j of the tuple t_i. We write $t_i.A_j$ or $t_i[A_j]$ to denote pv_{ij} and $[t_i]$ to replace $(V_{i1}, V_{i2}, \ldots, V_{ik})$, where $V_{ij} = \{v_{ij} | (v_{ij}, [l_{ij}, u_{ij}]) \in pv_{ij}\}$. For each set of attributes $H \subseteq \{A_1, A_2, \ldots, A_k\}$, the symbol $t_i[H]$ denotes the rest of the tuple t_i after eliminating the values of attributes not belonging to H. In addition, if we only care about a unique relation over a schema then we can unify the relation's name and its schema's name.

Example 3.1 In the database about patients at the clinic of a hospital, a simple IPRDB relation, named PATIENT, over the IPRDB schema **PATIENT**({P_ID, P_NAME, P_AGE, P_DISEASE, D_COST}, $\wp$) can be given as Table 2.

In the relation, the attributes P_ID, P_NAME, P_AGE, P_DISEASE, and D_COST describe the information about the identifier, name, age, disease, and daily treatment cost of each patient, respectively. In reality, while diagnosing, the physicians can be unsure of the disease of patients. Also, the daily treatment cost for patients is not sure even the patients learn about their diseases. For instance, the information of the patient John says that John's age is 65, the patient's disease may be lung cancer or tuberculosis with a probability of 0.5, and John has to pay the daily treatment cost of $30 with a probability between 0.3 and 0.6 or $35 with the probability between 0.4 and 0.7. Note that, for each attribute A in the schema **PATIENT**, $\wp(A)$ includes all probabilistic values on the domain of A (Definition 3.1). In addition, for simplicity, each probabilistic value $\{(v, [1, 1])\}$ will be represented as a single value v (such as probabilistic values for the attribute P_ID). Because if an attribute takes such a probabilistic value, then it only takes a value v with the probability of 1.0 (Definition 2.1). In other words, the attribute certainly takes the value v.

The IPRDB relational database is an extension of CRDB with IPRDB schemas and relations as defined below.

Definition 3.3 An *IPRDB relational database* over a set of attributes is a set of IPRDB relations corresponding to the set of their IPRDB schemas.

3.2 IPRDB Functional Dependencies

Functional dependencies play an essential role in CRDB. The probabilistic functional dependent concept in IPRDB is extended from that in CRDB [3] with probabilistic valued attributes. We first define the probability measure to determine the equal degree of two probabilistic values of the same attribute for two different tuples in an IPRDB relation.

Definition 3.4 Let $R(\boldsymbol{U}, \wp)$ be an IPRDB schema, r be a relation over R and t_1 and t_2 be two tuples in r, A be an attribute of $\boldsymbol{U}$, and $\otimes$ be a probabilistic conjunction strategy. The *probability interval* for the values of the attribute A of two tuples t_1 and t_2 to be equal under $\otimes$, denoted by $p(t_1.A =_\otimes t_2.A)$, is $\oplus_{i=1}^{m}\oplus_{j=1}^{n}$ $(([l_{1i}, u_{1i}] \otimes [l_{2j}, u_{2j}]).Pr(v_{1i} = v_{2j}))$, where $t_1.A = \{(v_{11}, [l_{11}, u_{11}]), \ldots, (v_{1m}, [l_{1m}, u_{1m}])\}$, $t_2.A = \{(v_{21}, [l_{21}, u_{21}]), \ldots, (v_{2n}, [l_{2n}, u_{2n}])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.

The probabilistic functional dependency in IPRDB is an extension of the functional dependency in CRDB with uncertain and imprecise valued attributes as below.

Definition 3.5 Let $R = (\boldsymbol{U}, \wp)$ be an IPRDB schema, r be any relation over $R, \otimes$ be a probabilistic conjunction strategy, $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two non-empty subsets of $\boldsymbol{U}$. An *IPRDB functional dependency* of $\boldsymbol{Y}$ on $\boldsymbol{X}$ under $\otimes$, denoted by $\boldsymbol{X} \rightarrow_\otimes \boldsymbol{Y}$, holds if and only if

$$\forall t_1, t_2 \in r : \otimes_{A \in \boldsymbol{X}} p(t_1.A =_\otimes t_2.A) \leq \otimes_{A \in \boldsymbol{Y}} p(t_1.A =_\otimes t_2.A).$$

One can see that this definition subsumes that of CRDB. Also, it is easy to see that for every IPRDB schema $R(\boldsymbol{U}, \wp)$, then $U \rightarrow_\otimes Y$ with $Y \subseteq U$ under all probabilistic conjunction strategies.

Example 3.2 In every relation r over the schema PATIENT with the set of attributes $\boldsymbol{U}$ = {P_ID, P_NAME, P_AGE, P_DISEASE, P_COST} in Example 3.1, the values of the attribute P_ID that describe the identifiers of patients are single and pairwise different. Thus, for two tuples $t_1, t_2 \in r$ and an attribute $A \in \boldsymbol{U}$, then $p(t_1.A =_\otimes t_2.A) \geq 0$ and $p(t_1.\text{P\_ID} =_\otimes t_2.\text{P\_ID}) = 0$. So, $p(t_1.\text{P\_ID} =_\otimes t_2.\text{P\_ID}) \leq \otimes_{A \in \boldsymbol{Y}} p(t_1.A =_\otimes t_2.A)$ with $\boldsymbol{Y} \subseteq \boldsymbol{U}$, by Definition 3.5, there is the IPRDB functional dependency P_ID $\rightarrow_\otimes \boldsymbol{Y}$ in the schema **PATIENT** under all probabilistic conjunction strategies.

As in CRDB [1-3], the keys of a schema in IPRDB are the basis for recognizing a tuple of a probabilistic relation. In the model and management systems of the conventional relational database [3], key attributes cannot take the null value. Similarly, in IPRDB, we assume that the value of each key attribute is always definite and unique. The concept of the key of IPRDB schemas is defined using the probabilistic functional dependency as follows.

Definition 3.6 Let $R(\boldsymbol{U}, \wp)$ be an IPRDB schema, r be any relation over R, and $\otimes$ be a probabilistic conjunction strategy. A set of attributes $K \subseteq \boldsymbol{U}$ is a *key* of R under $\otimes$ if the value of the attributes of K is definite and there is a probabilistic functional dependency $K \rightarrow_\otimes U$ such that there does not exist any proper subset of K holding this property.

Example 3.3 In the relation PATIENT above, if we assume that each patient has a unique identifier corresponding to the value of the attribute P_ID, then P_ID is a key of the schema **PATIENT** under all probabilistic conjunction strategies.

4. IPRDB ALGEBRA

As the CRDB algebra [1-3], the IPRDB algebra is a set of basic relational algebraic operations such as the selection, projection, Cartesian product, join, intersection, union, and difference. The IPRDB algebra or the probabilistic relational algebra, including basic probabilistic relational algebraic operations, is an extension of the CRDB algebra with probabilistic values of relational attributes to manipulate, handle, and query uncertain and imprecise information on the IPRDB data model.

4.1 Selection

The selection operation of IPRDB is an extension of that of CRDB with uncertain and imprecise valued attributes. Before defining the selection operation, we introduce the formal syntax and semantics of selection expressions and conditions as below.

Definition 4.1 Let R be an IPRDB schema, and X be a set of relational tuple variables. Then, *selection expressions* are inductively defined and have one of the following forms:

1. $x.A\theta c$, where $x \in X, A$ is an attribute in R, θ is a binary relation from $\{=, \neq, \leq, \geq, <, >, \Rightarrow\}$, and $c \in dom(A)$.
2. $x.A_1\theta_\otimes x.A_2$, where $x \in X, A_1$, and A_2 are two attributes in R, and $\otimes$ is a probabilistic conjunction strategy.
3. $\alpha \otimes \beta$, where α and β are selection expressions on the same relational tuple variable, and $\otimes$ is a probabilistic conjunction strategy.
4. $\alpha \oplus \beta$, where α and β are selection expressions on the same relational tuple variable, and $\oplus$ is a probabilistic disjunction strategy.

Example 4.1 Consider the schema **PATIENT** in Example 3.1. The selection of "all patients who get cirrhosis and pay the daily treatment cost over 5 USD" can be expressed by the selection expression $x.\text{P\_DISEASE} = \text{cirrhosis} \otimes x.\text{D\_COST} > 5$.

Now, selection conditions in IPRDB are formally defined based on selection expressions as follows.

Table 3: *Relation σ_{φ}(PATIENT).*

P_ID	P_NAME	P_AGE	P_DISEASE	D_COST
P218	Paul	{(43, [0.5, 0.5]), (44, [0.5, 0.5])}	{(hepatitis, [0.3, 0.5]), (cirrhosis, [0.5, 0.7])}	{($6, [0.4, 0.6]), ($7, [0.4, 0.6])}

Definition 4.2 Let R be an IPRDB schema. Then, *selection conditions* are inductively defined as follows:

1. If α is a selection expression and $[l, u]$ is a subinterval of [0, 1], then $(\alpha)[l, u]$ is a selection condition.
2. If φ and ω are selection conditions on the same tuple variable, then $\neg\varphi, (\varphi \wedge \omega), (\varphi \vee \omega)$ are selection conditions.

Example 4.2 Given the schema **PATIENT** in Example 3.1, the selection of "all patients who are over 40 years old with a probability of at least 0.8 or have tuberculosis and pay the daily treatment cost not less than 30 USD with a probability between 0.5 and 0.6" can be done using the selection condition $(x.\text{P\_AGE} > 40)[0.8, 1] \vee (x.\text{P\_DISEASE} = \text{tuberculosis} \otimes \text{x.D\_COST} \geq 30)[0.5, 0.6]$.

Definition 4.3 Let R be an IPRDB schema, r be a relation over R, x be a tuple variable, and t be a tuple in r. The *probabilistic interpretation of selection expressions* for R, r, and t, denoted by $Prob_{R,r,t}$, is the partial mapping from the set of all selection expressions to the set of all closed subintervals of [0, 1] that is inductively defined as follows:

1. $Prob_{R,r,t}(x.A\theta c) = \oplus_{i=1}^{k}([l_i, u_i].Pr(v_i\theta c))$, where $t.A = \{(v_1, [l_1, u_1]), \ldots, (v_k, [l_k, u_k])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.
2. $Prob_{R,r,t}(x.A_1\theta_{\otimes}x.A_2) = \oplus_{i=1}^{m}\oplus_{j=1}^{n}(([l_{1i}, u_{1i}]\otimes [l_{2j}, u_{2j}]).Pr(v_{1i}\theta v_{2j}))$, where $t.A_1$=$\{(v_{11}, [l_{11}, u_{11}]), \ldots, (v_{1m}, [l_{1m}, u_{1m}])\}$, $t.A_2 = \{(v_{21}, [l_{21}, u_{21}]), \ldots, (v_{2n}, [l_{2n}, u_{2n}])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.
3. $Prob_{R,r,t}(\alpha \otimes \beta) = Prob_{R,r,t}(\alpha) \otimes Prob_{R,r,t}(\beta)$.
4. $Prob_{R,r,t}(\alpha \oplus \beta) = Prob_{R,r,t}(\alpha) \oplus Prob_{R,r,t}(\beta)$.

We note that the mutual exclusion probabilistic disjunction operator $\oplus_{me}$ is used in items 1 and 2 of Definition 4.3 because the intervals $[l_1, u_1], \ldots, [l_k, u_k]$ represent a probability distribution function over $\{v_1, \ldots, v_k\}$, likewise for $[l_{11}, u_{11}], \ldots, [l_{1m}, u_{1m}]$ and $[l_{21}, u_{21}], \ldots, [l_{2n}, u_{2n}]$. Intuitively, $Prob_{R,r,t}(x.A\theta c)$ is the probability interval for the attribute A of the tuple t having a value v_i such that $v_i\ \theta\ c$, while $Prob_{R,r,t}(x.A_1\ \theta_{\otimes}\ x.A_2)$ is the probability interval for the attributes A_1 and A_2 of the tuple t having values v_{1i} and v_{2j}, respectively, such that $v_{1i}\theta v_{2j}$.

Example 4.3 Let R denote the schema **PATIENT** and r denote the relation PATIENT in Example 3.1. Consider the second tuple in r, denoted by t_2. We have

$Prob_{R,r,t_2}(x.\text{P\_DISEASE} = \text{cirrhosis})$
$= [0.3, 0.5].Pr(\text{hepatitis} = \text{cirrhosis})$
$\oplus_{me}\ [0.5, 0.7].Pr(\text{cirrhosis} = \text{cirrhosis})$
$= [0.3, 0.5] \times 0.0 \oplus_{me} [0.5, 0.7]\times 1.0$
$= [0, 0] \oplus_{me} [0.5, 0.7] = [0.5, 0.7]$

The satisfaction of a selection condition in IPRDB is an extension of that in CRDB with probability intervals as below.

Definition 4.4 Let R be an IPRDB schema, r be a relation over R and $t \in r$. The *satisfaction of selection conditions* under $Prob_{R,r,t}$ is defined as follows:

1. $Prob_{R,r,t} \models (\alpha)[l, u]$ if and only if (iff) $Prob_{R,r,t}(\alpha) \subseteq [l, u]$.
2. $Prob_{R,r,t} \models \neg\varphi$ iff $Prob_{R,r,t} \models \varphi$ does not hold.
3. $Prob_{R,r,t} \models \varphi \wedge \omega$ iff $Prob_{R,r,t} \models \varphi$ and $Prob_{R,r,t} \models \omega$.
4. $Prob_{R,r,t} \models \varphi \vee \omega$ iff $Prob_{R,r,t} \models \varphi$ or $Prob_{R,r,t} \models \omega$.

The selection on a relation in IPRDB is defined as follows.

Definition 4.5 Let R be an IPRDB schema, r be a relation over R, and φ be a selection condition over a tuple variable in r. The selection on r for φ, denoted by $\sigma_{\varphi}(r)$, is a relation r^* over R specified by

$$r^* = \{t \in r | Prob_{R,r,t} \models \varphi\}$$

Example 4.4 Let r denote the relation PATIENT in Example 3.1, and R denote its schema. The query "Find all patients who are over 40 years old with a probability of at least 0.9, and have cirrhosis and pay the daily treatment cost not less than 6 USD with a probability between 0.3 and 0.7" can be done by the selection operation σ_{φ}(PATIENT), where $\varphi = (x.\text{P\_AGE} > 40)[0.9, 1] \wedge (x.\text{P\_DISEASE} = \text{cirrhosis} \otimes_{in} x.\text{D\_COST} \geq 6)[0.3, 0.7]$.

Only the second tuple t_2 of the relation PATIENT in Example 3.1 satisfies φ because

$Prob_{R,r,t_2}(x.\text{P\_AGE} > 40)$
$= [0.5, 0.5]\times Pr(43>40)\oplus_{me}[0.5, 0.5]\times Pr(44>40)$
$= [0.5, 0.5]\times 1.0 \oplus_{me} [0.5, 0.5]\times 1.0$
$= [1.0, 1.0] \subseteq [0.9, 1]$.
$Prob_{R,r,t_2}(x.\text{D\_COST} \geq 6)$
$= [0.4, 0.6]\times Pr(6 \geq 6)\oplus_{me} [0.4, 0.6]\times Pr(7 \geq 6)$
$= [0.4, 0.6]\times 1.0 \oplus_{me} [0.4, 0.6]\times 1.0$
$= [0.4, 0.6] \oplus_{me} [0.4, 0.6]$
$= [0.8, 1]$.

From the result of the computation in Example 4.3, we have

$Prob_{R,r,t_2}(x.\text{P\_DISEASE=cirrhosis} \otimes_{in} x.\text{D\_COST} \geq 6)$
$= [0.5, 0.7] \otimes_{in} [0.8, 1] = [0.4, 0.7] \subseteq [0.3, 0.7]$.

For the other tuples, one has $Prob_{R,r,t_i}(x.\text{P\_DISEASE} = \text{cirrhosis} \otimes_{in} x.\text{D\_COST} \geq 6) = [0, 0] \not\subseteq [0.3, 0.7]$, $\forall_i \neq 2$. Thus, the result of the query is shown in Table 3.

Table 4: *Relation* $\prod_{\{P\_AGE,P\_DISEASE,D\_COST\}\oplus_{in}}$ *(PATIENT).*

P_AGE	P_DISEASE	D_COST
{(65, [1, 1])}	{(lung cancer, [0.5, 0.5]), (tuberculosis, [0.5, 0.5])}	{($30, [0.3, 0.6]), ($35, [0.4, 0.7])}
{(43, [0.5, 0.5]), (44, [0.5, 0.5])}	{(hepatitis, [0.3, 0.5]), (cirrhosis, [0.5, 0.7])}	{($6, [0.4, 0.6]), ($7, [0.4, 0.6])}
{(15, [1, 1])}	{(bronchitis, [1, 1])}	{($7, [1, 1)}
{(36, [1, 1])}	{(duodenitis, [0.7, 0.75]), (gastritis, [0.75, 0.8])}	{($8, [0.65, 0.75]), ($9, [0.75, 0.85])}

4.2 Projection

The projection of an IPRDB relation on a set of attributes is an extension of that of a CRDB relation with interval probabilities such that the projected tuples having the same value should be merged into a tuple in the result relation by a probabilistic disjunction strategy. The projection operation of an IPRDB relation is defined as follows.

Definition 4.6 Let $R(\boldsymbol{U}, \wp)$ be an IPRDB schema, r be a relation over R, $\boldsymbol{H}$ be a subset of attributes of $\boldsymbol{U}$, $\oplus$ be a probabilistic disjunction strategy. The *projection* of r on $\boldsymbol{H}$ under $\oplus$, denoted by $\prod_{\boldsymbol{H}\oplus}(r)$, is the relation r^* over the schema R^* determined by:

1. $R^* = (\boldsymbol{H}, \wp^*)$ and $\wp^*(A) = \wp(A), \forall A \in \boldsymbol{H}$.
2. $r^* = \{t^* | t^*.A = u.A \oplus \ldots \oplus w.A, \forall A \in \boldsymbol{H}, \exists u, \ldots, w \in r$ such that $[u[\boldsymbol{H}]] = \ldots = [w[\boldsymbol{H}]]\}$.

Example 4.5 Consider the relation PATIENT over the schema **PATIENT**({P_ID, P_NAME, P_AGE, P_DISEASE, D_COST}, $\wp$) as in Table 2, then the projection of it on the set of the attributes $\boldsymbol{H}$ = {P_AGE, P_DISEASE, D_COST} under $\oplus_{in}$ is the relation $\prod_{H\oplus_{in}}$(PATIENT) over the schema R^*({P_AGE, P_DISEASE, D_COST}, $\wp^*$) computed as in Table 4, where $\wp^*(A) = \wp(A), \forall A \in \boldsymbol{H}$.

Note that in the relation PATIENT, we have $[t_3[\boldsymbol{H}]] = [t_5[\boldsymbol{H}]]$, thus two tuples, t_3 and t_5, are projected on $\boldsymbol{H}$ and merged into the tuple t_4 under the independence probabilistic disjunction strategy $\oplus_{in}$ in Table 4.

4.3 Cartesian Product

For the Cartesian product of two IPRDB relations, as in CRDB, we assume the set of attributes of their schemas are disjoint, and every k-tuple $t = (pv_1, pv_2, \ldots, pv_k)$ of probabilistic values is an unordered list. The Cartesian product of two IPRDB relations is extended from that of two CRDB relations with uncertain and imprecise valued attributes as follows.

Definition 4.7 Let $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ be two sets of attributes that do not have any common element, $R_1(\boldsymbol{U}_1, \wp_1)$, $R_2(\boldsymbol{U}_2, \wp_2)$ be two IPRDB schemas, r_1, r_2 be two relations over R_1 and R_2, respectively. The *Cartesian product* of r_1 and r_2, denoted by $r_1 \times r_2$, is the relation r over R, determined by:

1. $R = (\boldsymbol{U}, \wp)$, where $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2$, $\wp(A) = \wp_1(A)$ if $A \in \boldsymbol{U}_1$ and $\wp(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_2$.
2. $r = \{t | t.A = t_1.A$ if $A \in \boldsymbol{U}_1, t.A = t_2.A$ if $A \in \boldsymbol{U}_2, t_1 \in r_1, t_2 \in r_2\}$.

4.4 Join

The join of two IPRDB relations is an extension of the natural join of two CRDB relations with probabilistic values as the following definition.

Definition 4.8 Let $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ be two sets of attributes such that if they have the same name attributes, respectively, in those two sets, then such attributes have the same value domain. Let $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ be two IPRDB schemas, r_1 and r_2 be two relations over R_1 and R_2, respectively, and $\otimes$ be a probabilistic conjunction strategy. The *join* of r_1 and r_2 under $\otimes$, denoted by $r_1 \bowtie_{\otimes} r_2$, is the relation r over the schema R, determined by:

Table 5: *Relation* $PATIENT_1$.

P_ID	P_DISEASE
P421	{(bronchitis, [0.35, 0.45]), (bronchiectasis, [0.55, 0.65)}
P829	{(pancreatitis, [1, 1])}

Table 6: *Relation* $PATIENT_2$.

P_NAME	P_DISEASE
Peter	{(bronchiectasis, [1, 1])}
Selena	{(pancreatitis, [0.4, 0.5]), (cirrhosis, [0.5, 0.6])}

Table 7: *Relation* $PATIENT_1 \bowtie_{\otimes_{in}} PATIENT_2$.

P_ID	P_NAME	P_DISEASE
P421	Peter	{(bronchiectasis, [0.55, 0.65)}
P829	**Selena**	{(pancreatitis, [0.4, 0.5])}

1. $R = (\boldsymbol{U}, \wp)$ where $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2, \wp(A) = \wp_1(A)$ if $A \in \boldsymbol{U}_1 - \boldsymbol{U}_2, \wp(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_2 - \boldsymbol{U}_1$ and $\wp(A) = \wp_1(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$.
2. $r = \{t | t.A = t_1.A$ if $A \in \boldsymbol{U}_1 - \boldsymbol{U}_2$, $t.A = t_2.A$ if $A \in \boldsymbol{U}_2 - \boldsymbol{U}_1, t.A = t_1.A \otimes t_2.A$ if $A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$, $t_1 \in r_1, t_2 \in r_2\}$.

Example 4.6 Given two IPRDB relations PATIENT_1 and PATIENT_2 as in Tables 5 and 6, then the result of the join of them under the probabilistic conjunction strategy $\otimes_{in}$ is the relation $\text{PATIENT}_1 \bowtie_{\otimes in} \text{PATIENT}_2$ computed as in Table 7. Here, the names of each relation and its schema are identical. The set $\wp(A)$ for each attribute A in the schemas consists of probabilistic values on $dom(A)$.

4.5 Intersection, Union and Difference

The intersection, union, and difference of two IPRDB relations over the same schema is an IPRDB relation over that schema, where two tuples that have the same key, respectively of those two relations, should be merged into a tuple in the result relation by a probabilistic combination strategy. Here, two tuples have the same key value like two identical tuples in the conventional relational database. Thus, the operations are the extensions of the intersection, union, and difference of two CRDB relations with probabilistic valued attributes. The intersection, union, and difference of two IPRDB relations are defined as follows.

Definition 4.9 Let $R(\boldsymbol{U}, \wp)$ be an IPRDB schema, r_1, and r_2 be two relations over R, K be a key of R, and $\otimes$ be a probabilistic conjunction strategy. The *intersection* of r_1 and r_2 under $\otimes$, denoted by $r_1 \cap_{\otimes} r_2$, is the IPRDB relation r over R defined by $r = \{t | t.A = t_1.A \otimes t_2.A,\ t_1 \in r_1,\ t_2 \in r_2,\ A \in \boldsymbol{U}$, such that $t_1[K] = t_2[K]\}$.

We note that the value of each key attribute is definite under the definition 3.6. Thus, the notation $t_1[K] = t_2[K]$ can be used in the definition 4.9. Moreover, we can uniquely determine a tuple of a relation under every key of the relation. So, the result relation is unique under all the keys.

Definition 4.10 Let $R(U, \wp)$ be an IPRDB schema, r_1 and r_2 be two relations over R, K be a key of R, $\oplus$ be a probabilistic disjunction strategy. The *union* of r_1 and r_2 under $\oplus$, denoted by $r_1 \cup_{\oplus} r_2$, is the IPRDB relation r over R defined by $r = \{t_1 \in r_1 | \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t_2 \in r_2 | \forall t_1 \in r_1, t_2[K] \neq t_1[K]\} \cup \{t | t.A = t_1.A \oplus t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1[K] = t_2[K]\}$.

Example 4.7 Given two IPRDB relations DIAGNOSE_1 and DIAGNOSE_2 over the same schema **DIAGNOSE**({P_ID, D_ID, P_DISEASE, D_COST}, $\wp$) as in Tables 8 and 9, where {P_ID, D_ID} is the key of this schema and the set $\wp(A)$ for each attribute A in **DIAGNOSE** consists of all probabilistic values on $dom(A)$. Then, the union of DIAGNOSE_1 and DIAGNOSE_2 under $\oplus_{in}$ is the relation $\text{DIAGNOSE}_1 \cup_{\oplus_{in}} \text{DIAGNOSE}_2$ computed as in Table 10.

We note that the tuple t_2 in Table 8 and the tuple t_2 in Table 9 have the same key value coalesced into the tuple t_4 under $\oplus_{in}$ in Table 10.

Definition 4.11 Let $R(\boldsymbol{U}, \wp)$ be an IPRDB schema, r_1 and r_2 be two relations over R, K be a key of R, and $\ominus$ be a probabilistic difference strategy. The *difference* of r_1 and r_2 under $\ominus$, denoted by $r_1 \cup_{\ominus} r_2$, is the IPRDB relation r over R defined by $r = \{t_1 \in r_1 | \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t | t.A = t_1.A \ominus$

Table 8: *Relation $DIAGNOSE_1$.*

P_ID	D_ID	P_DISEASE	D_COST
P226	D014	{(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])}	{($30, [0.3, 0.4]), ($35, [0.6, 0.7])}
P255	D020	{(hepatitis, [0.3, 0.8]), (pancreatitis, [0.2, 0.7])}	{($8, [0.6, 1])}

Table 9: *Relation $DIAGNOSE_2$.*

P_ID	D_ID	P_DISEASE	D_COST
P228	D016	{(lung cancer, [1, 1])}	{($30, [1, 1])}
P255	D020	{(hepatitis, [0.4, 0.8]), (cholecystitis, [0.2, 0.6])}	{($7, [0.2, 0.4]), ($8, [0.4, 0.8])}
P262	D022	{(dyspepsia, [1, 1])}	{($5, [1, 1])}

Table 10: *Relation $DIAGNOSE_1 \cup_{\oplus_{in}} DIAGNOSE_2$.*

P_ID	D_ID	P_DISEASE	D_COST
P226	D014	{(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])}	{($30, [0.3, 0.4]), ($35, [0.6, 0.7])}
P228	D016	{(lung cancer, [1, 1])}	{($30, [1, 1])}
P262	D022	{(dyspepsia, [1, 1])}	{($5, [1, 1])}
P255	D020	{(hepatitis, [0.58, 0.96]), (pancreatitis, [0.2, 0.7]), (cholecystitis, [0.2, 0.6])}	{($7, [0.2, 0.4]), ($8, [0.76, 1.0])}

$t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1[K] = t_2[K]\}$.

We note that the result relation in the definitions 4.10 and 4.11 does not depend on choosing the key of its schema.

4.6 Property of Algebraic Operations

The basic properties of the algebraic operations in IPRDB are the extensions of those in CRDB with probabilistic values. These properties say that the IPRDB model is sound and coherent.

Proposition 4.1 Let r be a relation over the schema R in IPRDB, and φ and ω be two selection conditions. Then,

$$\sigma_{\varphi}(\sigma_{\omega}(r)) = \sigma_{\omega}(\sigma_{\varphi}(r)) \tag{1}$$

Proof: Let $s = \sigma_{\omega}(r)$. By Definition 4.4 and 4.5, we have $\sigma_{\varphi}(\sigma_{\omega}(r)) = \{t \in s | Prob_{R,s,t} \models \varphi\}$
$= \{t \in r | (Prob_{R,r,t} \models \omega) \wedge (Prob_{R,s,t} \models \varphi)\}$
$= \{t \in r | (Prob_{R,r,t} \models \omega) \wedge (Prob_{R,r,t} \models \varphi)\}$
$= \{t \in r | Prob_{R,r,t} \models \varphi \wedge \omega\} = \sigma_{\varphi \wedge \omega}(r)$.

Thus, the equation $\sigma_{\varphi}(\sigma_{\omega}(r)) = \sigma_{\omega \wedge \omega}(r)$ is proven. The equation $\sigma_{\omega}(\sigma_{\varphi}(r)) = \sigma_{\omega \wedge \omega}(r)$ is similarly proven, since $\omega \wedge \varphi \Leftrightarrow \varphi \wedge \omega$. So, Proposition 4.1 is proven.

Proposition 4.2 Let R be an IPRDB schema, r be a relation over R, $\oplus$ be a probabilistic disjunction strategy, $\boldsymbol{A}$ and $\boldsymbol{B}$ be two subsets of attributes of $R, \boldsymbol{A} \subseteq \boldsymbol{B}$. Then,

$$\Pi_{\boldsymbol{A}\oplus}(\Pi_{\boldsymbol{B}\oplus}(r)) = \Pi_{\boldsymbol{A}\oplus}(r) \tag{2}$$

Proof: Because $\boldsymbol{A} \subseteq \boldsymbol{B}$, so $\boldsymbol{A} \cap \boldsymbol{B} = \boldsymbol{A}$ and sides of (2) are the relations over the same schema. From Definition 4.6, it is easy to see $\prod_{\boldsymbol{A}\oplus}(\prod_{\boldsymbol{B}\oplus}(r)) = \prod_{\boldsymbol{A}\cap\boldsymbol{B}\oplus}(r) = \prod_{\boldsymbol{A}\oplus}(r)$ under the probabilistic disjunction strategy $\oplus$. Thus, the equation (2) is proven.

Proposition 4.3 Let R_1, R_2, and R_3 be the IPRDB schemas such that if they have the same name attributes, then such attributes have the same value domain, r_1, r_2, and r_3 be relations over R_1, R_2, and R_3, respectively, $\otimes$ be a probabilistic conjunction strategy. Then,

$$r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1 \tag{3}$$
$$(r_1 \bowtie_{\otimes} r_2) \bowtie_{\otimes} r_3 = r_1 \bowtie_{\otimes} (r_2 \bowtie_{\otimes} r_3) \tag{4}$$

The equations (3) and (4) say that the join of IPRDB relations is commutative and associative.

Proof: It is easy to see that $r_1 \bowtie_{\otimes} r_2$ and $r_2 \bowtie_{\otimes} r_1$ are two relations over the same schema. By Definition 2.3, the conjunction of probabilistic values is commutative (due to the commutativity of probabilistic conjunction strategies). So, by Definition 4.8, it follows that $r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1$.

By Definition 4.8, the results of two sides of (4) are the relations over the same schema. Moreover, by Definition 2.3, the conjunction of probabilistic values is associative. By Definition 4.8 and from the associativity of the conventional relational natural join, it follows that the join of IPRDB relations is associative. Thus, it results in $(r_1 \bowtie_{\otimes} r_2) \bowtie_{\otimes} r_3 = r_1 \bowtie_{\otimes} (r_2 \bowtie_{\otimes} r_3)$.

Because the Cartesian product (Definition 4.7) is a particular case of the join, it yields the straight result of Proposition 4.3 below.

Corollary 4.1 Let R_1, R_2, and R_3 be IPRDB schemas such that they do not have the same name attributes, r_1, r_2, and r_3 be relations over R_1, R_2, and R_3, respectively. Then,

$$r_1 \times r_2 = r_2 \times r_1 \tag{5}$$
$$(r_1 \times r_2) \times r_3 = r_1 \times (r_2 \times r_3) \tag{6}$$

Proposition 4.4 Let R be an IPRDB schema, r_1, r_2, and r_3 be relations over R. Let $\otimes/\oplus$be a probabilistic conjunction/disjunction strategy. Then,

$$r_1 \cap_{\otimes} r_2 = r_2 \cap_{\otimes} r_1 \tag{7}$$
$$(r_1 \cap_{\otimes} r_2) \cap_{\otimes} r_3 = r_1 \cap_{\otimes} (r_2 \cap_{\otimes} r_3) \tag{8}$$
$$r_1 \cup_{\oplus} r_2 = r_2 \cup_{\oplus} r_1 \tag{9}$$
$$(r_1 \cup_{\oplus} r_2) \cup_{\oplus} r_3 = r_1 \cup_{\oplus} (r_2 \cup_{\oplus} r_3) \tag{10}$$

Equations of (7), (8), (9), and (10) say that the intersection and union of relations in IPRDB are commutative and associative.

Proof: From the commutativity and associativity of the probabilistic conjunction strategies, it follows that the conjunction of probabilistic values has the commutativity and associativity (Definition 2.3). So, the intersection of IPRDB relations r_1, r_2, and r_3 under the probabilistic conjunction strategy $\otimes$ and every chosen key also has commutativity and associativity. From that, by Definition 4.9, we have Equations (7) and (8).

From the commutativity and associativity of the probabilistic disjunction strategies, it follows that the disjunction of probabilistic values has the commutativity and associativity (Definition 2.4). So, the union of IPRDB relations r_1, r_2, and r_3 under the probabilistic disjunction strategy $\oplus$ and every chosen key also has commutativity and associativity. From that, by Definition 4.10, we have Equations (9) and (10).

5. RESULT AND DISCUSSION

We can see that IPRDB is an extension of CRDB and the second class PRDB models as in [18], [19], and [20] with probabilistic values (i.e., probabilistic intervals for values). In addition, IPRDB also has the capability of manipulating data more effectively than the second-class PRDB models as in [21], [22], and [23]. A more detailed discussion of the obtained results is as below.

5.1 Extension of IPRDB in representing data

There are two main classes of the PRDB model. The first class, denoted by C-1PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in [0, 1], such as [9] and [10]. Each attribute of a tuple can take a single value with an inferred probability from the membership degree of that tuple. The C-1PRDB algebraic operations are defined by directly extending the CRDB algebraic operations based on computing and combining probabilities of tuples in the C-1PRDB relations.

The second class, denoted by C-2PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in {0, 1}, such as [18] and [19], each relational tuple attribute is associated with a single probability value as (v, p) to say that the attribute may take the value v with the probability p. Some extended models of C-2PRDB such as [20], denoted by EC-2PRDB, where each relational tuple attribute is associated with a probability distribution as $\{(v_1, p_1), \ldots, (v_m, p_m)\}$ to say that the attribute may take one of values v_i with the probability p_i. The C-2PRDB and EC-2PRDB algebraic operations are defined by extending the CRDB algebraic operations and employing operators on single probabilities or probability distributions for computing and combining probabilities of attribute values in the C-2PRDB or EC-2PRDB relations.

As introduced in previous sections, our IPRDB model belongs to C-2PRDB. Each relational tuple attribute in IPRDB is associated with a probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ (as a distribution of probability intervals on a finite set of values) to say that the attribute may take one of values v_i with a probability in $[l_i, u_i]$. The IPRDB algebraic operations are defined by extending the CRDB algebraic operations using the probabilistic interpretations of binary relations on sets and the combination strategies of probabilistic intervals of attribute values (i.e., probabilistic values) in the C-2PRDB relations.

It is easy to see that a particular probabilistic value in IPRDB as $\{(v_1, [p_1, p_1]), \ldots, (v_m, [p_m, p_m])\}$ also is a probability distribution $\{(v_1, p_1), \ldots, (v_m, p_m)\}$ in the model [20]. Thus, the IPRDB model is an extension of C-2PRDB models, such as [19] and [20], with probabilistic values (Definition 2.1 and 3.2). Moreover, by associating probabilistic intervals with attribute values (in probabilistic values), IPRDB allows representing both the uncertainty of attribute values and the imprecision of the probability for that attribute values. In contrast, the models as [19] and [20] only allow representing the uncertainty of attribute values but do not allow expressing the imprecision of the probability for that attribute values (because in $\{(v_1, p_1), \ldots, (v_m, p_m)\}$, the probability for the value v_i is a precise number p_i). Figure 1 illustrates the extension of IPRDB in comparison with the CRDB, C-2PRDB, and EC-2PRDB models.

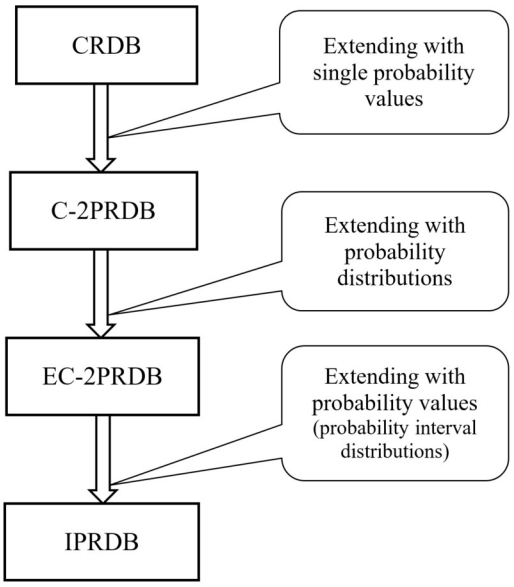

Fig.1: *Extension of IPRDB.*

5.2 Efficiency of IPRDB in manipulating data

Because the attribute value of IPRDB relations is a probabilistic value, the computation and manipulation of the IPRDB data model are more effective than those of the C-2PRDB data models of [21], [22], and [23], where the attribute value is the probability distribution function pairs of a set of values. The computing complexity of IPRDB algebraic operations is a polynomial under the size of probabilistic relations, and it is as effective as the computing complexity of CRDB and EC-2PRDB algebraic operations. Indeed, regarding the selection operation, since the computation time that a tuple holds or does not hold a selection condition is bounded above by some constant (Definition 4.3 and 4.4), then the cost for the selection of each tuple in an IPRDB relation (Definition 4.5) also is some constant or O(1). Thus, the computing time complexity of the selection operation on an IPRDB relation with n tuples is O(n). With the projection, from Definition 4.6, it is easy to see that the time for the probabilistic combination of the duplicate value tuples under a probabilistic disjunction strategy is a constant. Hence, the computing complexity of the projection on an IPRDB relation having n tuples is O(n). Similarly, the computing time complexity of Cartesian product, join, intersection, union, and difference operations on two IPRDB relations having n and m tuples is O(nm). Thus, the performance of the IPRDB model in computing and manipulating uncertain and imprecise information is good and can be applied in practice.

6. CONCLUSION

In this paper, we have proposed a new probabilistic relational database model, named IPRDB, that extends the CRDB model with interval probability valued attributes for uncertain and imprecise information. In IPRDB, each relation is defined as a set of tuples whose attributes are associated with interval probability values to represent uncertainty and imprecision of the value that these attributes may take. The fundamental concepts of the relational schema, probabilistic functional dependency, key as well as the set of basic probabilistic relational algebraic operations in IPRDB have been extended consistently with those in CRDB using the probabilistic interpretation of binary relations on sets, probabilistic combination strategies, and conjunction, disjunction, difference operations of probabilistic values. Basic properties of the probabilistic relational algebraic operations are proposed and proven ultimately to say that IPRDB is a sound and coherent model. The built IPRDB model can manipulate and deal with effectively uncertain and imprecise data.

For a complete database system of IPRDB, we are investigating the development of an IPRDB management system and its query language to apply the IPRDB model in practice.

ACKNOWLEDGEMENT

We acknowledge the Industrial University of Ho Chi Minh City for supporting this work.

AUTHOR CONTRIBUTIONS

Conceptualization, H. Nguyen and D.N. Le; methodology, H. Nguyen; validation, H. Nguyen and D.N. Le; formal analysis, H. Nguyen; investigation, H. Nguyen; data curation, H. Nguyen and D.N. Le; writing—original draft preparation, H. Nguyen; writing—review and editing, H. Nguyen and D.N. Le; visualization, H. Nguyen and D.N. Le; supervision, H. Nguyen; funding acquisition, H. Nguyen. All authors have read and agreed to the published version of the manuscript.

References

[1] E.F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol.13, no.6, pp.377-387, 1970.

[2] G. Özsoyoğlu, Z. M. Özsoyoğlu, and V. Matos, "Extending relational algebra and relational calculus with set-valued attributes and aggregate functions," *ACM Transactions on Database Systems*, vol.12, no.4, pp.566-592, 1987.

[3] A. Silberschatz, H.F. Korth, and S. Sudarshan, *Database system concepts*, Seventh Edition, McGraw-Hill, 2019.

[4] D. Dey and S. Sarkar, "A probabilistic relational model and algebra," *ACM Transactions on Database Systems*, vol.21, no.3, pp.339-369, 1996.

[5] D. Barbara, H. Garcia-Molina, and D. Porter, "The management of probabilistic data," *IEEE Transactions on Knowledge and Data Engineering*, vol.4, no.5, pp.487-502, 1992.

[6] J. Bernad, C. Bobed, and E. Mena, "Uncertain probabilistic range queries on multidimensional data," *Information Sciences*, vol. 537, pp.334-367, 2020.

[7] A. Ali, S. Talpur, and S. Narejo, "Detecting faulty sensors by analyzing the uncertain data using probabilistic databases," *Proceedings of 3rd International Conference on Computing, Mathematics and Engineering Technologies*, Sukkur, Pakistan, pp.143-150, 2020.

[8] V.V. Kheradkar and S. K. Shirgave, "Query processing over relational cross model in uncertain and probabilistic databases," *Proceedings of 3Th International Conference on Artificial Intelligence and Smart Energy*, Coimbatore, India, pp.763-769, 2023.

[9] N. Fuhr and T. Rolleke, "A probabilistic relational algebra for the integration of information retrieval and database systems," *ACM Transactions on Information Systems*, vol.15, no.1, pp.32-66, 1997.

[10] S. Zhang and C. Zhang, "A probabilistic data model and its semantics," *Journal of Research and Practice in Information Technology*, vol.35, no.4, pp.237-256, 2003.

[11] Z. Ma and L. Yan, *Advances in probabilistic databases for uncertain information management*, Springer-Verlag Berlin Heidelberg, 2013.

[12] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: A survey of theories and practices," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no.11, pp.2463-2482, 2013.

[13] I.I. Ceylan, A. Darwiche, and G.V.D. Broeck, "Open-world probabilistic databases: Semantics, algorithms, complexity," *Journal of Artificial Intelligence*, vol.295, no.11, pp.103474-103513, 2021.

[14] H. Debbi, "Explaining query answers in probabilistic databases," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.8, no.4, pp.140-152, 2023.

[15] L.V.S. Lakshmanan, N. Leone, R. Ross, and V.S. Subrahmanian, "Probview: A flexible probabilistic database system," *ACM Transactions on Database Systems*, vol.22, no.3, pp.419-469, 1997.

[16] W. Zhao, A. Dekhtyar, and J. Goldsmith, "Databases for interval probabilities," *International Journal of Intelligent Systems*, vol.19, no.9, pp.789-815, 2004.

[17] R. Ross and V.S. Subrahmanian, "Aggregate op-

erators in probabilistic databases," *Journal of the ACM*, vol.52, no.1, pp.54-101, 2005.

[18] D. Dey and S. Sarkar, "Generalized normal forms for probabilistic relational data," *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.3, pp.485-497, 1992.

[19] T. Eiter, T. Lukasiewicz, and M. Walter, "A data model and algebra for probabilistic complex values," *Annals of Mathematics and Artificial Intelligence*, vol.33, pp.205-252, 2001.

[20] S.K. Lee, "An extended relational database model for uncertain and imprecise information," *Proceedings of 18th Conference on Very Large Data Bases*, Vancouver, Canada, pp.211-220, 1992.

[21] H. Nguyen, "A probabilistic relational database model and algebra," *Journal of Computer Science and Cybernetics*, vol.31, no.4, pp.305-321, 2015.

[22] H. Nguyen, T.N. Nguyen, and T.T.N. Tran, "A probabilistic relational database model with uncertain multivalued attributes," *ICIC Express Letters*, vol. 16, no.3, pp.241-248, 2022.

[23] H. Nguyen, "Extending probabilistic relational database model with uncertain multivalued attributes," *International Journal of Innovative Computing, Information and Control*, vol.18, no.5, pp.1477–1492, 2022.

[24] V. Biazzo, R. Giugno, T. Lukasiewicz, and V. S. Subrahmanian, "Temporal probabilistic object bases," *IEEE Transactions on Knowledge and Data Engineering*, vol.15, no.4, pp. 921-939, 2003.

[25] H. Nguyen, "Extending relational database model for uncertain information," *Journal of Computer Science and Cybernetics*, vol.35, no.4, pp.355-372, 2019.

[26] T. Friedman and G. Broeck, "Symbolic querying of vector spaces: probabilistic databases meet relational embeddings," *Proceedings of 36th Conference on Uncertainty in Artificial Intelligence*, Toronto, Canada, vol.124, pp.1268-1277, 2020.

[27] A. Gilad, A. Imber, and B. Kimelfeld, "The consistency of probabilistic databases with independent cells," *Proceedings of 26th International Conference on Database Theory*, Ioannina, Greece, pp. 22:1-22:19, 2023.

[28] T.V. Bremen and K.S. Meel, "Probabilistic query evaluation: The combined FPRAS landscape," *Proceedings of 42th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, New York, USA, pp 339–347, 2023.

Hoa Nguyen received his Ph.D degree in Computer Science at Vietnam National University, Ho Chi Minh City, Vietnam, in 2008. Dr. Nguyen is currently an associate professor at Information Technology Faculty, Saigon University, Vietnam, he is also a visiting professor on databases at Industrial University of Ho Chi Minh City, Vietnam. His research interests include imprecise and uncertain knowledge representation, fuzzy databases and probabilistic dadabases.

Duy Nhat Le received his PhD degree in Computer Science from Moscow State Pedagogical University, Russia in 2013. He is currently a lecturer at Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam since 2013. His research areas of interest include computational intelligence and cryptography.