International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 03, No. 01, page 11 - 24
Submitted 8/1/2025; Accepted 14/2/2025; Published 20/2/2025

An Extended Relational Database Model for Interval Probability Set-Valued Attributes

[1] Hoa Nguyen and [2]Thi Nhi Tran
[1]Information Technology Faculty, Saigon University, VIETNAM
[2] Information Technology Faculty, Saigon University, VIETNAM
e-mail : [1]nguyenhoa@sgu.edu.vn, [2]ttnhi@hv.sgu.edu.vn

Corresponding Autor: Hoa Nguyen

Abstract

In this paper, we introduce a new probabilistic relational database model as an extension of the classical relational database model for interval probability set-valued attributes to represent and handle uncertain and imprecise information in practice. To develop the new model, we use extended probabilistic values for representing interval probability set-valued relational attributes and the probabilistic interpretation of binary relations on sets for computing uncertain degree of functional dependencies, keys and relations on attribute values, and propose the new combination strategies of extended probabilistic values for building probabilistic relational algebraic operations. A set of the properties of the basic probabilistic relational algebraic operations is also formulated and proven.

Keywords—Interval Probability Set-Valued Attribute, Probabilistic Interpretation, Extended Probabilistic Value, Probabilistic Relation, Probabilistic Relational Algebra.

1 Introduction

As we know, information in practice may be uncertain and imprecise, but the classical relational database model (CRDB) in [1], [2], and [3] is limited for representing and handling uncertain and imprecise information. Currently, there have been many non-classical database models, including probabilistic relational database models (PRDB), such as [4], [5], and [6], studied and built to overcome the limitation of CRDB. However, no model would be so universal that could include all measures and tackle all aspects of uncertainty of information in the real world.

PRDB models for uncertain and imprecise information are extended and built as extensions of CRDB based on the probability theory. There are two main types of PRDB models. The first one defines a probabilistic relation as a set of tuples such that each tuple is associated with a probability to express the uncertainty degree of it 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 represent the uncertainty degree of the values that it may take.

The first PRDB model type is the extension of CRDB at the relation level, as the works in [7], [8] and [9] thereby 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. However, in many natural situations, we cannot know precisely the probability that we can only estimate as an approximate number in a subinterval of [0, 1]. The models in [10-13] were extended with probability intervals associated with each tuple to overcome the shortcoming of the models in [7-9]. Nevertheless, the PRDB models did not express the uncertainty of attribute values of relations that it only was inferred from the uncertainty membership degree of tuple of the relations.

The second PRDB model type is the extension of CRDB at the attribute level, as the works in [14] and [15], thereby 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 generally, in [16], each attribute was associated with a probability distribution on a set of values to express the possibility that the attribute might take one of 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 that we can only estimate as an approximate number in a subinterval of [0, 1]. The model in [17] 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. Nevertheless, the model did not allow attributes to take set values and, thus it was limited in the real applications.

Recently, the model in [17] has been extended for uncertain multivalued attributes as in [18]. However, when the probabilistic relations in [17] and [18] have many attributes, the number of generated probability distribution functions is too large to lead to low performance in manipulating data. The model in [19] overcame the shortcoming of the model in [17] by using probability intervals on a set to represent attribute values. Howerver, the model in [19] did not allow multivalued attributes. For instance, the attribute P_DISEASE in [19] was represented by P_DISEASE: {(hepatitis, [0.3, 0.5]), (cirrhosis, [0.5, 0.7])} to say that the patient's disease might be hepatitis with a probability in the interval [0.3, 0.5] or cirrhosis with a probability in the interval [0.5, 0.7]. However, in practice, a patient may have both hepatitis and cirrhosis with a determined probability interval such as [0.4, 0.6], then the model in [19] cannot represent. The model in [20] overcame the shortcoming of the model in [19] by associating each relational attribute with a distribution of probability intervals on a set of value sets. However, in [20], the probabilistic functional dependency and schema key of relations haven't been defined. In addition, except the selection operation, other probabilistic relational algebraic operations haven't been built for the model in [20]. Thus, the ability of representing and dealing with uncertain information of it has been limited in the real world applications.

In this paper, we define notions of the probabilistic functional dependency and schema key of relations and extend the model in [20] with a full set of basic probabilistic relational algebraic operations to a new probabilistic relational database model to overcome the limitations of the models in [19] and [20]. The new probabilistic relational database model is abbreviated by EIPRDB to be an extension of the IPRDB model in [19] with interval probability set-valued attributes (i.e., interval probability multivalued attributes). Some properties of EIPRDB algebraic operations are also proposed, formulated and proven.

To build EIPRDB, we use extended probabilistic values in [20] for representing uncertain set-valued attributes of relations, employ probabilistic interpretations of binary relations on sets in [19], operators on probability intervals in [18], and propose new combination strategies of extended probabilistic values to define the probabilistic relational algebraic operations for computing and querying uncertain and imprecise information on EIPRDB relations. The built EIPRDB model is able to represent and manipulate effectively uncertain and imprecise information and can be applied to solve problems in real databases.

Basic probability definitions as a mathematical foundation for EIPRDB are presented in Section 2. The EIPRDB 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 EIPRDB and their properties. Section 5 presents the achieved results and discussions of the EIPRDB model. Finally, Section 6 concludes the paper and outlines further research directions.

2 Probability and Probabilistic Combination Strategies

In this section, some probability definitions and probabilistic combination strategies are presented as the basis for representing and handling uncertain information in EIPRDB.

2.1 *Extended Probabilistic Values*

Extended probabilistic values over a set of sets in [20] used to represent uncertain set-valued attributes of EIPRDB relations are defined as below.

Definition 1. Let τ be a data type and D be the domain of τ, an *extended 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 belongs to 2^D, v_i and v_j are disjointed and $0 \le l_i \le u_i \le 1$, for every $i, j = 1, 2, \ldots, m$.

Informally, an extended probabilistic value $pv = \{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ says that pv's value is exactly one member (set) v_i of the set $V = \{v_1,\ldots, v_m\}$ and the probability that pv's value is v_i belongs to the interval $[l_i, u_i]$. An extended 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 $p(v_i) \in [l_i, u_i]$, $i = 1,\ldots, m$ and $\Sigma_{v_i \in V} p(v_i) \le 1$.

Example 1. Suppose a patient's disease is diagnosed as hepatitis and cirrhosis with a probability between 0.3 and 0.5 or cholecystitis with a probability between 0.5 and 0.7. Then, this information may be represented by the extended probabilistic value {({hepatitis, cirrhosis}, [0.3, 0.5]), (cholecystitis, [0.5, 0.7])}.

We note that an element x in D is also considered as a special set $\{x\}$ on D, thus an extended probabilistic value $\{(\{x_1\}, [l_1, u_1]), (\{x_2\}, [l_2, u_2]),\ldots, (\{x_k\}, [l_k, u_k])\}$ can be written as $\{(x_1, [l_1, u_1]), (x_2, [l_2, u_2]),\ldots, (x_k, [l_k, u_k])\}$ for simplicity. Also, an extended probabilistic value can be denoted by $pv = \{(v, I)| v \in 2^D, I = [l, u] \subseteq [0, 1]\}$.

2.2 *Probabilistic Interpretation of Binary Relations on Sets*

The probabilistic interpretation of binary relations on sets in [20] used to compute the uncertain degree of relations on attribute values in EIPRDB and is defined as follows.

Definition 2. Let A and B be sets, U and V be value domains, and θ be a binary relation from $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$. The *probabilistic interpretation* of the relation $A\ \theta\ B$, denoted $Pr(A\ \theta\ B)$, 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. $Pr(A\ \theta\ B) = \begin{cases} p(u \in B|\ u{\in}A), \theta \text{ is the relation } \subseteq \\ p(u \in A|\ u{\in}B), \theta \text{ is the relation } \supseteq \end{cases}$

   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 $u{\in}A$ and $p(u \in A|\ u{\in}B)$ is the conditional probability for $u{\in}A$ given $u{\in}B$.

Example 2. Some probabilistic interpretations of the set relations on the domain consisting of natural numbers are computed as follows.

$Pr(\{4, 5\} = \{5, 6\}) = p(u = v|\ u \in \{4, 5\}, v \in \{5, 6\}) = 0.25.$

$Pr(4, 5\} < \{5, 6\}) = p(u < v\ |\ u \in\{4, 5\}, v \in \{5, 6\}) = 0.75.$

$Pr(\{4, 5\} \subseteq \{5, 6\}) = p(u \in \{5, 6\}\,|u \in\{4, 5\}) = 0.5.$

$Pr(\{4, 5\} \supseteq \{5\}) = p(u \in \{4, 5\}\,|u \in\{5\}) = 1.0.$

2.3 *Combination Strategies of Probability Intervals*

In many real situations, the probability of an event may not be defined or computed exactly. Then, a probability interval can be used 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 given in [20], where $\otimes$, $\oplus$, and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively and are defined as below.

1. Independence conjunction, disjunction, and difference strategies, denoted $\otimes_{in}$, $\oplus_{in}$, and $\ominus_{in}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{in}[l_2, u_2] = [l_1 \,.\, l_2, u_1 \,.\, u_2]$,
   - $[l_1, u_1] \oplus_{in}[l_2, u_2] = [l_1 + l_2 - (l_1 \,.\, l_2), u_1 + u_2 - (u_1 \,.\, u_2)]$,
   - $[l_1, u_1] \ominus_{in}[l_2, u_2] = [l_1 \,.\, (1 - u_2), u_1 \,.\, (1 - l_2)]$.
2. Mutual exclusion conjunction, disjunction, and difference strategies (when e_1 and e_2 are mutually exclusive), denoted $\otimes_{me}$, $\oplus_{me}$, and $\ominus_{me}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{me}[l_2, u_2] = [0, 0]$,
   - $[l_1, u_1] \oplus_{me}[l_2, u_2] = [min(1, l_1 + l_2), min(1, u_1 + u_2)]$,
   - $[l_1, u_1] \ominus_{me}[l_2, u_2] = [l_1, min(u_1, 1 - l_2)]$.
3. Positive correlation conjunction, disjunction, and difference strategies (when e_1 implies e_2, or e_2 implies e_1), denoted $\otimes_{pc}$, $\oplus_{pc}$, and $\ominus_{pc}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{pc}[l_2, u_2] = [min(l_1, l_2), min(u_1, u_2)]$,
   - $[l_1, u_1] \oplus_{pc}[l_2, u_2] = [max(l_1, l_2), max(u_1, u_2)]$,
   - $[l_1, u_1] \ominus_{pc}[l_2, u_2] = [max(0, l_1 - u_2), max(0, u_1 - l_2)]$.
4. Ignorance conjunction, disjunction, and difference strategies, denoted $\otimes_{ig}$, $\oplus_{ig}$, and $\ominus_{ig}$ respectively, are determined by:
   - $[l_1, u_1] \otimes_{ig}[l_2, u_2] = [max(0, l_1 + l_2 - 1), min(u_1, u_2)]$,
   - $[l_1, u_1] \oplus_{ig}[l_2, u_2] = [max(l_1, l_2), min(1, u_1 + u_2)]$,
   - $[l_1, u_1] \ominus_{ig}[l_2, u_2] = [max(0, l_1 - u_2), min(u_1, 1 - l_2)]$.

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 Extended Probabilistic Values*

To develop EIPRDB algebraic operations, we extend the conjunction, disjunction, and difference of probabilistic values in [19] for extended probabilistic values for combining the probability interval of set values of attributes in outcome relations of these algebraic operations as the following definitions.

Definition 3. Let pv_1 and pv_2 be two extended 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 extended probabilistic value pv defined by $pv = \{(v_1 \cap v_2, I_1 \otimes I_2) \mid (v_1, I_1) \in pv_1, (v_2, I_2) \in pv_2\}$.

Example 3. Let $pv_1 = \{$(hepatitis, [0.7, 0.8]), (cholecystitis, [0.2, 0.3])$\}$ and $pv_2 = \{$({hepatitis, cirrhosis}, [1.0, 1.0])$\}$ be extended probabilistic values, then $pv_1 \otimes_{in} pv_2$ under the independence probabilistic conjunction strategy is the extended probabilistic value $pv = \{$(hepatitis, [0.7, 0.8])$\}$.

Definition 4. Let pv_1 and pv_2 be two extended 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 extended probabilistic value pv defined by $pv = \{(v_1, I_1) \mid (v_1, I_1) \in pv_1$ and $\neg\exists\ (v_2, I_2) \in pv_2, v_2 \cap v_1 \neq \varnothing\} \cup \{(v_2, I_2) \mid (v_2, I_2) \in pv_2$ and $\neg\exists\ (v_1, I_1) \in pv_1, v_1 \cap v_2 \neq \varnothing\} \cup \{(v_1 \cup v_2, I_1 \oplus I_2) \mid (v_1, I_1) \in pv_1, (v_2, I_2) \in pv_2$ and $v_1 \cap v_2 \neq \varnothing\}$.

Example 4. Let $pv_1 = \{$({hepatitis, cirrhosis}, [0.2, 0.6]), (cholecystitis, [0.2, 0.6])$\}$ and $pv_2 = \{$({hepatitis, cirrhosis}, [0.3, 0.65]), (pancreatitis, [0.3, 0.65])$\}$ be extended probabilistic values, then $pv_1 \oplus_{in} pv_2$ under the independence probabilistic disjunction strategy is the extended probabilistic value $pv = \{$(cholecystitis, [0.2, 0.6]), (pancreatitis, [0.3, 0.65]), ({hepatitis, cirrhosis}, [0.44, 0.86])$\}$.

Definition 5. Let pv_1 and pv_2 be two extended 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 extended probabilistic value pv defined by $pv = \{(v_1, I_1) \mid (v_1, I_1) \in pv_1$ and $\neg\exists\ (v_2, I_2) \in pv_2, v_2 \cap v_1 \neq \varnothing\} \cup \{(v_1, I_1 \ominus I_2) \mid (v_1, I_1) \in pv_1, \exists\ (v_2, I_2) \in pv_2$ and $v_2 \cap v_1 \neq \varnothing\}$.

3 EIPRDB Data Model

EIPRDB data model consists of basic components such as the schema, probabilistic relation, and database to represent data and relationships between them.

3.1 *EIPRDB Schemas and Relations*

The EIPRDB schema is extended from that of CRDB with uncertain set-valued attributes as follows.

Definition 6. An *EIPRDB 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 extended probabilistic values on the domain of A.

We can use the notation $R(\boldsymbol{U}, \wp)$ and R to denote the schema $R = (\boldsymbol{U}, \wp)$ and $dom(A)$ to denote the domain of the attribute A.

An EIPRDB relation is an instance of an EIPRDB schema, where each relational attribute is associated with an extended probabilistic value to represent an uncertain value set that the attribute may take. The EIPRDB relation is extended from that of CRDB in [1] and [2] as the following definition.

Definition 7. Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes. An *EIPRDB 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 extended probabilistic values $pv_{ij} = \{(v_{ij}, [l_{ij}, u_{ij}]) \mid v_{ij} \in 2^{dom(A_j)}, [l_{ij}, u_{ij}] \subseteq [0, 1]\}, j = 1, 2, \ldots, k$ 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 attribute A_j of the tuple t_i may take a uncertain value set represented by pv_{ij}. 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 $V_{ij} = \{v_{ij} \mid (v_{ij}, [l_{ij}, u_{ij}]) \in pv_{ij}\}$. The symbol $t_i[H]$, where $H \subseteq \boldsymbol{U}$, denotes the rest of the tuple t_i after eliminating the values of attributes in $\boldsymbol{U}$ 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 5. A simple EIPRDB relation, named DIAGNOSE, over the EIPRDB schema **DIAGNOSE**({D_ID, P_ID, P_NAME, P_AGE, P_DISEASE, DATE, D_COST}, $\wp$) in the database about patients at the clinic of a hospital can be given as Table 1. 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 while D_ID and DATE represent the identifier of a doctor and the date that the doctor diagnoses the disease for a patient. In reality, while diagnosing the doctors 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 Blair says that the patient is 60 years old, diagnosed on 15/11/2024, may have lung cancer or tuberculosis with the probability 0.5 and has to pay the daily treatment cost $30 with the 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 **DIAGNOSE**, $\wp(A)$ includes all extended probabilistic values on the domain of A (Definition 6). In addition, for simplicity, each extended probabilistic value $\{(v, [1, 1])\}$, where $v \in dom(A)$, will be represented as a single value v (such as extended probabilistic values for the attribute P_ID). Because if an attribute takes such an extended probabilistic value, then it only takes a value v with the probability of 1.0 (Definition 1). In other words, the attribute certainly takes the value v.

Table 1. Relation DIAGNOSE

D_ID	P_ID	P_NAME	P_AGE	P_DISEASE	DATE	D_COST
DT093	P104	Blair	{(60, [1, 1])}	{(lung cancer, [0.5, 0.5]), (tuberculosis, [0.5, 0.5])}	15/11/2024	{($30, [0.3, 0.6]), ($35, [0.4, 0.7])}
DT102	P218	Oliver	{(46, [0.5, 0.5]), (47, [0.5, 0.5])}	{({hepatitis, cirrhosis}, [0.5, 0.7]), (cholecystitis, [0.3, 0.5])}	18/11/2024	{($8, [0.4, 0.5]), ($9, [0.5, 0.6])}
DT102	P325	Mary	{(36, [1, 1])}	{(duodenitis, [0.5, 0.5]), (gastritis, [0.5, 0.5])}	18/11/2024	{($8, [0.5, 0.5]), ($9, [0.5, 0.5])}
DT102	P412	Anna	{(15, [1, 1])}	{({bronchitis, angina}, [1, 1])}	18/11/2024	{($12, [0.5, 0.5]), ($13, [0.5, 0.5]}
DT025	P426	Bill	{(36, [1, 1])}	{(duodenitis, [0.4, 0.5]), (gastritis, [0.5, 0.6])}	19/11/2024	{($8, [0.3, 0.5]), ($9, [0.5, 0.7])}

The EIPRDB relational database is defined as an extension of CRDB with uncertain set-valued attributes as follows.

Definition 8. An *EIPRDB relational database* over a set of uncertain set-valued attributes is a set of EIPRDB relations corresponding to the set of their EIPRDB schemas.

3.2 *EIPRDB Functional Dependencies*

The functional dependency in EIPRDB is an extension of that in CRDB [3] with probabilistic valued attributes based on the probability measure for the equal degree of two extended probabilistic values of the same attribute for two different tuples in a relation as follows.

Definition 9. Let $R(\boldsymbol{U}, \wp)$ be an EIPRDB 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 $t_1.A = \{(v_{11}, [l_{11}, u_{11}]), \dots, (v_{1m}, [l_{1m}, u_{1m}])\}$, $t_2.A = \{(v_{21}, [l_{21}, u_{21}]), \dots, (v_{2n}, [l_{2n}, u_{2n}])\}$ and $\oplus$ is the mutual exclusion probabilistic disjunction operator.

Definition 10. Let $R = (\boldsymbol{U}, \wp)$ be an EIPRDB 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 *EIPRDB functional dependency* of $\boldsymbol{Y}$ on $\boldsymbol{X}$ under $\otimes$, denoted by $\boldsymbol{X} \rightarrow_{\otimes} \boldsymbol{Y}$, holds if and only if

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

It is easy to see that for every EIPRDB schema $R(\boldsymbol{U}, \wp)$, then $\boldsymbol{U} \rightarrow_{\otimes} \boldsymbol{Y}$ with $\boldsymbol{Y} \subseteq \boldsymbol{U}$ under all probabilistic conjunction strategies.

Example 6. In every relation r over the schema **DIAGNOSE** with the set of attributes $\boldsymbol{U}$ = {D_ID, P_ID, P_NAME, P_AGE, P_DISEASE, DATE, D_COST} in Example 5, the values of the attributes D_ID and P_ID that express the identifiers of doctors and patients, respectively are single and pairwise different. Thus, for two tuples $t_1, t_2 \in r$ and an attribute $A \in \boldsymbol{U}$, then $p(t_1.\text{D\_ID} =_{\otimes} t_2.\text{D\_ID}) \otimes p(t_1.\text{P\_ID} =_{\otimes} t_2.\text{P\_ID}) = 0$ and $p(t_1.A =_{\otimes} t_2.A) \geq 0$. So, $p(t_1.\text{D\_ID} =_{\otimes} t_2.\text{D\_ID}) \otimes p(t_1.\text{P\_ID} =_{\otimes} t_2.\text{P\_ID}) \leq \otimes_{A \in \boldsymbol{Y}} p(t_1.A =_{\otimes} t_2.A)$ with $\boldsymbol{Y} \subseteq \boldsymbol{U}$, by Definition 10, there is the EIPRDB functional dependency {D_ID, P_ID} $\rightarrow_{\otimes} \boldsymbol{Y}$ in the schema **DIAGNOSE** under all probabilistic conjunction strategies.

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

Definition 11. Let $R(\boldsymbol{U}, \wp)$ be an EIPRDB 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 $K \rightarrow_{\otimes} \boldsymbol{U}$ such that there does not exist any proper subset of K holding this property.

Example 7. From the result of Example 6, we have {D_ID, P_ID} $\rightarrow_{\otimes}$ $\boldsymbol{Y}$ for every $\boldsymbol{Y} \subseteq \boldsymbol{U}$. Thus {D_ID, P_ID} $\rightarrow_{\otimes}$ $\boldsymbol{U}$. Hence, {D_ID, P_ID} is a key of the schema **DIAGNOSE** under all probabilistic conjunction strategies.

4 EIPRDB Algebra

The EIPRDB algebra is a set of basic probabilistic relational algebraic operations such as the selection, projection, Cartesian product, join, intersection, union, and difference. The EIPRDB algebra or the probabilistic relational algebra is an extension of the CRDB algebra with probabilistic set values of relational attributes to manipulate, handle, and query uncertain and imprecise information on EIPRDB.

4.1 *Selection*

The selection is a basic algebraic operation in EIPRDB for querying on the relations of databases. The selection operation in EIPRDB is extended from that of CRDB taking into account uncertain set-valued relational attributes. Before defining the selection operation, we present the formal syntax and semantics of selection expressions and conditions as follows.

Definition 12. Let R be an EIPRDB 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 $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$, $c \in 2^D$, and D is $dom(A)$.
2. $x.A_1\ \theta_{\otimes} x.A_2$, where $x \in X$, A_1 and A_2 are two different 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 8. Consider the schema **DIAGNOSE** in Example 5, the selection of "all patients who have angina and pay the daily treatment cost over 9 USD" can be represented by the selection expression x.P_DISEASE = angina $\otimes$ x.D_COST > 9.

Selection conditions in EIPRDB are the extensions of those in CRDB and formally defined as below.

Definition 13. Let R be an EIPRDB 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 9. Given the schema **DIAGNOSE** in Example 5, the selection of "all patients who are not over 55 years old with a probability of at least 0.9 or have cholecystitis and pay the daily treatment cost not less than 8 USD with a probability from 0.5 to 0.7" can be done using the selection condition (x.P_AGE ≤ 55)[0.9, 1.0] $\vee$ (x.P_DISEASE = cholecystitis $\otimes$ x.D_COST ≥ 8)[0.5, 0.7].

The probabilistic interpretation of selection expressions in EIPRDB is defined on the probabilistic interpretation of binary relations of sets as below.

Definition 14. Let R be an EIPRDB 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* with respect to 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. $Prob_{R,r,t}(x.A\ \theta\ c) = \bigoplus_{i=1}^{k}[l_i, u_i].Pr(v_i\ \theta\ c)$, 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. $Prob_{R,r,t}(x.A_1\ \theta_{\otimes}\ x.A_2) = \bigoplus_{i=1}^{m}\bigoplus_{j=1}^{n}(([l_{1i}, u_{1i}] \otimes [l_{2j}, u_{2j}]).Pr(v_{1i}\ \theta\ v_{2j}))$, 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 the item 1 of Definition 14 because the extended probabilistic value $t.A = \{(v_1, [l_1, u_1]), \ldots, (v_k, [l_k, u_k])\}$ of the attribute A of the tuple t in r

represents a distribution function of probability intervals over the set $\{v_1,\ldots, v_k\}$ (Definition 1 and 7), likewise with the item 2 for the extended probabilistic values $t.A_1 = \{(v_{11}, [l_{11}, u_{11}]), \ldots, (v_{1m}, [l_{1m}, u_{1m}])\}$ and $t.A_2 = \{(v_{21}, [l_{21}, u_{21}]), \ldots, (v_{2n}, [l_{2n}, u_{2n}])\}$. Intuitively, $Prob_{R,r,t}(x.A\ \theta\ c)$ is the probability interval for the attribute A of the tuple t having a (set) 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 $v_{1i}\ \theta\ v_{2j}$ under $\otimes$.

Example 10. Let R denote the schema **DIAGNOSE** and r denote the relation DIAGNOSE in Example 5. Consider the first tuple in r, denoted by t_1. Applying the item 1 of Definition 14, using the probabilistic interpretation of the binary relations on sets in Definition 2 for the relations $30 \geq 32$ and $35 \geq 32$ and the mutual exclusion probabilistic disjunction $\oplus_{me}$ defined in the section 2.3 , we have

$$\begin{aligned} Prob_{R,r,t_1}(x.\text{ D\_COST} \geq 32) &= [0.3, 0.6].Pr(30 \geq 32) \oplus_{me}[0.4, 0.7].Pr(35 \geq 32)\\ &= [0.3, 0.6]\times 0.0 \oplus_{me} [0.4, 0.7]\times 1.0\\ &= [0, 0] \oplus_{me} [0.4, 0.7] = [0.4, 0.7]. \end{aligned}$$

The satisfaction of selection conditions in EIPRDB is defined on the probabilistic interpretation of selection expressions as below.

Definition 15. Let R be an EIPRDB 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. $Prob_{R,r,t} \vDash (\alpha)[l, u]$ if and only if (iff) $Prob_{R,r,t}(\alpha) \subseteq [l, u]$.
2. $Prob_{R,r,t} \vDash \neg\varphi$ iff $Prob_{R,r,t} \vDash \varphi$ does not hold.
3. $Prob_{R,r,t} \vDash \varphi \wedge \omega$ iff $Prob_{R,r,t} \vDash \varphi$ and $Prob_{R,r,t} \vDash \omega$.
4. $Prob_{R,r,t} \vDash \varphi \vee \omega$ iff $Prob_{R,r,t} \vDash \varphi$ or $Prob_{R,r,t} \vDash \omega$.

Now, the selection operation on a relation in EIPRDB is defined as follows.

Definition 16. Let R be an EIPRDB schema, r be a relation over R, and φ be a selection condition over a tuple variable x. The *selection* on r with respect to φ, denoted by $\sigma_\varphi(r)$, is the relation $r^* = \{t \in r \mid Prob_{R,r,t} \vDash \varphi\}$ over R, including all satisfied tuples of the selection condition φ.

Example 11. Let r denote the relation DIAGNOSE in Example 5 and R denote its schema. The query "Find all patients who are over 45 years old with a probability of at least 0.9, and have both hepatitis and cirrhosis and pay the daily treatment cost not less than 7 USD with a probability between 0.4 and 0.8" can be done by the selection operation σ_φ(DIAGNOSE), where $\varphi = (x.\text{P\_AGE} > 45)[0.9, 1.0] \wedge (x.\text{P\_DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D\_COST} \geq 7)[0.4, 0.8]$.

The selection σ_φ(DIAGNOSE) is implemented by checking the satisfaction of all tuples in the relation DIAGNOSE under Definition 15 and 16 for the selection condition φ. Applying Definition 14, we can see that only one patient denoted by the second tuple t_2 of the relation DIAGNOSE in Example 5 satisfies φ, because:

$$\begin{aligned} Prob_{R,r,t_2}(x.\text{P\_AGE} > 45) &= [0.5, 0.5]\times Pr(46 > 45) \oplus_{me} [0.5, 0.5]\times Pr(47 > 45)\\ &= [0.5, 0.5]\times 1.0 \oplus_{me} [0.5, 0.5]\times 1.0\\ &= [1.0, 1.0] \subseteq [0.9, 1.0]. \end{aligned}$$

$$\begin{aligned} &Prob_{R,r,t_2}(x.\text{P\_DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\})\\ &\qquad= [0.5, 0.7].Pr(\{\text{hepatitis, cirrhosis}\} \supseteq \{\text{hepatitis, cirrhosis}\})\\ &\qquad\quad \oplus_{me} [0.3, 0.5].Pr(\{\text{cholecystitis}\} \supseteq \{\text{hepatitis, cirrhosis}\})\\ &\qquad= [0.5, 0.7]\times 1.0 \oplus_{me} [0.3, 0.5]\times 0.0\\ &\qquad= [0.5, 0.7] \oplus_{me} [0, 0] = [0.5, 0.7]. \end{aligned}$$

$$\begin{aligned} Prob_{R,r,t_2}(x.\text{D\_COST} \geq 7) &= [0.4, 0.5]\times Pr(8 \geq 7) \oplus_{me} [0.5, 0.6]\times Pr(9 \geq 7)\\ &= [0.4, 0.5] \times 1.0 \oplus_{me} [0.5, 0.6]\times 1.0\\ &= [0.4, 0.5] \oplus_{me} [0.5, 0.6] = [0.9, 1.0]. \end{aligned}$$

$Prob_{R,r,t_2}(x.\text{P\_DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D\_COST} \geq 7) = [0.5, 0.7] \otimes_{in} [0.9, 1.0] = [0.45, 0.7] \subseteq [0.4, 0.8]$.

Hence, $Prob_{R,r,t_2} \vDash (x.\text{P\_AGE} > 45)[0.9, 1.0]$ and $Prob_{R,r,t_2} \vDash (x.\text{P\_DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D\_COST} \geq 7)[0.4, 0.8]$. Thus t_2 satisfies φ.

For the other tuples, one has $Prob_{R,r,t_i}(x.\text{P\_DISEASE} \supseteq \{\text{hepatitis, cirrhosis}\} \otimes_{in} x.\text{D\_COST} \geq 7) = [0, 0] \not\subset [0.4, 0.8]$, $\forall i \neq 2$. Thus, the result of the query is as Table 2.

Table 2. Relation σ_{φ}(DIAGNOSE)

D_ID	P_ID	P_NAME	P_AGE	P_DISEASE	DATE	D_COST
DT102	P218	Oliver	{(46, [0.5, 0.5]), (47, [0.5, 0.5])}	{({hepatitis, cirrhosis}, [0.5, 0.7]), (cholecystitis, [0.3, 0.5])}	18/11/2024	{($8, [0.4, 0.5]), ($9, [0.5, 0.6])}

4.2 *Projection*

The projection of an EIPRDB relation on a set of attributes is an extension of that of a CRDB relation with uncertain set-valued tuples such that the projected tuples having the same value should be merged into a tuple in the result relation by a probabilistic disjunction strategy.

Definition 17. Let $R(\boldsymbol{U}, \wp)$ be an EIPRDB schema, r be a relation over R, $\boldsymbol{X}$ be a subset of attributes of $\boldsymbol{U}$, $\oplus$ be a probabilistic disjunction strategy. The *projection* of r on $\boldsymbol{X}$ under $\oplus$, denoted by $\Pi_{X\oplus}(r)$, is the relation r^* over the schema R^* determined by:

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

Example 12. Consider the relation DIAGNOSE over the schema **DIAGNOSE**({D_ID , P_ID, P_NAME, P_AGE, P_DISEASE, DATE, D_COST}, $\wp$) as in Table 1, then the projection of it on the set of the attributes $\boldsymbol{X}$ = {P_AGE, P_DISEASE, D_COST} under $\oplus_{in}$ is the relation $\Pi_{X\oplus_{in}}$(DIAGNOSE) over the schema R^*({P_AGE, P_DISEASE, D_COST}, $\wp^*$) computed as in Table 3, where $\wp^*(A) = \wp(A)$, $\forall A \in \boldsymbol{X}$.

Table 3. Relation $\Pi_{\{\text{P\_AGE, P\_DISEASE, D\_COST}\}\oplus_{in}}$(DIAGNOSE)

P_AGE	P_DISEASE	D_COST
{(60, [1, 1])}	{(lung cancer, [0.5, 0.5]), (tuberculosis, [0.5, 0.5])}	{($30, [0.3, 0.6]), ($35, [0.4, 0.7])}
{(46, [0.5, 0.5]), (47, [0.5, 0.5])}	{({hepatitis, cirrhosis}, [0.5, 0.7]), (cholecystitis, [0.3, 0.5])}	{($8, [0.4, 0.5]), ($9, [0.5, 0.6])}
{(15, [1, 1])}	{({bronchitis, angina}, [1, 1])}	{($12, [0.5, 0.5]), ($13, [0.5, 0.5]}
{(36, [1, 1])}	{(duodenitis, [0.7, 0.75]), (gastritis, [0.75, 0.8])}	{($8, [0.65, 0.75]), ($9, [0.75, 0.85])}

Note that in the relation DIAGNOSE, we have $[t_3[\boldsymbol{X}]] = [t_5[\boldsymbol{X}]]$, thus two tuples, t_3 and t_5, are projected on $\boldsymbol{X}$ and merged into the tuple t_4 under the independence probabilistic disjunction strategy $\oplus_{in}$ in Table 3.

4.3 *Cartesian Product*

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

Definition 18. 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 EIPRDB 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. $r = \{t \mid t.A = t_1.A$ 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 EIPRDB relations is an extension of the natural join of two CRDB relations with uncertain set-valued tuples as below.

Definition 19. 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 EIPRDB 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:

1. $R = (\mathbf{U}, \wp)$ where $\mathbf{U} = \mathbf{U_1} \cup \mathbf{U_2}$, $\wp(A) = \wp_1(A)$ if $A \in \mathbf{U_1} - \mathbf{U_2}$, $\wp(A) = \wp_2(A)$ if $A \in \mathbf{U_2} - \mathbf{U_1}$ and $\wp(A) = \wp_1(A) = \wp_2(A)$ if $A \in \mathbf{U_1} \cap \mathbf{U_2}$.
2. $r = \{t \mid t.A = t_1.A$ if $A \in \mathbf{U_1} - \mathbf{U_2}$, $t.A = t_2.A$ if $A \in \mathbf{U_2} - \mathbf{U_1}$, $t.A = t_1.A \otimes t_2.A$ if $A \in \mathbf{U_1} \cap \mathbf{U_2}$, $t_1 \in r_1$, $t_2 \in r_2\}$.

Example 13. Let PATIENT$_1$ and PATIENT$_2$ be two EIPRDB relations as in Tables 4 and 5, then the result of the join of them under the probabilistic conjunction strategy $\otimes_{in}$ is the relation PATIENT$_1 \bowtie_{\otimes in}$ PATIENT$_2$ computed as in Table 6. Here, the name of each relation and its schema are identical and the set $\wp(A)$ for each attribute A in the schemas consists of all extended probabilistic values on *dom*(A).

Table 4. Relation PATIENT$_1$

P_ID	P_DISEASE
P325	{(bronchitis, [0.3, 0.4]), (bronchiectasis, [0.6, 0.7)}
P510	{({cholecystitis, gall-stone}, [1, 1])}

Table 5. Relation PATIENT$_2$

P_NAME	P_DISEASE
Peter	{(bronchiectasis, [1, 1])}
George	{({cholecystitis, gall-stone}, [0.5, 0.7]), (cirrhosis, [0.3, 0.5])}

Table 6. Relation PATIENT$_1$ $\bowtie_{\otimes_{in}}$ PATIENT$_2$

P_ID	P_NAME	P_DISEASE
P325	Peter	{(bronchiectasis, [0.6, 0.7)}
P510	George	{({cholecystitis, gall-stone}, [0.5, 0.7])}

4.5 *Intersection, Union and Difference*

The intersection, union, and difference of two EIPRDB relations over the same schema is an EIPRDB 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 CRDB. Thus, the operations are the extensions of those in CRDB with uncertain set-valued tuples. The intersection, union, and difference of two EIPRDB relations are defined as follows.

Definition 20. Let $R(\mathbf{U}, \wp)$ be an EIPRDB 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 EIPRDB relation r over R defined by $r = \{t \mid t.A = t_1.A \otimes t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \mathbf{U}$, such that $t_1[K] = t_2[K]\}$.

We note that the value of each key is definite under Definition 11. Thus, the notation $t_1[K] = t_2[K]$ can be used in Definition 20. 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 21. Let $R(\mathbf{U}, \wp)$ be an EIPRDB 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 EIPRDB relation r over R defined by $r = \{t_1 \in r_1 \mid \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t_2 \in r_2 \mid \forall t_1 \in r_1, t_2[K] \neq t_1[K]\} \cup \{t \mid t.A = t_1.A \oplus t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \mathbf{U}$ such that $t_1[K] = t_2[K]\}$.

Example 14. Let DIAGNOSE$_1$ and DIAGNOSE$_2$ be two EIPRDB relations over the same schema **DIAGNOSE**({P_ID, D_ID, P_DISEASE, D_COST}, $\wp$) as in Tables 7 and 8, 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 extended probabilistic values on *dom*(A). Then, the result of the union of them under $\oplus_{in}$ is the relation DIAGNOSE$_1 \cup_{\oplus_{in}}$DIAGNOSE$_2$ computed as in Table 9.

Table 7. Relation DIAGNOSE$_1$

P_ID	D_ ID	P_DISEASE	D_COST
P216	DT012	{(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])}	{($30, [0.3, 0.4]), ($35, [0.6, 0.7])}
P244	DT024	{({hepatitis, cirrhosis}, [0.2, 0.5]), (cholecystitis, [0.3, 0.6])}	{($8, [0.6, 1])}

Table 8. Relation DIAGNOSE$_2$

P_ID	D_ ID	P_DISEASE	D_COST
P218	DT012	{(lung cancer, [1, 1])}	{($30, [1, 1])}
P244	DT024	{({hepatitis, cirrhosis}, [0.3, 0.6]), (pancreatitis, [0.3, 0.7])}	{($7, [0.2, 0.5]), ($8, [0.5, 0.8])}
P252	DT025	{(dyspepsia, [1, 1])}	{($5, [1, 1])}

Table 9. Relation DIAGNOSE$_1 \cup_{\oplus_{in}}$DIAGNOSE$_2$

P_ID	D_ ID	P_DISEASE	D_COST
P216	DT012	{(lung cancer, [0.3, 0.6]), (tuberculosis, [0.4, 0.7])}	{($30, [0.3, 0.4]), ($35, [0.6, 0.7])}
P218	DT012	{(lung cancer, [1, 1])}	{($30, [1, 1])}
P252	DT025	{(dyspepsia, [1, 1])}	{($5, [1, 1])}
P244	DT024	{({hepatitis, cirrhosis}, [0.44, 0.8]), (cholecystitis, [0.3, 0.6]), (pancreatitis, [0.3, 0.7])}	{($7, [0.2, 0.5]), ($8, [0.8, 1])}

We note that the second tuple in Table 7 and the second tuple in Table 8 have the same key value coalesced into the fourth tuple under $\oplus_{in}$ in Table 9.

Definition 22. Let $R(\boldsymbol{U}, \wp)$ be an EIPRDB 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 EIPRDB relation r over R defined by $r = \{t_1 \in r_1 \mid \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t \mid t.A = t_1.A \ominus t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1[K] = t_2[K]\}$.

We note that, as for Definitions 20, the result relation in Definitions 21 and 22 does not depend on choosing the key of its schema.

4.6 *Property of Algebraic Operations*

The basic properties of EIPRDB algebra are extened from those of CRDB algebra with uncertain set-valued tuples (i.e., extended probabilistic values). These properties say that the IPRDB model is sound and coherent.

Proposition 1. Let R be an EIPRDB schema, r be a relation over R, and φ and ω be two selection conditions on r. Then,

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

Proof: Let $\rho = \sigma_\omega(r)$. By Definition 15 and 16, we have

$$\begin{aligned}\sigma_\varphi(\sigma_\omega(r)) &= \{t \in \rho \mid Prob_{R,\rho,t} \models \varphi\}\\ &= \{t \in r \mid (Prob_{R,r,t} \models \omega) \wedge (Prob_{R,\rho,t} \models \varphi)\}\\ &= \{t \in r \mid (Prob_{R,r,t} \models \omega) \wedge (Prob_{R,r,t} \models \varphi)\} \text{ (because } \rho \subseteq r)\\ &= \{t \in r \mid Prob_{R,r,t} \models \varphi \wedge \omega\} = \sigma_{\varphi \wedge \omega}(r).\end{aligned}$$

Thus, the equation $\sigma_\varphi(\sigma_\omega(r)) = \sigma_{\varphi\wedge\omega}(r)$ is proven. The equation $\sigma_\omega(\sigma_\varphi(r)) = \sigma_{\omega\wedge\varphi}(r)$ is similarly proven, since $\omega \wedge \varphi \Leftrightarrow \varphi \wedge \omega$. So, Proposition 1 is proven.

Proposition 2. Let R be an EIPRDB 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 17, it is easy to see $\Pi_{\boldsymbol{A}\oplus}(\Pi_{\boldsymbol{B}\oplus}(r)) = \Pi_{\boldsymbol{A}\cap\boldsymbol{B}\oplus}(r) = \Pi_{\boldsymbol{A}\oplus}(r)$ under the probabilistic disjunction strategy $\oplus$. Thus, the equation (2) is proven.

Proposition 3. Let R_1, R_2, and R_3 be the EIPRDB 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 EIPRDB 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 3, the conjunction of extended probabilistic values is commutative (due to the commutativity of probabilistic conjunction strategies). So, by Definition 19, it follows that $r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1$.

According to Definition 19, the results of two sides of (4) are the relations over the same schema. Moreover, according to Definition 3, the conjunction of extended probabilistic values is associative. Under Definition 19 and from the associativity of the classical relational natural join, it follows that the join of EIPRDB 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 18) is a particular case of the join, it yields the straight result of Proposition 3 below.

Corollary 1. Let R_1, R_2, and R_3 be EIPRDB 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,

$$r_1 \times r_2 = r_2 \times r_1 \quad (5)$$

$$(r_1 \times r_2) \times r_3 = r_1 \times (r_2 \times r_3) \quad (6)$$

Proposition 4. Let R be an EIPRDB schema, r_1, r_2, and r_3 be relations over R. Let $\otimes/\oplus$ be a probabilistic conjunction/disjunction strategy. Then,

$$r_1 \cap_{\otimes} r_2 = r_2 \cap_{\otimes} r_1 \quad (7)$$

$$(r_1 \cap_{\otimes} r_2) \cap_{\otimes} r_3 = r_1 \cap_{\otimes} (r_2 \cap_{\otimes} r_3) \quad (8)$$

$$r_1 \cup_{\oplus} r_2 = r_2 \cup_{\oplus} r_1 \quad (9)$$

$$(r_1 \cup_{\oplus} r_2) \cup_{\oplus} r_3 = r_1 \cup_{\oplus} (r_2 \cup_{\oplus} r_3) \quad (10)$$

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

Proof: From the commutativity and associativity of the probabilistic conjunction strategies, it follows that the conjunction of extended probabilistic values has the commutativity and associativity (Definition 3). So, the intersection of EIPRDB 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, according to Definition 20, we have Equations (7) and (8).

From the commutativity and associativity of the probabilistic disjunction strategies, it follows that the disjunction of extended probabilistic values has the commutativity and associativity (Definition 4). So, the union of EIPRDB 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, according to Definition 21, we have Equations (9) and (10).

5 Results and Discussions

As presented in previous sections, we can see that EIPRDB is an extension of CRDB and the second type PRDB models as in [14], [15], [16], [19], and [20] with extended probabilistic values (i.e., probabilistic intervals for value sets). In addition, EIPRDB also has the ability of querying and manipulating data more effectively than the second type PRDB models as in [17] and [18]. A more detailed discussion of the obtained results is as below.

5.1 *Extension of EIPRDB in representing and handling data*

As introduced above, there are two main types of the PRDB models. The first type one, named T-1PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in [0, 1], such as [7] and [9]. In the models, each relational attribute of a tuple is associated with a single value to say that the attribute may take the value with a probability computed and inferred from the membership degree of the tuple. The T-1PRDB algebraic operations are defined by directly extending the CRDB algebraic operations based on computing and combining probabilities of tuples in the T-1PRDB relations.

The second type one, named T-2PRDB, represents a probabilistic relation as a set of tuples whose membership degree is a probability in {0, 1}, such as [14] and [15]. In the models, each relational attribute of a tuple 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 T-2PRDB such as [16], named ET-2PRDB, where each relational attribute of a tuple is associated with a probability distribution as $\{(v_1, p_1),..., (v_m, p_m)\}$ to say that the attribute may take one of values v_i with the probability p_i. The T-2PRDB and ET-2PRDB algebraic operations are defined by extending the CRDB algebraic operations using the operators on single probabilities or probability distributions for computing and combining probabilities of attribute values in the T-2PRDB or ET-2PRDB relations.

Because, in some real cases, we cannot know precisely the probability p_i in the distribution function $\{(v_1, p_1),..., (v_m, p_m)\}$, thus the ET-2PRDB models are extended with probabilistic values to a new T-2PRDB model and named IPRDB as in [19]. In IPRDB, each relational attribute of a tuple is associated with a probabilistic value $\{(v_1, [l_1, u_1]), \ldots, (v_m, [l_m, u_m])\}$ to say that the attribute may take one of single values v_i with a probability in the interval $[l_i, u_i]$. The IPRDB algebraic operations are defined by extending the CRDB algebraic operations employing the operators on interval probabilities for computing and combining probabilities of attribute values in the IPRDB relations.

However, the IPRDB model cannot represent and handle set-valued attributes (i.e., multivalued attributes), consequently, it is extended to a new T-2PRDB model and named UIRDB as in [20] that can express and manipulate set-valued attributes. In UIRDB, each relational attribute of a tuple is associated with an extended probabilistic value $\{(V_1, [l_1, u_1]), \ldots, (V_m, [l_m, u_m])\}$ to say that the attribute may take one of value sets V_i with a probability in the interval $[l_i, u_i]$. The UIRDB selection operation is defined by extending the IPRDB selection operation using the probabilistic interpretations of binary relations on sets and the combination strategies of probabilistic intervals of extended probabilistic values in the UIRDB relations. Nevertheless, the probabilistic functional dependency, schema key of relations as well as other probabilistic relational algebraic operations haven't been defined in UIRDB. Thus, the ability of representing and dealing with uncertain information of it has been limited in the real world applications.

As presented in previous sections, we can see that the proposed EIPRDB model belongs to the second type PRDB models (T-2PRDBs), where each relational attribute of a tuple is associated with an extended 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 value sets) to say that the attribute may take one set of values V_i with a probability in $[l_i, u_i]$. Thus, the EIPRDB model is an extension of the IPRDB model in [19] with set-valued attributes. Moreover, the EIPRDB model is also an extension of the UIRDB model in [20] with the probabilistic functional dependency (PFD), schema key of relations and a full set of probabilistic relational algebraic operations. The EIPRDB algebraic operations are defined by extending the IPRDB algebraic operations employing the probabilistic interpretations of binary relations on sets and the combination strategies of probabilistic intervals of extended probabilistic values in the EIPRDB relations. Figure 1 illustrates the extension of EIPRDB in comparison with the CRDB, other T-2PRDB models.

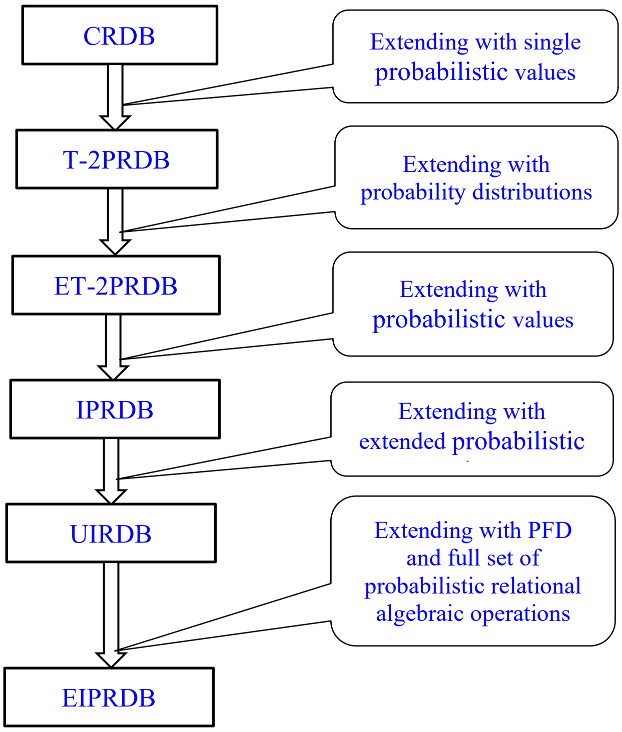
CRDB
Extending with single probabilistic values
T-2PRDB
Extending with probability distributions
ET-2PRDB
Extending with probabilistic values
IPRDB
Extending with extended probabilistic
UIRDB
Extending with PFD and full set of probabilistic relational algebraic operations
EIPRDB

Figure 1. Extension of EIPRDB

5.2 *Efficiency of EIPRDB in computing and manipulating data*

As we have known, in CRDB model, the computing complexity of relational algebraic operations is O(n) for the selection and projection on a relation having n tuples and O(nm) for the Cartesian product, join, intersection, union, and difference on two relations having n and m tuples.

In ET-2PRDB models, such as the model in [16], since each relational attribute is represented by a probability distribution function of a set of values, the computing complexity of relational algebraic operations is O(*kn*) for the selection and projection on a relation having *n* tuples and O(*knm*) for the Cartesian product, join, intersection, union, and difference on two relations having *n* and *m* tuples, where *k* is the cardinality of the domain of the distribution function.

In EIPRDB, UIRDB and IPRDB models, since each relational attribute is represented by a list of some values or data associated with probability intervals (i.e., an extended probabilistic value or a probabilistic value), the computation and manipulation on the EIPRDB, UIRDB, and IPRDB data models are more effective than those on the T-2PRDB data models in [17] and [18], where the relational attribute value is the probability distribution function pairs of a set of values.

The computing complexity of EIPRDB algebraic operations is a polynomial under the size of probabilistic relations, and it is as effective as the computing complexity of CRDB 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 14 and 15), then the cost for the selection of each tuple in an EIPRDB relation (Definition 16) also is some constant or O(1). Thus, the computing time complexity of the selection operation on an EIPRDB relation with *n* tuples is O(*n*). With the projection, from Definition 17, 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 EIPRDB relation having *n* tuples is O(*n*). Similarly, the computing time complexity of Cartesian product, join, intersection, union, and difference operations on two EIPRDB relations having *n* and *m* tuples is O(*nm*).

From the discussions above, we can say that the performance of the EIPRDB model in computing and manipulating uncertain and imprecise information is good and can be applied in practice.

6 Conclusions

We have proposed a new probabilistic relational database model, named EIPRDB, that extends the CRDB model with interval probability set-valued attributes for uncertain information. A defined EIPRDB relation includes tuples whose attributes may take an extended probabilistic value to represent uncertainty and imprecision of information of objects in the real world. The fundamental concepts of EIPRDB such as the relational schema, probabilistic functional dependency, key and probabilistic database have been defined as the extensions of those of CRDB with interval probability set-valued attributes and tuples. The EIPRDB algebra has been built using the probabilistic interpretation of binary relations on sets, probabilistic combination strategies, and conjunction, disjunction, difference operations of extended probabilistic values. Basic properties of EIPRDB algebra have been proposed and proven formally and coherently. The EIPRDB model is consistent with the CRDB model and can express, manipulate, and deal with effectively uncertain and imprecise data.

Towards applying EIPRDB model, we will build a management system for EIPRDB with the familiar querying and manipulating language like SQL that is able to represent and handle uncertain and imprecise information in the real world. To build the management system for EIPRDB, a data type for extended probabilistic values will be defined, then a compiler will be developed to compile the EIPRDB algebraic language (i.e., probabilistic relational algebraic operations) to probabilistic relational queries and manipulations in SQL.

BIBLIOGRAPHY

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

[3] 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.

[4] Z. Ma and L. Yan, *Advances in Probabilistic Databases for Uncertain Information Management*, Springer-Verlag Berlin Heidelberg, 2015.

[5] V.V. Kheradkar and S. K. Shirgave, “Query Processing over Relationalcross Model in Uncertain and Probabilistic Databases,” *Proceedings of 3Th International Conference on Artificial Intelligence and Smart*

Energy, Coimbatore, India, pp.763-769, 2023.

[6] D. Suciu, "Probabilistic Databases for All," *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems,* USA, pp.19–31, 2020.

[7] 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.

[8] T. Friedman, G. Broeck, "Symbolic Querying of Vector Spaces: Probabilistic Databases Meets Relational Embeddings," *Proceedings of 36th Conference on Uncertainty in Artificial Intelligence*, Canada, vol.124, pp.1268-1277, 2020.

[9] H. Debbi, "Explaining Query Answers in Probabilistic Databases," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no.4, pp.140-152, 2023.

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

[11] R. Ross and V.S. Subrahmanian, "Aggregate Operators in Probabilistic Databases," *Journal of the ACM*, vol.52, no.1, pp.54-101, 2005.

[12] H. Nguyen, "Extending Relational Database Model for Uncertain Information," *Journal of Computer Science and Cybernetics*, vol.35, no.4, pp.355-372, 2019.

[13] C. Zhang, Z. Mei, B. Wu, Z. Zhao, J. Yu, Q. Wang, "Query with Assumptions for Probabilistic Relational Databases," *Technical gazette*, vol. 27, no. 3, pp.923-932, 2020.

[14] J. Bernad, C. Bobed and E. Mena, "Uncertain Probabilistic Range Queries on Multidimensional Data," *Information Sciences*, vol. 537, pp.334-367, 2020.

[15] K. Papaioannou, M. Theobald, and M. Böhlen, "Supporting Set Operations in Temporal-Probabilistic Databases," *Proceedings of the 34th IEEE International Conference on Data Engineering*, France, pp. 1180-1191, 2018.

[16] S.K. Lee, "An Extended Relational Database Model for Uncertain and Imprecise Information," *Proceedings of 18th Conference on Very Large Data Bases*, Canada, pp.211-220, 1992.

[17] H. Nguyen, "A Probabilistic Relational Database Model and Algebra," *Journal of Computer Science and Cybernetics*, vol. 31, no.4, pp.305-321, 2015.

[18] 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.

[19] H. Nguyen, and D.N. Le, "A Relational Database Model with Interval Probability Valued Attributes for Uncertain and Imprecise Information," *ECTI Transactions on Computer and Information Technology,* vol.18, no.3, pp.307- 318, 2024.

[20] H. Nguyen, and T.N. Tran, "A Relational Database Model with Probability Intervals for Uncertain Set-Valued Attributes," *Malaysian Journal of Science and Advanced Technology*, vol.4, no.4, pp. 456-463, 2024.