Journal of Computer Science and Cybernetics, V.35, N.4 (2019), 355–372
DOI 10.15625/1813-9663/35/4/13907

EXTENDING RELATIONAL DATABASE MODEL FOR UNCERTAIN INFORMATION

NGUYEN HOA

Department of Information Technology, Saigon University

Faculty of Information Technology, Industrial University of Ho Chi Minh City

nguyenhoa@sgu.edu.vn

Abstract. In this paper, we propose a new probabilistic relational database model, denoted by PRDB, as an extension of the classical relational database model where the uncertainty of relational attribute values and tuples are respectively represented by finite sets and probability intervals. A probabilistic interpretation of binary relations on finite sets is proposed for the computation of their probability measures. The combination strategies on probability intervals are employed to combine attribute values and compute uncertain membership degrees of tuples in a relation. The fundamental concepts of the classical relational database model are extended and generalized for PRDB. Then, the probabilistic relational algebraic operations are formally defined accordingly in PRDB. In addition, a set of the properties of the algebraic operations in this new model also are formulated and proven.

Keywords. Probability Interval; Probabilistic Combination Strategy; Probabilistic Relation; Probabilistic Functional Dependency; Probabilistic Relational Algebraic Operation.

1. INTRODUCTION

Although the classical relational database model [3, 4], denoted by CRDB, is very useful for modeling, designing and implementing large-scale systems, it is restricted for representing and handling uncertain and imprecise information that are pervasive in the real world [6, 11, 13]. For example, applications of the CRDB model can neither deal with queries as "find all patients who are young"; nor "find all patients who are at least 90% likely to catch either hepatitis or cirrhosis", etc. Here, "young" is a vague concept that can be defined by a fuzzy set [28] or a possibility distribution [17], and "hepatitis or cirrhosis" uncertainly expresses a patient's possible diseases that can be represented by the discrete set comprising of the two diseases. Meanwhile, "90%" is the uncertainty degree, i.e., probability, of that whole fact about the patient. To overcome the shortcoming of CRDB, this model has to be extended for uncertain and imprecise information.

For building database models, uncertainty and imprecision are two different aspects of information that require respective theories and methods to handle. In particular, the fuzzy set theory is employed to express and handle imprecise information and extend CRDB to fuzzy relational database (FRDB) models, meanwhile the probability theory is used to represent and manipulate uncertain information and develop CRDB to probabilistic relational database (PRDB) models.

Up to now, many FRDB models have been built, e.g. in [1, 17, 20], and a large number of PRDB models have been proposed, e.g. in [2, 5, 6, 9, 10, 14, 16, 22, 24, 27], respectively for representing and handling uncertain and imprecise information. However, no model would be so universal that could include all measures and tackle all facets of uncertain and imprecise information. Thus, new databases model still continue to be developed for modeling data objects of the real world.

PRDB models have been extended from CRDB in these two main directions [11] (1) at the attribute level, where uncertain values of an attribute are defined by a probability associating with a value on the domain of that attribute; Or (2) at the relational tuple level, where attribute values are precise, but each tuple in a relation is associated with a probability measure that expresses the uncertainty degree of that tuple in the relation.

For instances, in [2, 6, 9, 13, 15], the value of an attribute was assigned to a probability to represent the uncertain level for that attribute to take the value. The models in [22, 27] allowed the value of each attribute associated with a probability interval to represent the uncertain degree of both the probability and the value that the attribute could take. More flexibly, the model in [7] represented the value of each attribute as a probability distribution on a set. It means that each attribute associated with a set of values and a probability distribution expressing possibility that the attribute can take one of values of the set with a probability computed from the distribution. The models in [18, 19] extended more the model in [7], where a pair of lower and upper bound probability distributions is used instead of a probability distribution as in [7]. In [10, 26], each tuple in a relation had an uncertainty degree, measured by a probability value for it belonging to the relation. The model in [5] extended the models in [10, 26], where used a pair of lower and upper bound probabilities [23] instead of a probability to represent the uncertain degree for a tuple belonging to a relation.

The models mentioned above are extensions with probability of the CRDB model in different levels to represent uncertain information of objects in practice. However, these models still have the restrictions. Particularly, regarding the models in [2, 6, 9, 10, 13, 15, 26], the probability associated with each tuple or each attribute value is not always determined exactly in practice. The models in [7, 18, 19, 22, 27] overcame the shortcoming of the models in [2, 6, 9, 13, 15] by estimating a probability interval or a pair of lower and upper bound probabilities for each attribute value of relations. However, in [7, 18, 19, 22, 27], the uncertain degree of each tuple in a relation was not represented. Meanwhile, in contrast for the models in [5, 10, 26], each tuple had a probability for it belonging to a relation but the attribute value of the tupe is single and the probability for that attribute taking the single value was not known.

In this paper, we propose a new probabilistic relational database model (PRDB) that combines the representable relevance and strength of both the relational attribute level and the relational tuple level for dealing with uncertain information. To build the PRDB model, we express the value of an attribute as a finite set and associate each relational tuple with a probability interval, then we propose a probabilistic interpretation of binary relations on sets and use the combination strategies on probability intervals in [8] to define all the basic concepts and probabilistic relational algebraic operations for PRDB. It is due to combining both of the representable levels for uncertain information, our model can overcome the shortcomings of the above mentioned models to represent and manipulate uncertain information in practice.

Basic probability definitions as a mathematical foundation for PRDB are presented in Section 2. The PRDB model including the fundamental concepts such as schema, relation and probabilistic functional dependency is introduced in Section 3. Section 4 and 5 present probabilistic relational algebraic operations and their properties in PRDB. Finally, Section 6 concludes the paper and outlines further research directions in the future.

2. BASIC PROBABILITY DEFINITIONS

This section presents some basic probability definitions to build PRDB for representing and handling uncertain information.

2.1. Probabilistic interpretation of relations on sets

For computing the probability of a binary relation on atrribute values in PRDB, we propose the probabilistic interpretation of binary relations on sets as following definitions.

Definition 1. 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 $prob(A\ \theta\ B)$, is a value in $[0,1]$ that is defined by

1. $prob(A\ \theta\ B) = p(u\ \theta\ v | u \in A,\ v \in B)$, 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. $prob(A \Rightarrow B) = p(u \in B | u \in A)$, 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 "$y \in B$", $prob(A\ \theta\ B)$ is the probability for $x\ \theta\ y$ being true. Meanwhile $prob(A \Rightarrow B)$ is that, given a proposition "$x \in A$" being true, $prob(A \Rightarrow B)$ is the probability for "$x \in B$" being true.

Example 1. Let $A = \{3, 4\}$ and $B = \{4, 5\}$ be two sets on the domain $\{1, 2, 3, 4, 5, 6\}$. Then

1. $prob(A = B) = p(u = v | u \in A, v \in B)$
   $= p(u = v | u \in \{3, 4\}, v \in \{4, 5\}) = 0.25.$
2. $prob(A \Rightarrow B) = p(u \in B | u \in A)$
   $= p(u \in \{4, 5\} | u \in \{3, 4\}) = 0.5.$

2.2. Probabilistic combination strategies

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 $e_1 \vee e_2$, or 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 [8, 19] as presented in Table 1, where $\otimes$, $\oplus$ and $\ominus$ denote the conjunction, disjunction, and difference operators, respectively.

Table 1. Definitions of probabilistic combination strategies

Strategy	Operators
Ignorance	$([L_1, U_1] \otimes_{ig} [L_2, U_2]) \equiv [\max(0, L_1 + L_2 - 1), \min(U_1, U_2)]$ $([L_1, U_1] \oplus_{ig} [L_2, U_2]) \equiv [\max(L_1, L_2), \min(1, U_1 + U_2)]$ $([L_1, U_1] \ominus_{ig} [L_2, U_2]) \equiv [\max(0, L_1 - U_2), \min(U_1, 1 - L_2)]$
Independence	$([L_1, U_1] \otimes_{in} [L_2, U_2]) \equiv [L_1.L_2, U_1.U_2]$ $([L_1, U_1] \oplus_{in} [L_2, U_2]) \equiv [L_1 + L_2 - (L_1.L_2), U_1 + U_2 - (U_1.U_2)]$ $([L_1, U_1] \ominus_{in} [L_2, U_2]) \equiv [L_1.(1 - U_2), U_1.(1 - L_2)]$
Positive correlation (when e_1 implies e_2, or e_2 implies e_1)	$([L_1, U_1] \otimes_{pc} [L_2, U_2]) \equiv [\min(L_1, L_2), \min(U_1, U_2)]$ $([L_1, U_1] \oplus_{pc} [L_2, U_2]) \equiv [\max(L_1, L_2), \max(U_1, U_2)]$ $([L_1, U_1] \ominus_{pc} [L_2, U_2]) \equiv [\max(0, L_1 - U_2), \max(0, U_1 - L_2)]$
Mutual exclusion (when e_1 and e_2 are mutually exclusive)	$([L_1, U_1] \otimes_{me} [L_2, U_2]) \equiv [0, 0]$ $([L_1, U_1] \oplus_{me} [L_2, U_2]) \equiv [\min(1, L_1 + L_2), \min(1, U_1 + U_2)]$ $([L_1, U_1] \ominus_{me} [L_2, U_2]) \equiv [L_1, \min(U_1, 1 - L_2)]$

3. PROPOSED PRDB MODEL

As for CRDB, the schema, relation, functional dependency and key are the fundamental concepts in the PRDB model.

3.1. PRDB schemas

A *PRDB schema* consists of a set of attributes (as in CRDB) associated with a membership function representing the lower-bound and upper-bound probabilities for an instance tuple of the relational attributes being true and is defined as follows.

Definition 2. A PRDB schema is a pair $R = (\boldsymbol{U}, \wp)$, where

1. $\boldsymbol{U} = \{A_1, A_2, ..., A_k\}$ is a set of pairwise different relational attributes.

2. $\wp$ is a function that maps each $(v_1, v_2, ..., v_k) \in 2^{D_1} \times 2^{D_2} \times ... \times 2^{D_k}$ to a subinterval of the interval $[0, 1]$, D_i is the domain of the attribute A_i, $i = 1, ..., k$.

We note that a precise value can be considered as a special set. That is, each precise value $v \in D$ can be defined as a set $\{v\}$ on D. Therefore, the above definition can accommodate relational attributes whose values are precise as in CRDB. Also, the PRDB schema is actually a generalization of the probabilistic relational database schemas in [6, 26, 27], where relational attributes could take only precise and single values.

As in CRDB, the notations $R(\boldsymbol{U}, \wp)$ and R can be used to replace $R = (\boldsymbol{U}, \wp)$. In addition, each $t = (v_1, v_2, ..., v_k)$ is called a tuple on the set of attributes $A_1, A_2, ..., A_k$, the domain of each attribute A is denoted by $dom(A)$.

Example 2. Suppose a PRDB schema **PATIENT**(P_ID, P_NAME, P_AGE, P_DISEASE, P_COST, $\wp$), where the attributes P_ID, P_NAME, P_AGE, P_DISEASE, and P_COST respectively describe the information about the identifier, name, age, disease, and daily treatment cost of each patient, $\wp$ maps each information tuple of patients to an interval $[\alpha, \beta] \subseteq [0, 1]$.

3.2. PRDB relations

A PRDB relation is an instance of a PRDB schema in which each relational attribute takes a value set in its domain and each tuple takes a probability interval on $[0, 1]$ as the definition below.

Definition 3. Let $\boldsymbol{U} = \{A_1, A_2, ..., A_k\}$ be a set of k pairwise different attributes. A *PRDB relation* r over the schema $R = (\boldsymbol{U}, \wp)$ is a finite set $\{t | t = (v_1, v_2, ..., v_k) \in 2^{D_1} \times 2^{D_2} \times ... \times 2^{D_k}$, $\wp(t) = [\alpha, \beta] \subseteq [0, 1]\}$ where $\wp(t)$ represents the probabilistic membership degree of t in r and D_i is the domain of the attribute A_i for every $i = 1, 2, ..., k$.

We note that each component v_i of the tuple $t = (v_1, v_2, ..., v_k)$ in a PRDB relation r is a set in 2^{D_i} but the attribute A_i only takes one of the values in v_i and $\wp(t)$ expresses the uncertain membership degree of the tuple t, that is a probability between α and β. Definition 3 is a proper extension of the definitions of relations in CRDB and the models in [6, 26, 27], where the value of an attribute was certain and the membership degree of a tuple was 0, 1 or a single probability. As in [3, 7, 26], the PRDB model adopts the closed world assumption (CWA). It means, for every tuple $t = (v_1, v_2, ..., v_k)$ on the set of attributes $\boldsymbol{U} = \{A_1, A_2, ..., A_k\}$ of the schema $R(\boldsymbol{U}, \wp)$ such that $\wp(t) = [0, 0]$ (Definition 2) then there does not exist any relation r over R including t.

For each probabilistic tuple t, we write $t.A_i$ and $t.\wp$ to denote the value (set) v_i and the probability interval $[\alpha, \beta]$, respectively. For each set of attributes $\boldsymbol{X} \subseteq \{A_1, A_2, ...A_k\}$, the notation $t[\boldsymbol{X}]$ is used to denote the rest of t after eliminating the value of attributes not belonging to $\boldsymbol{X}$.

Example 3. Table 2 shows an example relation PATIENT over the schema **PATIENT** in Example 2. For the attributes P_ID and P_NAME, their values are assumed to be single, certain. In reality, while being diagnosed, the actual disease of a patient may still be uncertain. Similarly, the treatment cost for patients is also not known definitely even as the patients know their diseases. Therefore, for the attribute P_DISEASE, its values can be as certain as "tuberculosis" or as uncertain as {hepatitis, cirrhosis} meaning the patients disease could be either "hepatitis" or "cirrhosis". For the attribute P_COST, the value "85" means 85 USD per day, $\{175, 200\}$ means that the daily treatment cost can be either 175 or 200 USD.

Meanwhile, the probability interval [0.8, 1], for instance, expresses that the uncertain degree for the tuple (P442, Mary, 16, {hepatitis, cirrhosis}, $\{7, 8\}$) belonging to the relation PATIENT is between 0.8 and 1.

Table 2. Relation PATIENT

P_ID	P_NAME	P_AGE	P_DISEASE	P_COST	$\wp$
P115	John	65	tuberculosis	$\{175, 200\}$	[0.9, 1]
P226	Anna	50	bronchitis	6	[1, 1]
P338	Bill	30	cholecystitis	85	[0.7, 1]
P442	Mary	16	{hepatitis, cirrhosis}	$\{7, 8\}$	[0.8, 1]
P555	Paul	$\{45, 46\}$	diabetes	$\{5, 6\}$	[1, 1]

Now, the notion of a probabilistic relational database is defined as follows.

Definition 4. A *probabilistic relational database* over a set of attributes is a set of probabilistic relations corresponding with the set of their probabilistic relational schemas.

Note that, if we only care about a unique relation over a schema then we can unify its symbol name with its schemas name.

3.3. PRDB value-equivalent tuples

A relational database model, either being classical or non-classical, does not allow redundant tuples in a relation, i.e., those whose respective attribute values are equal. For the model in [6], where relational attributes could take only precise values and the uncertain membership degree of tuples was a single probability value, the authors introduced the notion of value-equivalence. Two tuples were said to be value-equivalent if and only if their respective relational attribute values are equal. Then they should be coalesced into a single tuple with the same relational attribute values and the combined uncertain membership degree as the sum of their ones. Similarly, in [7], identical tuples as the result of the projection operation were also coalesced.

In [26], the authors added the notion of ε-equality. Two tuples were said to be ε-equal if and only if they are value-equivalent, as defined in [6], and the absolute difference of their probabilistic attribute values is less than ε.

In our proposed PRDB, since relational attribute values can be proper sets, the value-equivalence of two tuples is not the matter of "to be or not to be" as in [6] but to a certain degree. To be coherent with the probabilistic framework of the model, we evaluate the likelihood of the value-equivalence of two tuples and introduce the notion of ε-value-equivalence as follows.

Definition 5. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, $\boldsymbol{X}$ be a subset of $\boldsymbol{U}$, t_1 and t_2 be two tuples on $\boldsymbol{X}$, $\varepsilon \in [0,1]$. Then t_1 and t_2 are said to be $\varepsilon - value - equivalent$ with respect to a probabilistic conjunction strategy $\otimes$, denoted by $t_1 \approx_{\varepsilon\otimes} t_2$ if and only if $\otimes_{A\in X} prob\,(t_1.A = t_2.A) \geq \varepsilon$.

We note that, by Definition 1, $prob(t_1.A = t_2.A)$ is the probability for the values of attribute A in t_1 and t_2 being equal. The introduction of ε-value-equivalence is to coalesce two PRDB tuples under some probabilistic combination strategy if their equality likelihood is greater than or equal to a certain threshold ε, or not to coalesce them otherwise. The definition of value-equivalence in [6] could be considered as a special case of our definition with $\varepsilon = 1$.

Example 4. Suppose there is a new piece of information coming for John and the following tuple is added to the relation in Example 3: $\langle$(P115 John, 65, tuberculosis, 175), $[0.8, 1]\rangle$. Then the value-equivalence likelihood of these two tuples about John, namely t_1 and t_2, under the independence probabilistic conjunction strategy $\otimes_{in}$ is

$$\otimes_{A\in\{\text{P\_ID, P\_NAME, P\_AGE, P\_DISEASE, P\_COST}\}} prob(t_1.A = t_2.A) = 1 \times 1 \times 1 \times 1 \times 0.5 = 0.5$$

and t_1 and t_2 can be coalesced into the tuple t under the equivalent threshold $\varepsilon = 0.5$ and the independence probabilistic disjunction strategy $\oplus_{in}$, where $t.A = t_1.A \cap t_2.A$, $\wp(t) = \wp\,(t_1) \oplus_{in} \wp\,(t_2)$. That is

$$t = (\text{P115, John, 65, tuberculosis, 175}) \text{ with } \wp(t) = [0.9, 1] \oplus_{in} [0.8, 1] = [0.98, 1].$$

3.4. PRDB functional dependencies

Functional dependencies play an important role in CRDB. In [18, 19] a probabilistic functional dependency was defined under the probability degree for values of two attributes being equal. Meanwhile, functional dependencies were not formally defined in previous works. For PRDB, our definition is as follows.

Definition 6. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r be a relation over R, $\otimes$ be a probabilistic conjunction strategy, $\boldsymbol{X} \subseteq \boldsymbol{U}$ and $\boldsymbol{Y} \subseteq \boldsymbol{U}$. A *PRDB functional dependency* of $\boldsymbol{Y}$ on $\boldsymbol{X}$ under $\otimes$, denoted by $\boldsymbol{X} \rightarrow_{\otimes} \boldsymbol{Y}$, holds if and only if

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

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

Example 5. 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, 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 $A \in \boldsymbol{U}$, $prob(t_1.\text{P\_ID} = t_2.\text{P\_ID}) = 0$, while $prob\,(t_1.A = t_2.A) \geq 0$. So, $\otimes_{A \in Y}$ $prob\,(t_1.A = t_2.A) \geq 0$ with $\boldsymbol{Y} \subseteq \boldsymbol{U}$, by Definition 6, there is the PRDB functional dependency P_ID$\rightarrow_{\otimes}$ $\boldsymbol{Y}$ in the schema **PATIENT** under all probabilistic conjunction strategies.

As for CRDB, the values of the key attributes of a schema in PRDB are the basis to identify a tuple in a relation, as defined below.

Definition 7. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r be a relation over R, and $\otimes$ be a probabilistic conjunction strategy. A non-empty set of attributes $\mathbb{K} \subseteq \boldsymbol{U}$ is a *key* of R under $\otimes$ if and only if there is a probabilistic functional dependency $\mathbb{K} \rightarrow_{\otimes} \boldsymbol{U}$ such that there does not exist any proper subset of $\mathbb{K}$ holding this property.

Example 6. 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.

Note that, by Definition 7, for every PRDB schema $R(\boldsymbol{U}, \wp)$, the set of attributes $\boldsymbol{U}$ is a key of the schema.

4. PRDB ALGEBRAIC OPERATIONS

As for CRDB [3, 4], the basic operations on PRDB are the selection, projection, Cartesian product, join, intersection, union, and difference. We now extend those operations of CRDB for PRDB taking into account set values and uncertain tuple membership degrees in relations.

4.1. Selection

The selection is a basic algebraic operation and is used to query information in relational databases. Before defining the selection operation for PRDB, we present the formal syntax and semantics of selection conditions by extending those definitions of CRDB with probability and set values. We start with the syntax of selection expressions as the following definition.

Definition 8. Let R be a PRDB schema and X be a set of its tuple variables. Then *selection expressions* are inductively defined to have one of the following forms:

1. $x.A\ \theta\ c$, where $x \in X$, A is a relational attribute in R, θ is a binary relation from $\{=,\ \neq,\ \leq, <,\ \geq,\ >,\ \Rightarrow\}$, and c is a finite set in $dom(A)$.

2. $x.A_1 = x.A_2$, where $x \in X$, A_1 and A_2 are two different relational attributes in R with $dom(A_1) = dom(A_2)$.

3. $E_1 \otimes E_2$, where E_1 and E_2 are selection expressions on the same tuple variable and $\otimes$ is a probabilistic conjunction strategy.

4. $E_1 \oplus E_2$, where E_1 and E_2 are selection expressions on the same tuple variable and $\oplus$ is a probabilistic disjunction strategy.

Example 7. Consider the schema **PATIENT** in Example 2, the selection of "all patients who get cirrhosis and pay the daily treatment cost over 7 USD" can be expressed by the selection expression x.P_DISEASE = cirrhosis $\otimes$ x.D_COST > 7.

Each selection condition is defined as a logical combination of selection expressions with probability intervals to be satisfied.

Definition 9. Let R be a PRDB schema. Then *selection conditions* are inductively defined as follows:

1. If E is a selection expression and $[L, U]$ is a subinterval of $[0, 1]$, then $(E)[L, U]$ is a selection condition.

2. If ϕ and ψ are selection conditions on the same tuple variable, then $\neg\phi$, $(\phi \wedge \psi)$, $(\phi \vee \psi)$ are selection conditions.

Example 8. Given the schema **PATIENT** in Example 3, the selection of "all patients who are over 25 years old with a probability of at least 0.9 or have hepatitis and pay the daily treatment cost not less than 7 USD with a probability from 0.4 to 0.6" can be done using the selection condition (x.P_AGE$>$ 25)$[0.9, 1] \vee$ (x.P_DISEASE = hepatitis $\otimes$ x.D_COST$\geq$ 7)$[0.4, 0.6]$.

The probabilistic interpretation (i.e., semantics) of selection expressions is defined by extending those definitions of CRDB with probability and set values as below.

Definition 10. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB 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. $prob_{R,r,t}(x.A\ \theta\ c) = [\alpha.prob(t.A\ \theta\ c), \beta.prob(t.A\ \theta\ c)]$, where $\wp(t) = [\alpha, \beta]$.

2. $prob_{R,r,t}(x.A_1 = x.A_2) = [\alpha.prob(t.A_1 = t.A_2), \beta.prob(t.A_1 = t.A_2)]$, where $\wp(t) = [\alpha, \beta]$.

3. $prob_{R,r,t}(E_1 \otimes E_2) = prob_{R,r,t}(E_1) \otimes prob_{R,r,t}(E_2)$.

4. $prob_{R,r,t}(E_1 \oplus E_2) = prob_{R,r,t}(E_1) \oplus prob_{R,r,t}(E_2)$.

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 such that $v\theta c$. Meanwhile, $prob_{R,r,t}(x.A_1 = x.A_2)$ is the probability interval for the attributes A_1 and A_2 of the tuple t having values v_1 and v_2, respectively, such that $v_1 = v_2$.

Example 9. Let r denote the relation PATIENT in Example 3 and R denote the schema of PATIENT, regarding the fourth tuple t_4 in r, one has

$$prob_{R,r,t_4}(x.\text{P\_COST} \geq 7) = [0.8 \times prob(\{7, 8\} \geq 7), 1 \times prob(\{7, 8\} \geq 7)] = [0.8, 1].$$

Definition 10 is different from the probabilistic interpretation in [19] because, unlike that model, our PRDB contains the probabilistic interval for tuples in a relation. On the basis of the probabilistic interpretation of selection expressions, the satisfaction (i.e., semantics) of selection conditions in PRDB is defined below.

Definition 11. Let R be a PRDB 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. $prob_{R,r,t} \vDash (E)[L, U]$ if and only if (iff) $prob_{R,r,t}(E) \subseteq [L, U]$.

2. $prob_{R,r,t} \vDash \neg\phi$ iff $prob_{R,r,t} \vDash \phi$ does not hold.

3. $prob_{R,r,t} \vDash \phi \wedge \psi$ iff $prob_{R,r,t} \vDash \phi$ and $prob_{R,r,t} \vDash \psi$.

4. $prob_{R,r,t} \vDash \phi \vee \psi$ iff $prob_{R,r,t} \vDash \phi$ or $prob_{R,r,t} \vDash \psi$.

Example 10. Consider the selection condition $(x.\text{P\_DISEASE} = \text{tuberculosis} \oplus_{in} x.\text{P\_COST} \geq 180)[0.9, 1]$ for the relation PATIENT, denoted by r, in Example 3. With the first tuple $t_1 = (\text{P115, John, 65, tuberculosis}, \{175, 200\})$, where $\wp(t_1) = [0.9, 1]$, one has

$$\begin{aligned}
&prob_{R,r,t_1}(x.\text{P\_DISEASE} = \text{tuberculosis} \oplus_{in} x.\text{P\_COST} \geq 180)\\
&= [0.9 \times prob(\text{tuberculosis} = \text{tuberculosis}), 1 \times prob(\text{tuberculosis} = \text{tuberculosis})]\\
&\oplus_{in}[0.9 \times prob(\{175, 200\} \geq 180), 1 \times prob(\{175, 200\} \geq 180)]\\
&= [0.9 \times 1, 1 \times 1] \oplus_{in} [0.9 \times 0.5, 1 \times 0.5] = [0.9, 1] \oplus_{in} [0.45, 0.5] = [0.945, 1] \subseteq [0.9, 1].
\end{aligned}$$

Consequently, $prob_{R,r,t_1} \vDash (x.\text{P\_DISEASE} = \text{tuberculosis} \oplus_{in} x.\text{P\_COST} \geq 180)[0.9, 1]$. Now, the selection operation on a relation in PRDB is defined as follows.

Definition 12. Let R be a PRDB 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 $\sigma_\phi(r)$, is the relation $r^* = \{t \in r | prob_{R,r,t} \vDash \phi\}$ over R, including all those tuples that satisfy the selection condition ϕ.

Example 11. Let r denote the relation PATIENT in Example 3 and R denote its schema. The query "Find all patients who are not greater than 16 years old with a probability of at least 0.8, and have hepatitis and pay over 6 USD for the daily treatment cost with a probability between 0.3 and 0.6" can be done by the selection operation σ_{ϕ}(PATIENT), where

$$\phi = (x.\text{P\_AGE} \leq 16)[0.8, 1] \wedge (x.\text{P\_DISEASE} = \text{hepatitis} \otimes_{in} x.\text{P\_COST} > 6)[0.3, 0.6].$$

Only the fourth tuple $t_4 = (\text{P442}, Mary, 16, \{\text{hepatitis, cirrhosis}\}, \{7, 8\})$ with $\wp(t_4) = [0.8, 1]$, in Example 3 satisfies ϕ, because

$$\begin{aligned} prob_{R,r,t_4}(x.\text{P\_AGE} \leq 16) &= [0.8 \times prob(16 \leq 16), 1 \times prob(16 \leq 16)] = [0.8 \times 1, 1 \times 1] \\ &= [0.8, 1] \subseteq [0.8, 1] \end{aligned}$$

and $$\begin{aligned} & prob_{R,r,t_4}(x.\text{P\_DISEASE} = \text{hepatitis} \otimes_{in} x.\text{P\_COST} > 6) \\ &= [0.8 \times prob(\{\text{hepatitis, cirrhosis}\} = \text{hepatitis}), 1 \times prob(\{\text{hepatitis, cirrhosis}\} = \\ &\quad \text{hepatitis})] \otimes_{in} [0.8 \times prob(\{7, 8\} > 6), 1 \times prob(\{7, 8\} > 6)] \\ &= [0.8 \times 0.5, 1 \times 0.5] \otimes_{in} [0.8 \times 1, 1 \times 1] = [0.32, 0.5] \subseteq [0.3, 0.6]. \end{aligned}$$

For the other tuples, one has $prob_{R,r,t_i}(x.\text{P\_AGE} \leq 16) = [0, 0] \nsubseteq [0.8, 1]$, $\forall i \neq 4$. Thus, those tuples do not satisfy ϕ.

4.2. Projection

A projection of a PRDB relation on a set of attributes is a new PRDB relation where only the attributes in that set are considered for each tuple of the new relation. Moreover, equivalent tuples under a chosen threshold should be coalesced into a tuple in the result relation by probabilistic combination strategies. The projection operation of a PRDB relation is extended from the projection operation of a CRDB relation with set values and uncertain tuple membership degrees and is defined as follows.

Definition 13. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r be a relation over R and $\boldsymbol{L}$ be a subset of attributes of $\boldsymbol{U}$, $\otimes$ and $\oplus$ be probabilistic disjunction and conjunction strategies with respect to the same combination alternative, $\varepsilon \in [0, 1]$ be an equivalent threshold on $\boldsymbol{L}$. The *projection* of r on $\boldsymbol{L}$ under $\oplus$, $\otimes$ and ε, denoted by $\Pi_{\boldsymbol{L}\oplus\varepsilon\otimes}(r)$, is the probabilistic relation r^* over the schema R^* determined by:

1. $R^* = (\boldsymbol{L}, \wp^*)$, where $\wp^*$ is the mapping from $2^{D1} \times 2^{D2} \times \ldots \times 2^{Dm}$ to the set of all intervals on $[0, 1]$, $m = |\boldsymbol{L}|, D_i$ is the value domain of $A_i \in \boldsymbol{L}, i = 1, ..., m$.

2. $r^* = \{t^* | t^*.A = u.A \cap ... \cap w.A, \wp^*(t^*) = \wp(u) \oplus ... \oplus \wp(w), \forall A \in \boldsymbol{L}, \exists u, ..., w \in r$ such that $u[\boldsymbol{L}] \approx_{\varepsilon\otimes} ... \approx_{\varepsilon\otimes} w[\boldsymbol{L}]\}$.

We note that the combination alternative of a probabilistic combination strategy can be the "ignorance", "independence", "positive correlation" or "mutual exclusion" as in Table 1.

Example 12. Consider the relation DIAGNOSE over the schema **DIAGNOSE**$(\boldsymbol{U}, \wp)$ as in Table 3, where $\boldsymbol{U} = \{\text{P\_ID, D\_ID, P\_AGE, P\_DISEASE}\}$. Then the projection of DIAGNOSE on $\boldsymbol{L} = \{\text{D\_ID, P\_AGE, P\_DISEASE}\}$ under $\oplus_{in}$, $\otimes_{in}$ and the equivalent threshold $\varepsilon = 0.5$ is the relation $r^* = \Pi_{\boldsymbol{L}\oplus_{in}0.5\otimes_{in}}(\text{DIAGNOSE})$ over the schema $R^* = (\boldsymbol{L}, \wp^*)$ computed as in Table 4.

Table 3. Relation DIAGNOSE

P_ID	D_ID	P_AGE	P_DISEASE	$\wp$
P388	D102	60	tuberculosis	[0.9, 1]
P245	D025	{40, 42}	cholecystitis	[1, 1]
P237	D102	60	{lung cancer, tuberculosis}	[0.8, 1]

Table 4. Relation $\Pi_{L\oplus_{in}0.5\otimes_{in}}$(DIAGNOSE)

D_ID	P_AGE	P_DISEASE	$\wp^*$
D102	60	tuberculosis	[0.98, 1]
D025	$\{40, 42\}$	cholecystitis	[1, 1]

We note that two tuples t_1 and t_3 in Table 3 are equivalent on $\boldsymbol{L} = \{$D_ID, P_AGE, P_DISEASE$\}$ under the threshold $\varepsilon = 0.5$ and the independence probabilistic conjunction strategy $\otimes_{in}$ and they are projected on $\boldsymbol{L}$ and coalesced into the tuple t_1 under the independence probabilistic disjunction strategy $\oplus_{in}$ with $\wp^*(t_1) = [0.98, 1]$ in Table 4.

4.3. Cartesian product

For the Cartesian product of two PRDB relations, as in CRDB, we assume the set of attributes of their schemas are disjoint and every k-tuple $t = (v_1, v_2, ..., v_k)$ is an un-ordered list. The Cartesian product of two PRDB relations is extended from the Cartesian product of two CRDB relations as follows.

Definition 14. Let $\boldsymbol{U}_1, \boldsymbol{U}_2$ be two sets of attributes that have not any common element, $R_1 = (\boldsymbol{U}_1, \wp_1)$, $R_2 = (\boldsymbol{U}_2, \wp_2)$ be two PRDB schemas, r_1, r_2 be two relations over R_1 and R_2, respectively and $\otimes$ be a probabilistic conjunction strategy. The *Cartesian product* of r_1 and r_2 under $\otimes$, denoted by $r_1 \times_\otimes r_2$, is the probabilistic relation r over R, determined by:

1. $R = (\boldsymbol{U}, \wp)$, where $\boldsymbol{U} = \boldsymbol{U_1} \cup \boldsymbol{U_2}, \wp$ is the mapping from $2^{D_1} \times 2^{D_2} \times ... \times 2^{D_n}$ to the set of all intervals on $[0, 1], n = |\boldsymbol{U}|$, D_i is the value domain of $A_i \in \boldsymbol{U},\ i = 1, ..., n$.

2. $r = \{t \mid t.A = t_1.A \text{ if } A \in \boldsymbol{U_1},\ t.A = t_2.A \text{ if } A \in \boldsymbol{U_2},\ t_1 \in r_1,\ t_2 \in r_2,\ \wp(t) = \wp_1(t_1) \otimes \wp_2(t_2)\}$.

4.4. Join

The join of two PRDB relations is extended from the natural join of two CRDB relations with probability and set values as following definition.

Definition 15. 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 PRDB schemas, r_1, r_2 be two relations over R_1 and R_2, respectively and $\otimes$ be a probabilistic conjunction strategy. The *natural join* of r_1 and r_2 under $\otimes$, denoted by $r_1 \bowtie_\otimes r_2$, is the probabilistic relation r over the schema R, determined by:

1. $R = (\boldsymbol{U}, \wp)$, where $\boldsymbol{U} = \boldsymbol{U_1} \cup \boldsymbol{U_2}, \wp$ is the mapping from $2^{D_1} \times 2^{D_2} \times ... \times 2^{D_n}$ to the set of all intervals on $[0, 1], n = |\boldsymbol{U}|$, D_i is the value domain of $A_i \in \boldsymbol{U},\ i = 1, ..., n$.

2. $r = \{t|t.A = t_1.A$ if $A \in \boldsymbol{U_1} - \boldsymbol{U_2},\ t.A = t_2.A$ if $A \in \boldsymbol{U_2} - \boldsymbol{U_1},\ t.A = t_1.A \cap t_2.A$ if $A \in \boldsymbol{U_1} \cap \boldsymbol{U_2}$ and $t_1.A \cap t_2.A \neq \varnothing,\ \wp(t) = \wp_1(t_1) \otimes \wp_2(t_2),\ t_1 \in r_1,\ t_2 \in r_2\}$.

Example 13. Given two PRDB relations DOCTOR$_1$ and DOCTOR$_2$ as in Tables 5 and 6. Then, the result of the join of them under the probabilistic conjunction strategy $\otimes_{in}$ is the relation DOCTOR computed as in Table 7.

Table 5. Relation DOCTOR$_1$

D_ID	D_AGE	$\wp_1$
D005	45	$[1,1]$
D093	30	[0.9, 1]
D102	$\{55,56\}$	[0.8, 1]

Table 6. Relation DOCTOR$_2$

D_NAME	D_AGE	$\wp_2$
Alice	$\{30,31\}$	[0.7, 1]
George	52	[1, 1]
Peter	$\{54,55\}$	[0.9, 1]

Table 7. Relation $DOCTOR = DOCTOr_1 \bowtie_{\otimes in}$ DOCTOR$_2$

D_ID	D_NAME	D_AGE	$\wp$
D093	Alice	30	[0.63, 1]
D102	Peter	55	[0.72, 1]

4.5. Intersection, union, and difference

The intersection, union and difference of two PRDB relations over the same schema is a PRDB relation over that schema, where two equivalent tuples under a threshold ε, respectively of those two relations are coalesced into a tuple in the result relation by a probabilistic combination strategy. Thus, the operations are an extension of the intersection, union and difference of two CRDB relations with probability and set values. The intersection, union and difference of two PRDB relations in turn are defined as below.

Definition 16. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r_1 and r_2 be two relations over R, $\otimes$ be a probabilistic conjunction strategy, and $\varepsilon \in [0,1]$ be an equivalent threshold on $\boldsymbol{U}$. The *intersection* of r_1 and r_2 under $\otimes$ and ε, denoted by $r_1 \cap_{\varepsilon\otimes} r_2$, is the probabilistic relation r over R defined by $r = \{t|t.A = t_1.A \cap t_2.A,\ \wp(t) = \wp(t_1) \otimes \wp(t_2),\ t_1 \in r_1,\ t_2 \in r_2, A \in \boldsymbol{U},$ such that $t_1 \approx_{\varepsilon\otimes} t_2$ and $t_1.A \cap t_2.A \neq \varnothing\}$.

Example 14. Consider two relations DIAGNOSE$_1$ and DIAGNOSE$_2$ over the same schema DIAGNOSE$(\boldsymbol{U}, \wp)$ as in Tables 8 and 9, where $\boldsymbol{U}$ = {P_ID, D_ID, P_AGE, P_DISEASE}. Then the intersection of DIAGNOSE$_1$ and DIAGNOSE$_2$ under $\otimes_{in}$ and the equivalent threshold $\varepsilon = 0.25$ is the relation DIAGNOSE computed as in Table 10.

Table 8. Relation DIAGNOSE$_1$

P_ID	D_ID	P_AGE	P_DISEASE	$\wp$
P215	D093	$\{60,62\}$	{lung cancer, tuberculosis}	[1, 1]
P234	D102	$\{40,41\}$	hepatitis	[0.9, 1]

Table 9. Relation DIAGNOSE$_2$

P_ID	D_ID	P_AGE	P_DISEASE	$\wp$
P383	D102	60	lung cancer	[0.9, 1]
P234	D102	$\{41, 42\}$	{hepatitis, gall-stone}	[0.8, 1]
P242	D025	17	cholecystitis	[1, 1]

Table 10. Relation DIAGNOSE = DIAGNOSE$_1\cap_{0.25\otimes_{in}}$DIAGNOSE$_2$

P_ID	D_ID	P_AGE	P_DISEASE	$\wp$
P234	D102	41	hepatitis	[0.72, 1]

We note that the tuple t_2 in Table 8 and the tuple t_2 in Table 9 are equivalent on $\boldsymbol{U} = \{$P_ID, D_ID, P_AGE, P_DISEASE$\}$ under the threshold $\varepsilon = 0.25$ and the independence probabilistic conjunction strategy $\otimes_{in}$, consequently they are coalesced into the tuple t_1 under $\otimes_{in}$ with $\wp(t_1) = [0.72, 1]$ in the Table 10.

Definition 17. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r_1 and r_2 be two relations over R, $\oplus$ and $\otimes$ be probabilistic disjunction and conjunction strategies with respect to the same combination alternative, and $\varepsilon \in [0, 1]$ be an equivalent threshold on $\boldsymbol{U}$. The *union* of r_1 and r_2 under $\otimes$, $\oplus$ and ε, denoted by $r_1 \cup_{\varepsilon\oplus\otimes} r_2$, is the probabilistic relation r over R defined by $r = \{t = t_1 \in r_1|$ there is not any tuple $t_2 \in r_2$ such that $t_1 \approx_{\varepsilon\otimes} t_2,\ \wp(t) = \wp(t_1)\} \cup \{t = t_2 \in r_2|$ there is not any tuple $t_1 \in r_1$ such that $t_2 \approx_{\varepsilon\otimes} t_1,\ \wp(t) = \wp(t_2)\} \cup \{t|t.A = t_1.A \cap t_2.A,\ \wp(t) = \wp(t_1) \oplus \wp(t_2),\ t_1 \in r_1,\ t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1 \approx_{\varepsilon\otimes} t_2,$ and $t_1.A \cap t_2.A \neq \varnothing\}$.

Definition 18. Let $R = (\boldsymbol{U}, \wp)$ be a PRDB schema, r_1 and r_2 be two relations over R, $\ominus$ and $\otimes$ be probabilistic difference and conjunction strategies with respect to the same combination alternative, and $\varepsilon \in [0, 1]$ be an equivalent threshold on $\boldsymbol{U}$. The *difference* of r_1 and r_2 under $\ominus$, $\otimes$ and ε, denoted by $r_1 -_{\varepsilon\ominus\otimes} r_2$, is the probabilistic relation r over R defined by $r = \{t = t_1 \in r_1|$ there is not any tuple $t_2 \in r_2$ such that $t_1 \approx_{\varepsilon\otimes} t_2,\ \wp(t) = \wp(t_1)\} \cup \{t|t.A = t_1.A \cap t_2.A,\ \wp(t) = \wp(t_1) \ominus \wp(t_2),\ t_1 \in r_1,\ t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1 \approx_{\varepsilon\otimes} t_2$ and $t_1.A \cap t_2.A \neq \varnothing\}$.

Example 15. Given two PRDB relations DIAGNOSE$_1$ and DIAGNOSE$_2$ over the same schema **DIAGNOSE**$(\boldsymbol{U}, \wp)$ as in Tables 8 and 9 of Example 14. Then the difference of DIAGNOSE$_1$ and DIAGNOSE$_2$ under $\ominus_{in}$, $\otimes_{in}$ and the equivalent threshold $\varepsilon = 0.25$ is the relation DIAGNOSE computed as in Table 11.

Table 11. Relation DIAGNOSE = DIAGNOSE$_1$ $-_{0.25\ominus\otimes in}$ DIAGNOSE$_2$

P_ID	D_ID	P_AGE	P_DISEASE	$\wp$
P215	D093	$\{60, 62\}$	{lung cancer, tuberculosis}	[1, 1]
P234	D102	41	hepatitis	[0, 0.2]

We note that the tuple t_2 in Table 8 and the tuple t_2 in Table 9 are equivalent on $\boldsymbol{U} = \{$P_ID, D_ID, P_AGE, P_DISEASE$\}$ under the threshold $\varepsilon = 0.25$ and the independence probabilistic conjunction strategy $\otimes_{in}$, consequently they are coalesced into the tuple t_2 under $\ominus_{in}$ with $\wp(t_2) = [0, 0.2]$ in the Table 11.

5. PROPERTY OF PRDB ALGEBRAIC OPERATIONS

In this section, we propose some properties of the PRDB algebraic operations as an extension from those in CRDB. Clearly, these properties say that our PRDB model is coherent and consistent.

Proposition 1. *Let R be a PRDB schema, r be a relation over R, ϕ_1 and ϕ_2 be two selection conditions. Then*

$$\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r) \tag{1}$$

where, the last expression assumes that ϕ_1 and ϕ_2 have the same tuple variable.

Proof. Let $r_1 = \sigma_{\phi_1}(r),\ r_2 = \sigma_{\phi_2}(r)$ and $r_{1\wedge 2} = \sigma_{\phi_1 \wedge \phi_2}(r)$. Then for each $t \in r$, we have

$$\begin{aligned}
\sigma_{\phi_1}(\sigma_{\phi_2}(r)) &= \{t \in r_2 | prob_{R,r_2,t} \models \phi_1\} \\
&= \{t \in r | (prob_{R,r,t} \models \phi_2) \wedge (prob_{R,r_2,t} \models \phi_1)\} \\
&= \{t \in r | (prob_{R,r,t} \models \phi_2) \wedge (prob_{R,r,t} \models \phi_1)\} \text{ (because of } r_2 \subseteq r) \\
&= \{t \in r | prob_{R,r,t} \models \phi_1 \wedge \phi_2\} \text{ (Definition 11)} \\
&= \sigma_{\phi_1 \wedge \phi_2}(r).
\end{aligned}$$

So, $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$ is proven. The equation $\sigma_{\phi 2}(\sigma_{\phi 1}(r)) = \sigma_{\phi_2 \wedge \phi 1}(r)$ is proven similarly. Since $\phi_1 \wedge \phi_2 \Leftrightarrow \phi_2 \wedge \phi_1$ (the logical conjunction of selection conditions are commutative), hence $\sigma_{\phi_1 \wedge \phi_2}(r) = \sigma_{\phi_2 \wedge \phi_1}(r)$. Therefore, we have $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r))$ and so $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$. Thus, Proposition 1 is proven. ■

Proposition 2. *Let R be a PRDB schema, r be a relation over R, $\oplus$ and $\otimes$ be probabilistic disjunction and conjunction strategies with respect to the same combination alternative, $\boldsymbol{A}$ and $\boldsymbol{B}$ be two subsets of attributes of $R, \boldsymbol{A} \subseteq \boldsymbol{B}$ and $\varepsilon \in [0,1]$ be an equivalent threshold on $\boldsymbol{B}$. Then*

$$\Pi_{A \oplus \varepsilon \otimes}(\Pi_{B \oplus \varepsilon \otimes}(r)) = \Pi_{A \oplus \varepsilon \otimes}(r). \tag{2}$$

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 (Definition 13). Moreover, it is due to $\boldsymbol{A} \subseteq \boldsymbol{B}$, so ε-value-equivalent tuples on $\boldsymbol{B}$ are also ε-value-equivalent on $\boldsymbol{A}$ with respect to $\otimes$ (Definition 5). From that, we are easy to see that $\Pi_{A \oplus \varepsilon \otimes}(\Pi_{B \oplus \varepsilon \otimes}(r)) = \Pi_{A \cap B \oplus \varepsilon \otimes}(r) = \Pi_{A \oplus \varepsilon \otimes}(r)$ under the equivalent threshold ε and the same combination alternative of $\oplus$ and $\otimes$. Thus, the equation (2) is proven. ■

Proposition 3. *Let R_1, R_2 and R_3 be the PRDB 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*

$$r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1, \tag{3}$$

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

Equation (3) and (4) say that the join operation of PRDB relations is commutative and associative.

Proof. Clearly, $r_1 \bowtie_{\otimes} r_2$ and $r_2 \bowtie_{\otimes} r_1$ are two relations over the same schema. Since the intersection of sets and the conjunction of probability intervals are commutative. So, by Definition 15, the join of PRDB relations are commutative, we have $r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1$.

By Definition 15, the results of two sides of (4) are the relations over the same schema. Moreover, the intersection of sets and the conjunction of probability intervals have the associativity. From the associativity of the join of classical relations and by Definition 15, it is easy to see that the join of PRDB 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 is a particular case of the join (Definition 14 and Definition 15), we have the straight corollary of the Proposition 3 below.

Corollary 1. *Let R_1, R_2 and R_3 be PRDB schemas such that each pair of them has not any common attribute, 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*

$$r_1 \times_{\otimes} r_2 = r_2 \times_{\otimes} r_1, \tag{5}$$

$$(r_1 \times_{\otimes} r_2) \times_{\otimes} r_3 = r_1 \times_{\otimes} (r_2 \times_{\otimes} r_3). \tag{6}$$

Proposition 4. *Let R be a PRDB schema, r_1, r_2 and r_3 be relations over R, $\otimes$ and $\oplus$ be probabilistic conjunction and disjunction strategies with respect to the same combination alternative, $\varepsilon \in [0, 1]$. Then*

$$r_1 \cap_{\varepsilon\otimes} r_2 = r_2 \cap_{\varepsilon\otimes} r_1, \tag{7}$$

$$(r_1 \cap_{\varepsilon\otimes} r_2) \cap_{\varepsilon\otimes} r_3 = r_1 \cap_{\varepsilon\otimes} (r_2 \cap_{\varepsilon\otimes} r_3), \tag{8}$$

$$r_1 \cup_{\varepsilon\oplus\otimes} r_2 = r_2 \cup_{\varepsilon\oplus\otimes} r_1, \tag{9}$$

$$(r_1 \cup_{\varepsilon\oplus\otimes} r_2) \cup_{\varepsilon\oplus\otimes} r_3 = r_1 \cup_{\varepsilon\oplus\otimes} (r_2 \cup_{\varepsilon\oplus\otimes} r_3). \tag{10}$$

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

Proof. For every equivalent threshold ε chosen, then the equivalent tuples in relations do not change. Moreover, from the commutativity and associativity of the intersection of sets and of the conjunction of probability intervals, by Definition 16, it follows the commutativity and associativity of the intersection of PRDB relations under the equivalent threshold ε and the probabilistic conjunction strategy $\otimes$. Consequently, we have equations (7) and (8).

As for the equations (7) and (8), under an equivalent threshold ε chosen, then the equivalent tuples in relations do not change. From the commutativity and associativity of the intersection of sets and of the conjunction and disjunction of probability intervals, by Definition 17, it follows the union of PRDB relations is commutative and associative under the equivalent threshold ε and the same combination alternative of $\oplus$ and $\otimes$. Thus, we have the equations (9) and (10).

6. CONCLUSION

In this paper, we have proposed a probabilistic relational database model, denoted by PRDB, as a straight extension and generalization of the classical relational database model. As compared to the existing probabilistic relational database models, the uniqueness of our proposed PRDB is that it can represent and handle both uncertain relational tuples associated with probability intervals and imprecise attribute values defined by sets. Computing the set value of attributes and combining the probabilistic membership degrees of tuples in manipulating of the algebraic operations are implemented by the probabilistic interpretations of binary relations on sets and the combination strategies of probability intervals. A notion of the equivalence of relational tuples has been proposed for their coalescence. The data model and basic relational algebraic operations for PRDB have been defined formally and consistently. A set of basic properties of the PRDB algebraic operations has also been proposed as theorems and proven completely.

For a full-fledged model and algebra of PRDB, we are formulating and defining other algebraic operations including *theta join* (the join operation with a general join condition) and *division* ones. Besides, we will extend the properties of the functional dependency and the normalization of relations in CRDB for PRDB. Towards applying PRDB in practice, we will build a management system for PRDB with the familiar querying and manipulating language like SQL that is able to represent and handle uncertain information in the real world.

REFERENCES

[1] P. Bosc, D. Kraft, F. Petry, "Fuzzy sets in database and information systems: status and opportunities", *Journal of Fuzzy Sets and Systems*, vol. 156, pp. 418–426, 2005.

[2] I.I. Ceylan, A. Darwiche, G.V.D Broeck, "Open-world probabilistic databases", *Proceedings of Fifteenth International Conference on Principles of Knowledge Representation and Reasoning*, Cape Town, South Africa, April 25–29, 2016, pp. 339–348.

[3] 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.

[4] C. J. Date, "An Introduction to Database Systems, *8th ed. Addison-Wesley Publishers*, 2008.

[5] A. Dekhtyar, R. Ross, V. S. Subrahmanian "Probabilistic temporal databases, I: algebra", *ACM Transactions on Database Systems*, vol. 26, pp. 41–95, 2001.

[6] D. Dey, S. A. Sarkar, "A probabilistic relational model and algebra", *ACM Transactions on Database Systems*, vol. 21, pp. 339–369, 1996.

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

[8] T. Eiter, J.J. Lu, T. Lukasiewicz, V.S. Subrahmanian, "Probabilistic object bases", *ACM Transactions on Database Systems*, vol. 26, no. 3, pp. 264–312, 2001.

[9] N. Ettouzi, Ph. Leray, M.B. Messaoud, "An exact approach to learning probabilistic relational model", *Proceedings of the 8th Conference on Probabilistic Graphical Models*, Lugano, Switzerland, September 6-9, 2016, pp. 171182.

[10] N. Fuhr, T. Rolleke, "A probabilistic relational algebra for the integration of information retrieval and database systems", *ACM Transactions on Information Systems*, vol. 15, pp. 32–66, 1997.

[11] T. Ge, A. Dekhtyar, J. Goldsmith, "Uncertain data: Representations, query processing, and applications", in *Studies in Fuzziness and Soft Computing*, Springer, 2013, pp. 67–108.

[12] L. V. S. Lakshmanan, N. Leone, R. Ross, V. S. Subrahmanian, "Probview: A flexible probabilistic database system", *ACM Transactions on Database Systems*, vol. 22, pp. 419–469, 1997.

[13] Y. Li, J. Chen, L. Feng, "Dealing with uncertainty: a survey of theories and practices", *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, pp. 2463–2482, 2013.

[14] C. Linda, V. D. Gaag, L. Philippe, "Qualitative probabilistic relational models", *Proceedings of 12th International Conference on Scalable Uncertainty Management*, Milan, Italy, October 3–5, 2018, pp. 276–289.

[15] L. H. Mormille, F. G. Cozman, "Learning probabilistic relational models: A simplified framework, a case study, and a package", *Proceedings of 5th Symposium on Knowledge Discovery, Mining and Learning*, Uberlndia, Minas Gerais, Brazil, October 2–4, 2017, pp.129–136.

[16] Z. Ma, L. Yan, *Advances in probabilistic databases for uncertain information management*, Springer, vol. 304, 2013.

[17] Z. Ma, L. Yan, "Modeling fuzzy data with RDF and fuzzy relational database models", *International Journal of Intelligent Systems*, vol. 33, pp. 1534–1554, 2018.

[18] H. Nguyen, D.H. Tran, "A probabilistic relational data model for uncertain information", *Proceedings of 3rd IEEE International Conference on Information Science and Technology*, Yangzhou, China, March 23–25, 2013, pp. 607–613.

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

[20] H. Nguyen, "A type-2 fuzzy relational database model", *Journal of Information & Communication Technology: Research and Development on Information & Communication Technology*, vol. E3, no. 14, pp. 19–26, 2017.

[21] K. Papaioannou, M. Theobald, M. Bhlen, "Supporting set operations in temporal-probabilistic databases", *Proceedings of the 34th IEEE International Conference on Data Engineering*, Paris, France, April 16–19, 2018, pp. 1180–1191.

[22] R. Ross, V.S. Subrahmanian, "Aggregate operators in probabilistic databases", *Journal of the ACM*, vol. 52, no. 1, pp. 54–101, 2005.

[23] G. Sanfilippo, "Lower and upper probability bounds for some conjunctions of two conditional events", *Proceedings of 12th International Conference on Scalable Uncertainty Management*, Milan, Italy, October 3–5, 2018, pp. 260–275.

[24] D. Suciu, D. Olteanu, C. R, C. Koch, *Probabilistic Databases*, Morgan & Claypool Publishers, 2011.

[25] R. Tang, R. Cheng, H. Wu, S. Bressan, *A Framework for Conditioning Uncertain Relational Data*, Springer-Verlag Berlin Heidelberg, 2012, pp. 71-87.

[26] S. Zhang, C. Zhang, "A probabilistic data model and its semantics", *Journal of Research and Practice in Information Technology*, vol. 35, pp. 237–256, 2003.

[27] W. Zhao, A. Dekhtyar, J. Goldsmith, “Databases for interval probabilities”, *International Journal of Intelligent Systems*, vol. 19, no. 9, pp. 789–815, 2004.

[28] L.A. Zadeh, “Fuzzy sets”, *Journal of Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

Received on June 30, 2019
Revised on August 08, 2019