International Journal of Innovative Computing, Information and Control
Volume **18**, Number **5**, October **2022**

ICIC International ©2022 ISSN 1349-4198
pp. **1477–1492**

EXTENDING PROBABILISTIC RELATIONAL DATABASE MODEL WITH UNCERTAIN MULTIVALUED ATTRIBUTES

Hoa Nguyen[1,2]

[1]Information Technology Faculty
Saigon University
No. 273, An Duong Vuong, District 5, Ho Chi Minh City 72710, Vietnam
nguyenhoa@sgu.edu.vn

[2]Faculty of Information Technology
Industrial University of Ho Chi Minh City
No. 12, Nguyen Van Bao, Ward 4, Go Vap District, Ho Chi Minh City 71408, Vietnam

Received April 2022; revised July 2022

Abstract. *In this paper, we introduce a new relational database model and algebra as an extension of the probabilistic relational database model with uncertain multivalued attributes for representing and handling uncertain information. To develop this new model, the probabilistic interpretation of binary relations on sets is used for computing uncertain degree of functional dependencies, keys and relations on attribute values, the extended probabilistic triples are employed for representing multivalued relational attributes, and the new combination strategies of extended probabilistic triples are defined for building probabilistic relational algebraic operations. A set of the properties of the basic probabilistic relational algebraic operations is also formulated and proven. The new probabilistic relational database model and algebra are coherently and consistently with the classical relational database model and algebra.*

Keywords: Uncertain multivalued attribute, Probabilistic interpretation, Probabilistic triple, Probabilistic relation, Probabilistic relational algebraic operation

1. **Introduction.** Although the classical relational database model (CRDB) is very useful for modeling, designing and implementing large-scale systems [1,2], it is restricted for representing and handling uncertain and imprecise information that are pervasive in the real world [3-5,9-15,21-27,29,30]. For example, applications of the CRDB model can neither deal with queries as "find all patients who are old and have to pay a high treatment cost" nor "find all patients who are at least 95% likely to catch either hepatitis or cirrhosis", etc. Here, "*old*" and "*high*" are vague concepts that can be defined by a fuzzy set [3,4] or a possibility distribution [27], and "hepatitis or cirrhosis" uncertainly expresses a patient's possible diseases that can be represented by the discrete set comprising the two diseases. Meanwhile, "95%" 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.

DOI: 10.24507/ijicic.18.05.1477

Currently, many FRDB models have been built (e.g., [3-6,27]), and a large number of PRDB models have been proposed (e.g., [7-20,23-30]) 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 database models still continue to be developed for modeling data objects of the real world.

PRDB models have been developed from CRDB by two main directions corresponding to two extended levels [26]: 1) at the relation level, each relation is defined by a set of tuples that each tuple is associated with a probability to represent the uncertainty degree of it in the relation; or 2) at the attribute level, each attribute in a relation is associated with a probability to define the uncertainty degree of the values that it may take.

At the relation level, as the works in [7-13,29], each tuple of a relation was associated with a probability in the interval $[0, 1]$ to express the uncertainty membership degree of that tuple for the relation. The uncertainty degree of the attribute values of a tuple was inferred from the uncertainty membership degree of that tuple. However, in many real situations, we do not know exactly the probability as a number in the interval $[0, 1]$ but only can estimate it as an approximate number in a subinterval of $[0, 1]$. The models in [14-17,20,25] were extended with probability intervals associated with each tuple to overcome the shortcoming.

At the attribute level, as in [18,19], each value of an attribute was associated with a probability in the interval $[0, 1]$ to represent the uncertain level for that attribute taking the value. More flexibly, the model in [23] represented the value of each attribute as a probability distribution on a set. It means that each attribute was associated with a set of values and a probability distribution expressing the possibility that the attribute might take one of values of the set with a probability computed from the distribution. The model in [24] extended more the model in [23], where a pair of lower and upper bound probability distributions is used instead of a probability distribution as in [23].

In mentioned probabilistic database models including object-oriented ones [21,22], the attribute of a tuple or an object only took a single, unique value in a set of values with some probability. For instance, the attribute SOIL of a plant in [21] represented by SOIL: $\langle\{loamy, swampy\}, 0.8u, 1.2u\rangle$ said that the suitable soil type for the plant to flourish might be *loamy* or *swampy* with a probability in the interval $[0.4, 0.6]$. However, in practice, a plant may also be conformable both loamy and swampy soil types with a determined probability interval to grow, and then the model in [21] cannot express.

Recently, in [26], the authors introduced a probabilistic relational database model with uncertain multivalued attributes, (called URDB), to overcome the shortcoming of above mentioned models. However, in this model, the probabilistic functional dependency and the relational schema key have not been defined and except the selection operation, other probabilistic relational algebraic operations have not been built. Thus, the ability of representing and dealing with uncertain information of URDB has been limited in the real world applications.

In this paper, we define notions of the probabilistic functional dependency and the relational schema key for URDB and extend it with a full set of basic probabilistic relational algebraic operations. Some properties of URDB algebraic operations are also proposed, formulated and proven. This new extension for URDB is also consistent with the classical relational database model in [1,2] and the decision making support system in [28], where tuples and objects can have multivalued attributes, and the heterogeneous nonlinear non-affine multi-agent system in [31], where control directions can be uncertain.

Basic probability definitions as a mathematical base for URDB are presented in Section 2. The URDB model including fundamental concepts as the schema, relation, database,

probabilistic functional dependency and the relational schema key is introduced in Section 3. Section 4 presents probabilistic relational algebraic operations and their properties. Finally, Section 5 concludes the paper and outlines further research directions in the future.

2. Probability and Probabilistic Combination Strategies.

The URDB model is developed on a probability base including probability definitions and probabilistic combination strategies for representing and handling uncertain information.

2.1. Probability distribution functions and probabilistic triples.

For expressing uncertain information in URDB, we use probability distribution functions and probabilistic triples over a set in [21,24]. More, probabilistic triples over a set are extended to probabilistic triples over a set of sets for representing multivalued attributes as in [26]. Probability distribution functions and extended probabilistic triples are defined as below.

Definition 2.1. *Let V be a finite set, a probability distribution function α over V is a mapping α: $V \to [0,1]$ such that $\Sigma_{x\in V}\alpha(x) \leq 1$.*

An important probability distribution function which we often encounter in practice is the uniform distribution $u(x) = 1/|V|$, $\forall x \in V$.

Definition 2.2. *Let S be a finite set, a probabilistic triple $\langle V, \alpha, \beta\rangle$ over S consists of a subset V of the set 2^S (i.e., the set of all subsets of S) whose elements are disjointed, a probability distribution function α: $V \to [0,1]$, and a function β: $V \to [0,1]$ such that $\alpha(x) \leq \beta(x)$, $\forall x \in V$ and $\Sigma_{x\in V}\beta(x) \geq 1$ hold.*

Informally, a probabilistic triple $\langle V, \alpha, \beta\rangle$ assigns each $x \in V$ a probability interval $[\alpha(x), \beta(x)]$ to express the uncertainty degree of x in V. This assignment is consistent in the sense that each $x \in V$ is assigned a probability $p(x) \in [\alpha(x), \beta(x)]$ such that $\Sigma_{x\in V}p(x) = 1$.

Example 2.1. *When examining a patient, a doctor may be unsure about what disease the patient is suffered from. However, if the doctor is sure that the patient's diseases are duodenitis and gastroenteritis or dyspepsia with a probability between 60% and 80%, then this knowledge may be encoded by the extended probabilistic triple $\langle\{\{$duodenitis, gastroenteritis$\}, \{$dyspepsia$\}\}, 1.2u, 1.6u\rangle$. Here, u is the uniform distribution function over $\{\{$duodenitis, gastroenteritis$\}, \{$dyspepsia$\}\}$, $1.2u$ and $1.6u$ are probability distribution functions α and β respectively with $\alpha(x) = 1.2u(x) = 1.2(1/2) = 0.6$ and $\beta(x) = 1.6u(x) = 1.6(1/2) = 0.8$, $\forall x \in \{\{$duodenitis, gastroenteritis$\}, \{$dyspepsia$\}\}$.*

We note that an element e in S is also considered as a special set $\{e\}$ on S; thus a probabilistic triple $\langle\{\{e_1\}, \{e_2\}, \ldots, \{e_k\}\}, \alpha, \beta\rangle$ can be written as $\langle\{e_1, e_2, \ldots, e_k\}, \alpha, \beta\rangle$ for simplicity. Also, "an extended probabilistic triple" is called "a probabilistic triple".

2.2. Probabilistic interpretation of binary relations on sets.

For computing uncertain degree of binary relations on attribute values in URDB, the probabilistic interpretation of binary relations on sets in [26] is extended from [25] as follows.

Definition 2.3. *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 $u \in B$.

2.3. **Probabilistic combination strategies.** In this work, we employ the combination strategies of probability intervals in [21,26] to compute the probability intervals of the conjunction, disjunction or difference event of two events. 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* as follows:

1) Independence conjunction, disjunction and difference strategies, denoted $\otimes_{in}$, $\oplus_{in}$, and $\ominus_{in}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{in} [L_2, U_2] \equiv [L_1.L_2, U_1.U_2]$
   - $[L_1, U_1] \oplus_{in} [L_2, U_2] \equiv [L_1 + L_2 - (L_1.L_2), U_1 + U_2 - (U_1.U_2)]$
   - $[L_1, U_1] \ominus_{in} [L_2, U_2] \equiv [L_1.(1 - U_2), U_1.(1 - L_2)]$
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
   - $[L_1, U_1] \otimes_{me} [L_2, U_2] \equiv [0, 0]$
   - $[L_1, U_1] \oplus_{me} [L_2, U_2] \equiv [\min(1, L_1 + L_2), \min(1, U_1 + U_2)]$
   - $[L_1, U_1] \ominus_{me} [L_2, U_2] \equiv [L_1, \min(U_1, 1 - L_2)]$
3) 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
   - $[L_1, U_1] \otimes_{pc} [L_2, U_2] \equiv [\min(L_1, L_2), \min(U_1, U_2)]$
   - $[L_1, U_1] \oplus_{pc} [L_2, U_2] \equiv [\max(L_1, L_2), \max(U_1, U_2)]$
   - $[L_1, U_1] \ominus_{pc} [L_2, U_2] \equiv [\max(0, L_1 - U_2), \max(0, U_1 - L_2)]$
4) Ignorance conjunction, disjunction and difference strategies, denoted $\otimes_{ig}$, $\oplus_{ig}$, and $\ominus_{ig}$ respectively, are determined by
   - $[L_1, U_1] \otimes_{ig} [L_2, U_2] \equiv [\max(0, L_1 + L_2 - 1), \min(U_1, U_2)]$
   - $[L_1, U_1] \oplus_{ig} [L_2, U_2] \equiv [\max(L_1, L_2), \min(1, U_1 + U_2)]$
   - $[L_1, U_1] \ominus_{ig} [L_2, U_2] \equiv [\max(0, L_1 - U_2), \min(U_1, 1 - L_2)]$

In the following sections, the notation $[L_1, U_1] \leq [L_2, U_2]$ is used to denote $L_1 \leq L_2$ and $U_1 \leq U_2$ whereas the notation $[L_1, U_1] \subseteq [L_2, U_2]$ is for $L_2 \leq L_1$ and $U_1 \leq U_2$. Also, a single probability value p can be treated as the probability interval $[p, p]$.

2.4. **Conjunction, disjunction and difference of probabilistic triples.** For building probabilistic relational algebraic operations in URDB such as the projection, join, intersection, union and difference, we extend the conjunction, disjunction and difference of probabilistic triples in [21,24] to new ones of extended probabilistic triples as the basis for combining the probability of uncertain multivalued attribute values in outcome relations of these algebraic operations. First, the conjunction of extended probabilistic triples is defined as follows.

Definition 2.4. *Let $pt_1 = \langle V_1, \alpha_1, \beta_1 \rangle$ and $pt_2 = \langle V_2, \alpha_2, \beta_2 \rangle$ be two probabilistic triples, and $\otimes$ be a probabilistic conjunction strategy. The conjunction of pt_1 and pt_2 under $\otimes$, denoted by $pt_1 \otimes pt_2$, is the probabilistic triple $pt = \langle V, \alpha, \beta \rangle$, such that*

1) $V = \{v = v_1 \cap v_2 | v_1 \in V_1, v_2 \in V_2$ *and* $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)] \neq [0, 0]\}$, *and*

2) $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)]$, $\forall v = v_1 \cap v_2 \in V$, $v_1 \in V_1$, $v_2 \in V_2$.

Example 2.2. *Let* $pt_1 = \langle\{\{hepatitis, cholecystitis\}, 0.8u, 1.4u\rangle$ *and* $pt_2 = \langle\{\{\{hepatitis, cirrhosis\}\}, u, u\rangle$ *be probabilistic triples, and then* $pt_1 \otimes_{in} pt_2$ *under the independence probabilistic conjunction strategy is the probabilistic triple* $pt = \langle\{hepatitis\}, 0.4u, 0.7u\rangle$.

Next, the disjunction and difference of probabilistic triples in turn are defined as below.

Definition 2.5. *Let* $pt_1 = \langle V_1, \alpha_1, \beta_1\rangle$ *and* $pt_2 = \langle V_2, \alpha_2, \beta_2\rangle$ *be two probabilistic triples, and* $\oplus$ *be a probabilistic disjunction strategy. The disjunction of* pt_1 *and* pt_2 *under* $\oplus$*, denoted by* $pt_1 \oplus pt_2$*, is the probabilistic triple* $pt = \langle V, \alpha, \beta\rangle$*, such that*

1) $V = H \cup Q \cup T$*, where* $H = \{v_1 \in V_1 | \neg\exists v_2 \in V_2, v_1 \cap v_2 \neq \varnothing\}$, $Q = \{v_2 \in V_2 | \neg\exists v_1 \in V_1, v_1 \cap v_2 \neq \varnothing\}$, $T = \{v = v_1 \cup v_2 | v_1 \in V_1, v_2 \in V_2, v_1 \cap v_2 \neq \varnothing\}$*, and*

2)
$$[\alpha(v), \beta(v)] = \begin{cases} [\alpha_1(v), \beta_1(v)], & \forall v \in H, \\ [\alpha_2(v), \beta_2(v)], & \forall v \in Q, \\ [\alpha_1(v_1), \beta_1(v_1)] \oplus [\alpha_2(v_2), \beta_2(v_2)], & \forall v = v_1 \cup v_2 \in T, \\ & v_1 \in V_1, v_2 \in V_2. \end{cases}$$

Example 2.3. *Let* $pt_1 = \langle\{\{hepatitis, cirrhosis\}, cholecystitis\}, 0.4u, 1.2u\rangle$ *and* $pt_2 = \langle\{\{\{hepatitis, cirrhosis\}, pancreatitis\}, 0.6u, 1.3u\rangle$ *be probabilistic triples, then* $pt_1 \oplus_{in} pt_2$ *under the independence probabilistic disjunction strategy is the probabilistic triple* $pt = \langle\{cholecystitis, pancreatitis, \{hepatitis, cirrhosis\}\}, \alpha, \beta\rangle$*, where* $\alpha(cholecystitis) = 0.2$, $\beta(cholecystitis) = 0.6$, $\alpha(pancreatitis) = 0.3$, $\beta(pancreatitis) = 0.65$, $\alpha(\{hepatitis, cirrhosis\}) = 0.44$, $\beta(\{hepatitis, cirrhosis\}) = 0.86$.

Definition 2.6. *Let* $pt_1 = \langle V_1, \alpha_1, \beta_1\rangle$ *and* $pt_2 = \langle V_2, \alpha_2, \beta_2\rangle$ *be two probabilistic triples, and* $\ominus$ *be a probabilistic difference strategy. The difference of* pt_1 *and* pt_2 *under* $\ominus$*, denoted by* $pt_1 \ominus pt_2$*, is the probabilistic triple* $pt = \langle V, \alpha, \beta\rangle$*, such that*

1) $V = H \cup T$*, where* $H = \{v_1 \in V_1 | \neg\exists v_2 \in V_2, v_1 \cap v_2 \neq \varnothing\}$, $T = \{v_1 \in V_1 | \exists v_2 \in V_2, v_1 \cap v_2 \neq \varnothing$ *and* $[\alpha_1(v_1), \beta_1(v_1)] \ominus [\alpha_2(v_2), \beta_2(v_2)] \neq [0, 0]\}$*, and*

2)
$$[\alpha(v), \beta(v)] = \begin{cases} [\alpha_1(v), \beta_1(v)], & \forall v \in H, \\ [\alpha_1(v_1), \beta_1(v_1)] \ominus [\alpha_2(v_2), \beta_2(v_2)], & \forall v = v_1 \in T, v_1 \in V_1, \\ & v_2 \in V_2, v_1 \cap v_2 \neq \varnothing. \end{cases}$$

Example 2.4. *Let* pt_1 *and* pt_2 *be probabilistic triples given as in Example 2.3, and then* $pt_1 \ominus_{in} pt_2$ *under the independence probabilistic difference strategy is the probabilistic triple* $pt = \langle\{cholecystitis, \{hepatitis, cirrhosis\}\}, \alpha, \beta\rangle$*, where* $\alpha(cholecystitis) = 0.2$, $\beta(cholecystitis) = 0.6$, $\alpha(\{hepatitis, cirrhosis\}) = 0.07$, $\beta(\{hepatitis, cirrhosis\}) = 0.42$.

Now, the probabilistic definitions in Section 2 above are used to build the URDB model and the probabilistic relational algebraic operations of it in Sections 3 and 4. The probabilistic triples are employed for representing the URDB model (i.e., probabilistic schemas and relations) while the probabilistic interpretation of binary relations on sets, conjunction, disjunction and difference of probabilistic triples and probabilistic combination strategies are for defining probabilistic relational algebraic operations on the model.

3. **URDB Model.** As CRDB model, URDB model is a structure with fundamental concepts, such as the schema, relation and database to represent data and the relationship between them. URDB model is extended from the model in [26] with the probabilistic functional dependency and relational schema key.

3.1. **URDB schemas.** A URDB schema consists of a set of relational attributes respectively associated with domains defining probabilistic triples representing uncertain values of those attributes. The URDB schema is extended from that of the model [24] with uncertain multivalued attributes (cf. [26]) as follows.

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

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

2) $\wp$ is a function that maps each attribute $A \in \boldsymbol{U}$ to the set of all probabilistic triples on the value domain of A.

For simplicity, the notation $R(\boldsymbol{U}, \wp)$ and then R can be used to denote $R = (\boldsymbol{U}, \wp)$.

3.2. **URDB relations.** A URDB relation is an instance of a URDB schema, where each relational attribute may take more than one uncertain value represented by a probabilistic triple. The URDB relation is extended from that of the model in [24] with uncertain multivalued attributes (cf. [26]) as the following definition.

Definition 3.2. *Let $\boldsymbol{U} = \{A_1, A_2, \ldots, A_k\}$ be a set of k pairwise different attributes. A URDB relation r over the schema $R(\boldsymbol{U}, \wp)$ is a finite set of elements $\{t_1, t_2, \ldots, t_n\}$, where each element $t_i = (\langle V_{i1}, \alpha_{i1}, \beta_{i1}\rangle, \langle V_{i2}, \alpha_{i2}, \beta_{i2}\rangle, \ldots, \langle V_{ik}, \alpha_{ik}, \beta_{ik}\rangle)$ is a list of k probabilistic triples such that $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ belongs to the set $\wp(A_j)$ and $V_{ij} \neq \varnothing$, for every $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$.*

Each element t in the relation r over $R(\boldsymbol{U}, \wp)$ is called a tuple on $\boldsymbol{U}$. For each tuple t_i, the probabilistic triple $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ represents an uncertain valued set of the attribute A_j of the tuple t_i. We write $t_i.A_j$ or $t_i[A_j]$ to denote $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ and $[t_i]$ to replace $(V_{i1}, V_{i2}, \ldots, V_{ik})$.

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

Example 3.1. *In the database about patients at the clinic of a hospital, a simple URDB relation, named PATIENT, over the URDB schema* ***PATIENT***$(\{$*P_ID, P_NAME, P_AGE, P_DISEASE, D_COST*$\}, \wp)$ *can be given as Table 1.*

Table 1. Relation PATIENT

P_ID	P_NAME	P_AGE	P_DISEASE	D_COST
P165	John	$\langle\{55\}, u, u\rangle$	$\langle\{$lung cancer, tuberculosis$\}$, $0.6u, 1.2u\rangle$	$\langle\{30, 35\}, 0.7u, 1.3u\rangle$
P224	Paul	$\langle\{47, 48\}, u, u\rangle$	$\langle\{\{$hepatitis, cirrhosis$\}$, $\{$cholecystitis$\}\}, 0.9u, 1.3u\rangle$	$\langle\{6, 7\}, 0.8u, 1.4u\rangle$
P336	Ann	$\langle\{15\}, u, u\rangle$	$\langle\{$dyspepsia, cholelithiasis$\}$, $08u, 1.4u\rangle$	$\langle\{7\}, u, u\rangle$
P448	Selena	$\langle\{52\}, u, u\rangle$	$\langle\{\{$duodenitis, gastroenteritis$\}$, $\{$dyspepsia$\}\}, 1.2u, 1.6u\rangle$	$\langle\{7, 8\}, 0.8u, 1.2u\rangle$
P512	Helen	$\langle\{42, 43\}, u, u\rangle$	$\langle\{$dyspepsia, cholelithiasis$\}$, $0.6u, 1.2u\rangle$	$\langle\{7\}, u, u\rangle$

In the relation, the attributes P_ID, P_NAME, P_AGE, P_DISEASE and D_COST describe the information about the identifier, name, age, disease and daily treatment cost of each patient, respectively. In reality, while diagnosing, the disease of each patient is not always determined certainly by the doctors. Similarly, the daily treatment cost for patients is also not known definitely even as the patients know about their diseases. Here,

u is the uniform distribution function as presented in Definition 2.1, each tuple in the relation PATIENT represents uncertain information of a patient, for instance, the tuple t_1 (the first tuple) in this relation expresses that the clinic has a patient whose identifier, name and age are P165, John and 55, respectively. The patient's disease is lung cancer or tuberculosis with a probability between 0.3 and 0.6 and the daily treatment cost is 30 or 35 (USD) with a probability between 0.35 and 0.65. Note that, for each attribute A in the schema **PATIENT**, $\wp(A)$ includes all probabilistic triples on the domain of A (Definition 3.1). In addition, for simplicity, each probabilistic triple $\langle V, u, u\rangle$, where $V = \{v\}$, v is not a set, will be represented as a single value v (such as probabilistic triples for the attribute P_ID). Because if an attribute takes such a probabilistic triple, then, actually it only takes a value v with the probability as 1 (Definition 2.2). In other words, the attribute certainly takes the value v.

The URDB database is defined as an extension of CRDB and the probabilistic relational database in [24] with uncertain multivalued attributes as follows.

Definition 3.3. *A URDB database over a set of attributes is a set of URDB relations corresponding to the set of their URDB schemas.*

3.3. **URDB functional dependencies.** Functional dependencies play an important role in CRDB. The functional dependent concept in URDB is extended from that in [24] with uncertain multivalued attributes. We first define the probability measure to determine the equal degree of two values of the same attribute for two different tuples in a URDB relation as follows.

Definition 3.4. *Let $R(\boldsymbol{U}, \wp)$ be a URDB schema, r be a relation over $R(\boldsymbol{U}, \wp)$ 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 $[\Sigma_{v\in V}\alpha(v).Pr(v_1 = v_2), \min(1, \Sigma_{v\in V}\beta(v).Pr(v_1 = v_2))]$, where $t_1.A = \langle V_1, \alpha_1, \beta_1\rangle$, $t_2.A = \langle V_2, \alpha_2, \beta_2\rangle$ and $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)]$, $\forall v = (v_1, v_2) \in V = V_1 \times V_2$.*

Now, the probabilistic functional dependency in URDB as an extension of the functional dependency in [24] with uncertain multivalued attributes and is defined as below.

Definition 3.5. *Let $R = (\boldsymbol{U}, \wp)$ be a URDB schema, r be any relation over R, $\otimes$ be a probabilistic conjunction strategy, $\boldsymbol{X} = \{A_i, \ldots, A_l\}$ and $\boldsymbol{Y} = \{A_j, \ldots, A_m\}$ be two subsets of $\boldsymbol{U}$. A URDB functional dependency of $\boldsymbol{Y}$ on $\boldsymbol{X}$ under $\otimes$, denoted by $\boldsymbol{X} \to_\otimes \boldsymbol{Y}$, holds if and only if*

$$\forall t_1, t_2 \in r,\ p(t_1[\boldsymbol{X}] =_\otimes t_2[\boldsymbol{X}]) \le p(t_1[\boldsymbol{Y}] =_\otimes t_2[\boldsymbol{Y}]),$$

where $p(t_1[\boldsymbol{X}] =_\otimes t_2[\boldsymbol{X}]) = p(t_1.A_i =_\otimes t_2.A_i) \otimes \cdots \otimes p(t_1.A_l =_\otimes t_2.A_l)$ and $p(t_1[Y] =_\otimes t_2[Y]) = p(t_1.A_j =_\otimes t_2.A_j) \otimes \cdots \otimes p(t_1.A_m =_\otimes t_2.A_m)$.

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

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

As in the classical relational database, the keys of a schema in URDB are the basis for recognizing a tuple of a probabilistic relation. In the model and management systems of the classical relational database, key attributes are constrained not to take the value NULL [1,2]. Similarly, in URDB, we assume that the value of each key attribute is always certain and definite. The key concept of URDB schemas is defined using the probabilistic functional dependency as follows.

Definition 3.6. *Let $R(\boldsymbol{U}, \wp)$ be a URDB schema, r be any relation over R and $\otimes$ be a probabilistic conjunction strategy, a set of attributes $K \subseteq \boldsymbol{U}$ is called a key of R under $\otimes$ if the value of each attribute of K is always certain in r 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 3.3. *In the relation PATIENT above, if we assume that each patient has a unique identifier corresponding to the value of the attribute P_ID, then P_ID is a key of the schema **PATIENT** under all probabilistic conjunction strategies.*

4. **URDB Algebra.** As the CRDB algebra [1,2], the URDB algebra is a set of basic operations to manipulate, handle and query data. In [26], a URDB algebra was presented. However, that algebra only consisted of a defined selection operation, meanwhile other operations such as the projection, Cartesian product, join, intersection, union and difference are missing. In this work, we extend the URDB algebra in [26] to a new URDB algebra with a full set of basic relational algebraic operations taking account of uncertain multivalued attributes to manipulate, handle and query uncertain information in practice.

4.1. **Selection.** The selection operation in URDB is extended from that in [24] (cf. [26]) to allow querying with uncertain multivalued relational attributes. Before presenting the selection operation, the selection expressions and conditions in turn are defined as below.

Definition 4.1. *Let R be a URDB 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 $\{=, \neq, \leq, \geq, <, >, \subseteq, \supseteq\}$, and c is a single value or a set of values.*
2) *$x.A_1 =_{\otimes} x.A_2$, where $x \in X$, A_1 and A_2 are two different attributes in R, and $\otimes$ is a probabilistic conjunction strategy.*
3) *$E_1 \otimes E_2$, where E_1 and E_2 are selection expressions on the same relational 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 relational tuple variable, and $\oplus$ is a probabilistic disjunction strategy.*

Definition 4.2. *Let R be a URDB 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 4.1. *Given the schema **PATIENT** in Example 3.1, the selection of "all patients who are over 50 years old with a probability of at least 0.9 or have lung cancer and pay the daily treatment cost not less than 30 USD with a probability between 0.6 and 0.7" can be done using the selection condition $(x.P\_AGE > 50)[0.9, 1] \vee (x.P\_DISEASE = \text{lung cancer} \otimes x.D\_COST \geq 30)[0.6, 0.7]$, where $x.P\_AGE > 50$ and $x.P\_DISEASE = \text{lung cancer} \otimes x.D\_COST \geq 30$ are selection expressions under Definition 4.1.*

The probabilistic interpretation (as a probabilistic measure) of selection expressions in URDB (cf. [26]) is extended from that in [24] with the probabilistic interpretation of binary relations on sets as follows.

Definition 4.3. *Let R be a URDB 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) $Prob_{R,r,t}(x.A\,\theta\,c) = [\Sigma_{v\in V}\alpha(v).Pr(v\,\theta\,c), \min(1, \Sigma_{v\in V}\beta(v).Pr(v\,\theta\,c))]$, *where* $t.A = \langle V, \alpha, \beta\rangle$.
2) $Prob_{R,r,t}(x.A_1 =_{\otimes} x.A_2) = [\Sigma_{v\in V}\alpha(v).Pr(v_1 = v_2), \min(1, \Sigma_{v\in V}\beta(v).Pr(v_1 = v_2))]$, *where* $t.A_1 = \langle V_1, \alpha_1, \beta_1\rangle$, $t.A_2 = \langle V_2, \alpha_2, \beta_2\rangle$ *and* $[\alpha(v), \beta(v)] = [\alpha_1(v_1), \beta_1(v_1)] \otimes [\alpha_2(v_2), \beta_2(v_2)]$, $\forall v = (v_1, v_2) \in V = V_1 \times V_2$.
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 $v\,\theta\,c$, while $Prob_{R,r,t}(x.A_1 =_{\otimes} x.A_2)$ 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 4.2. *Let R denote the schema **PATIENT** and r denote the relation PATIENT in Example 3.1. Consider the fourth tuple in r, denoted by t_4. We have*

$$
\begin{aligned}
& Prob_{R,r,t_4}(x.P\_DISEASE \supseteq \{duodenitis, gastroenteritis\}) \\
= \; & [1.2u(\{duodenitis, gastroenteritis\}).Pr(\{duodenitis, gastroenteritis\} \\
& \supseteq \{duodenitis, gastroenteritis\}) + 1.2u(\{dyspepsia\}).Pr(\{dyspepsia\} \\
& \supseteq \{duodenitis, gastroenteritis\}), \\
& \min(1, 1.6u(\{duodenitis, gastroenteritis\}).Pr(\{duodenitis, gastroenteritis\} \\
& \supseteq \{duodenitis, gastroenteritis\}) + 1.6u(\{dyspepsia\}).Pr(\{dyspepsia\} \\
& \supseteq \{duodenitis, gastroenteritis\}))] \\
= \; & [1.2 \times 0.5 \times 1.0 + 1.2 \times 0.5 \times 0.0, \min(1, 1.6 \times 0.5 \times 1.0 + 1.6 \times 0.5 \times 0.0)] \\
= \; & [0.6, 0.8].
\end{aligned}
$$

The satisfaction (i.e., semantics) of selection conditions in URDB (cf. [26]) is defined as below.

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

1) $Prob_{R,r,t} \models (E)[L,U]$ *if and only if (iff)* $Prob_{R,r,t}(E) \subseteq [L,U]$.
2) $Prob_{R,r,t} \models \neg\phi$ *iff* $Prob_{R,r,t} \models \phi$ *does not hold.*
3) $Prob_{R,r,t} \models \phi \wedge \psi$ *iff* $Prob_{R,r,t} \models \phi$ *and* $Prob_{R,r,t} \models \psi$.
4) $Prob_{R,r,t} \models \phi \vee \psi$ *iff* $Prob_{R,r,t} \models \phi$ *or* $Prob_{R,r,t} \models \psi$.

Now, the notion of the satisfaction of selection condition is use to define the selection operation on a URDB relation as follows (cf. [26]).

Definition 4.5. *Let R be a URDB 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 $r^* = \{t \in r | Prob_{R,r,t} \models \phi\}$ over R, including all satisfied tuples of the selection condition ϕ.*

Example 4.3. *Let r denote the relation PATIENT in Example 3.1 and R denote its schema. The query "Find all patients who are over 50 years old with a probability of*

at least 0.9, have both duodenitis and gastroenteritis 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 $\sigma_\phi(PATIENT)$, *where* $\phi = (x.P\_AGE > 50)[0.9, 1] \wedge (x.P\_DISEASE \supseteq \{duodenitis, gastroenteritis\} \otimes_{in} x.D\_COST \geq 7)[0.4, 0.8]$.

Only the fourth tuple t_4 *of the relation PATIENT in Example 3.1 satisfies* ϕ, *because*

$Prob_{R,r,t_4}(x.P\_AGE > 50) = [u(52) \times Pr(52 > 50), \min(1, u(52) \times Pr(52 > 50))] = [1 \times 1, \min(1, 1 \times 1)] = [1, 1] \subseteq [0.9, 1]$,

$Prob_{R,r,t_4}(x.D\_COST \geq 7) = [0.8u(7) \times Pr(7 \geq 7) + 0.8u(8) \times Pr(8 \geq 7), \min(1, 1.2u(7) \times Pr(7 \geq 7) + 1.2u(8) \times Pr(8 \geq 7))] = [0.8 \times 0.5 \times 1 + 0.8 \times 0.5 \times 1, \min(1, 1.2 \times 0.5 \times 1 + 1.2 \times 0.5 \times 1)] = [0.8, 1]$.

From the result of the computation in Example 4.2, we have

$Prob_{R,r,t_4}(x.P\_DISEASE \supseteq \{duodenitis, gastroenteritis\} \otimes_{in} x.D\_COST \geq 7) = [0.6, 0.8] \otimes_{in} [0.8, 1] = [0.48, 0.8] \subseteq [0.4, 0.8]$.

For the other tuples, one has $Prob_{R,r,t_i}(x.P\_DISEASE \supseteq \{duodenitis, gastroenteritis\} \otimes_{in} x.D\_COST \geq 7) = [0, 0] \not\subset [0.4, 0.8]$, $\forall i \neq 4$.

Thus, the result of the query is the relation $\sigma_\phi(PATIENT)$ *that consists of one tuple* $t_4 = (P448, Selena, \langle\{52\}, u, u\rangle, \langle\{\{duodenitis, gastroenteritis\}, \{dyspepsia\}\}, 1.2u, 1.6u\rangle, \langle\{7, 8\}, 0.8u, 1.2u\rangle)$.

As presented above, except the selection, other basic algebraic operations as the projection, Cartesian product, join, intersection, union and difference have not been built for URDB in [26]. In the next sections, we develop a complete algebra for URDB as an extension of the probabilistic relational algebra in [24] with uncertain multivalued attributes.

4.2. **Projection.** A projection of a URDB relation on a set of attributes is a new URDB relation computed similarly to the projection of a CRDB relation. However, since the value of relational attributes may be uncertain, the projected tuples that have the same valued set should be coalesced into a tuple in the result relation by a probabilistic combination strategy. The projection operation of a URDB relation is defined as follows.

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

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

Example 4.4. *Consider the relation PATIENT over the schema* $\boldsymbol{PATIENT}(\{P\_ID, P\_NAME, P\_AGE, P\_DISEASE, D\_COST\}, \wp)$ *as in Table 1, then the projection of it on the set of the attributes* $\boldsymbol{L} = \{P\_DISEASE, D\_COST\}$ *under* $\oplus_{in}$ *is the relation* $\Pi_{\{P\_DISEASE, D\_COST\}\oplus_{in}}(PATIENT)$ *over the schema* $R^*(\{P\_DISEASE, D\_COST\}, \wp^*)$ *computed as in Table 2, where* $\wp^*(A) = \wp(A)$, $\forall A \in \boldsymbol{L}$.

Note that in the relation PATIENT, we have $[t_3[\boldsymbol{L}]] = [t_5[\boldsymbol{L}]]$; *thus two tuples* t_3 *and* t_5 *are projected on* $\boldsymbol{L}$ *and coalesced into the tuple* t_4 *under the independence probabilistic disjunction strategy* $\oplus_{in}$ *in Table 2.*

4.3. **Cartesian product.** For the Cartesian product of two URDB relations, as in CRDB, we assume the set of attributes of their schemas is disjoint and every k-tuple $t = (\langle V_1, \alpha_1, \beta_1\rangle, \ldots, \langle V_k, \alpha_k, \beta_k\rangle)$ is an un-ordered list. The Cartesian product of two URDB relations is extended from the Cartesian product of two CRDB relations with uncertain multivalued attributes as follows.

TABLE 2. Relation $\Pi_{\{P\_DISEASE,D\_COST\}\oplus_{in}}$(PATIENT)

P_DISEASE	D_COST
$\langle\{\text{lung cancer, tuberculosis}\}, 0.6u, 1.2u\rangle$	$\langle\{30, 35\}, 0.7u, 1.3u\rangle$
$\langle\{\{\text{hepatitis, cirrhosis}\}, \{\text{cholecystitis}\}\}, 0.9u, 1.3u\rangle$	$\langle\{6, 7\}, 0.8u, 1.4u\rangle$
$\langle\{\{\text{duodenitis, gastroenteritis}\}, \{\text{dyspepsia}\}\}, 1.2u, 1.6u\rangle$	$\langle\{7, 8\}, 0.8u, 1.2u\rangle$
$\langle\{\text{dyspepsia, cholelithiasis}\}, 1.16u, 1.76u\rangle$	$\langle\{7\}, u, u\rangle$

Definition 4.7. *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 URDB schemas, and r_1, r_2 be two relations over R_1 and R_2, respectively. The Cartesian product of r_1 and r_2, denoted by $r_1 \times r_2$, is the relation r over R, determined by*

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

4.4. **Join.** The join of two URDB relations is extended from the natural join of two probabilistic relations in [24] with uncertain multivalued attributes as following definition.

Definition 4.8. *Let $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ be two sets of attributes such that if they have the same name attributes, respectively in those two sets then such attributes have the same value domain. Let $R_1(\boldsymbol{U}_1, \wp_1)$ and $R_2(\boldsymbol{U}_2, \wp_2)$ be two URDB schemas, r_1, 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) *$R = (\boldsymbol{U}, \wp)$ where $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2$, $\wp(A) = \wp_1(A)$ if $A \in \boldsymbol{U}_1 - \boldsymbol{U}_2$, $\wp(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_2 - \boldsymbol{U}_1$ and $\wp(A) = \wp_1(A) = \wp_2(A)$ if $A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$.*
2) *$r = \{t | t.A = t_1.A$ if $A \in \boldsymbol{U}_1 - \boldsymbol{U}_2, t.A = t_2.A$ if $A \in \boldsymbol{U}_2 - \boldsymbol{U}_1, t.A = t_1.A \otimes t_2.A$ if $A \in \boldsymbol{U}_1 \cap \boldsymbol{U}_2$ and $t_1.A \otimes t_2.A \neq \langle \varnothing, \alpha, \beta \rangle, t_1 \in r_1, t_2 \in r_2\}$.*

Example 4.5. *Given two URDB relations $PATIENT_1$ and $PATIENT_2$ as in Tables 3 and 4, 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 5. Here, the names of each relation and its schema are identical, and the set $\wp(A)$ for each attribute A in the schemas consists of extended probabilistic triples on $dom(A)$.*

TABLE 3. Relation PATIENT_1

P_ID	P_DISEASE
P0421	$\langle\{\text{bronchitis, bronchiectasis}\}, 0.9u, 1.2u\rangle$
P3829	$\langle\{\{\text{cholecystitis, gall-stone}\}\}, u, u\rangle$

TABLE 4. Relation PATIENT_2

P_NAME	P_DISEASE
Peter	$\langle\{\text{bronchiectasis}\}, u, u\rangle$
George	$\langle\{\{\text{cholecystitis, gall-stone}\}, \text{cirrhosis}\}, 0.8u, 1.4u\rangle$

4.5. **Intersection, union and difference.** The intersection, union and difference of two URDB relations over the same schema is a URDB relation over that schema, where two tuples that have the same key, respectively of those two relations should be coalesced into a tuple in the result relation by a probabilistic combination strategy. Here, two tuples

TABLE 5. Relation $\text{PATIENT}_1 \bowtie_{\otimes_{in}} \text{PATIENT}_2$

P_ID	P_NAME	P_DISEASE
P0421	Peter	$\langle\{\text{bronchiectasis}\}, 0.45u, 0.6u\rangle$
P3829	George	$\langle\{\{\text{cholecystitis, gall-stone}\}\}, 0.4u, 0.7u\rangle$

that have the same key value are similar to two tuples that are identical in the classical relational database model. Thus, the operations are an extension of the intersection, union and difference of two CRDB relations with uncertain multivalued attributes. The intersection, union and difference of two URDB relations in turn are defined as below.

Definition 4.9. *Let $R(\boldsymbol{U}, \wp)$ be a URDB 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 URDB relation r over R defined by $r = \{t | t.A = t_1.A \otimes t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$, such that $t_1[K] = t_2[K]$ and $t_1.A \otimes t_2.A \neq \langle\varnothing, \alpha, \beta\rangle\}$.*

It is noted that, the notation $t_1[K] = t_2[K]$ is used in the definition due to the value of each key attribute assumed to be certain and definite as in Definition 3.6. Moreover, each tuple is uniquely determined by every key of a relation. So, the result relation is unique by all the keys.

Example 4.6. *Given two URDB relations $DIAGNOSE_1$ and $DIAGNOSE_2$ over the same schema $\boldsymbol{DIAGNOSE}(\boldsymbol{U}, \wp)$ as in Tables 6 and 7, where $\boldsymbol{U} = \{$P_ID, D_ID, P_DISEASE, D_COST$\}$, $\{$P_ID, D_ID$\}$ is the key of $\boldsymbol{DIAGNOSE}$ and the set $\wp(A)$ for each attribute A in $\boldsymbol{U}$ consists of all probabilistic triples on $dom(A)$. Then the intersection of $DIAGNOSE_1$ and $DIAGNOSE_2$ under $\otimes_{in}$ is the relation $DIAGNOSE_1 \cap_{\otimes_{in}} DIAGNOSE_2$ computed as in Table 8.*

TABLE 6. Relation DIAGNOSE_1

P_ID	D_ID	P_DISEASE	D_COST
P216	D012	$\langle\{\text{lung cancer, tuberculosis}\}, 0.8u, 1.2u\rangle$	$\langle\{30, 35\}, 0.7u, 1.3u\rangle$
P244	D024	$\langle\{\{\text{hepatitis, cirrhosis}\}, \text{pancreatitis}\}, 0.6u, 1.3u\rangle$	$\langle\{8, 9\}, 0.5u, 1.2u\rangle$

TABLE 7. Relation DIAGNOSE_2

P_ID	D_ID	P_DISEASE	D_COST
P218	D012	$\langle\{\text{lung cancer}\}, u, u\rangle$	$\langle\{30\}, u, u\rangle$
P244	D024	$\langle\{\{\text{hepatitis, cirrhosis}\}, \text{cholecystitis}\}, 0.4u, 1.2u\rangle$	$\langle\{7, 8\}, 0.5u, 1.2u\rangle$
P252	D025	$\langle\{\text{dyspepsia}\}, u, u\rangle$	$\langle\{5\}, u, u\rangle$

TABLE 8. Relation $\text{DIAGNOSE}_1 \cap_{\otimes_{in}} \text{DIAGNOSE}_2$

P_ID	D_ID	P_DISEASE	D_COST
P244	D024	$\langle\{\{\text{hepatitis, cirrhosis}\}\}, 0.06u, 0.39u\rangle$	$\langle\{8\}, 0.0625u, 0.36u\rangle$

Definition 4.10. *Let $R(\boldsymbol{U}, \wp)$ be a URDB schema, r_1 and r_2 be two relations over R, K be a key of R, and $\oplus$ 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 URDB relation r over $R(\boldsymbol{U}, \wp)$ defined by $r = \{t_1 \in r_1 | \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t_2 \in r_2 | \forall t_1 \in r_1, t_2[K] \neq t_1[K]\} \cup \{t | t.A = t_1.A \oplus t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1[K] = t_2[K]\}$.*

Definition 4.11. *Let $R(\boldsymbol{U}, \wp)$ be a URDB schema, r_1 and r_2 be two relations over R, K be a key of R, $\ominus$ 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 URDB relation r over $R(\boldsymbol{U}, \wp)$ defined by $r = \{t_1 \in r_1 | \forall t_2 \in r_2, t_1[K] \neq t_2[K]\} \cup \{t | t.A = t_1.A \ominus t_2.A, t_1 \in r_1, t_2 \in r_2, A \in \boldsymbol{U}$ such that $t_1[K] = t_2[K]$ and $t_1.A \ominus t_2.A \neq \langle \varnothing, \alpha, \beta \rangle\}$.*

We note that as for Definition 4.9, the result relations in Definitions 4.10 and 4.11 do not depend on choosing the keys of their relational schema.

Example 4.7. *Given two URDB relations $DIAGNOSE_1$ and $DIAGNOSE_2$ over the same schema $\boldsymbol{DIAGNOSE}(\boldsymbol{U}, \wp)$ as in Tables 6 and 7 of Example 4.6. Then the union of $DIAGNOSE_1$ and $DIAGNOSE_2$ under $\oplus_{in}$ is the relation $DIAGNOSE_1 \cup_{\oplus_{in}} DIAGNOSE_2$ computed as in Table 9.*

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

P_ID	D_ID	P_DISEASE	D_COST
P216	D012	$\langle\{\text{lung cancer, tuberculosis}\}, 0.8u, 1.2u\rangle$	$\langle\{30, 35\}, 0.7u, 1.3u\rangle$
P218	D012	$\langle\{\text{lung cancer}\}, u, u\rangle$	$\langle\{30\}, u, u\rangle$
P252	D025	$\langle\{\text{dyspepsia}\}, u, u\rangle$	$\langle\{5\}, u, u\rangle$
P244	D024	$\langle\{\text{pancreatitis, cholecystitis, }\{\text{hepatitis, cirrhosis}\}\}, \alpha, \beta\rangle$, where $\alpha(\text{pancreatitis}) = 0.3$, $\beta(\text{pancreatitis}) = 0.65$, $\alpha(\text{cholecystitis}) = 0.2$, $\beta(\text{cholecystitis}) = 0.6$, $\alpha(\{\text{hepatitis, cirrhosis}\}) = 0.44$, $\beta(\{\text{hepatitis, cirrhosis}\}) = 0.86$.	$\langle\{7, 8, 9\}, \alpha, \beta\rangle$, where $\alpha(7) = 0.25$, $\beta(7) = 0.6$, $\alpha(9) = 0.25$, $\beta(9) = 0.6$, $\alpha(8) = 0.4375$, $\beta(8) = 0.84$.

4.6. **Property of algebraic operations.** The properties of the algebraic operations in URDB are extended from those in CRDB. Clearly, these properties say that our URDB model is sound and coherent.

Proposition 4.1. *Let R be a URDB schema, r be a relation over R, ϕ_1 and ϕ_2 be two selection conditions. Then*

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

Proof: Let $s = \sigma_{\phi_2}(r)$, we have

$$\begin{aligned}
& \sigma_{\phi_1}(\sigma_{\phi_2}(r)) \\
= \ & \{t \in s | Prob_{R,s,t} \models \phi_1\} \quad \text{(Definition 4.5)} \\
= \ & \{t \in r | (Prob_{R,r,t} \models \phi_2) \wedge (Prob_{R,s,t} \models \phi_1)\} \\
= \ & \{t \in r | (Prob_{R,r,t} \models \phi_2) \wedge (Prob_{R,r,t} \models \phi_1)\} \quad \text{(because } s \subseteq r\text{)} \\
= \ & \{t \in r | Prob_{R,r,t} \models \phi_1 \wedge \phi_2\} \quad \text{(Definition 4.4)} \\
= \ & \sigma_{\phi_1 \wedge \phi_2}(r).
\end{aligned}$$

Thus, the equation $\sigma_{\phi_1}(\sigma_{\phi_2}(r)) = \sigma_{\phi_1 \wedge \phi_2}(r)$ is proven. The equation $\sigma_{\phi_2}(\sigma_{\phi_1}(r)) = \sigma_{\phi_2 \wedge \phi_1}(r)$ is similarly proven. Since $\phi_1 \wedge \phi_2 \Leftrightarrow \phi_2 \wedge \phi_1$. So, Proposition 4.1 is proven.

Proposition 4.2. *Let R be a URDB 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*

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

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

Proposition 4.3. *Let R_1, R_2 and R_3 be the URDB 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, and $\otimes$ be a probabilistic conjunction strategy. Then*

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

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

Equations (3) and (4) say that the join operation of URDB 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. By Definition 2.4, the conjunction of probabilistic triples is commutative (due to the commutativity of probabilistic conjunction strategies and the intersection of sets). Therefore, by Definition 4.8, it yields $r_1 \bowtie_{\otimes} r_2 = r_2 \bowtie_{\otimes} r_1$.

By Definition 4.8, the results of two sides of (4) are the relations over the same schema. Moreover, the intersection of sets has the associativity, by Definition 2.4, it follows that the conjunction of probabilistic triples is associative. From the associativity of the classical relational join and by Definition 4.8, it is easy to see that the join of URDB relations is associative. Thus, it results in $(r_1 \bowtie_{\otimes} r_2) \bowtie_{\otimes} r_3 = r_1 \bowtie_{\otimes} (r_2 \bowtie_{\otimes} r_3)$.

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

Corollary 4.1. *Let R_1, R_2 and R_3 be URDB 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. Then*

$$r_1 \times r_2 = r_2 \times r_1 \tag{5}$$

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

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

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

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

$$r_1 \cup_{\oplus} r_2 = r_2 \cup_{\oplus} r_1 \tag{9}$$

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

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

Proof: From commutativity and associativity of the intersection of sets, it follows that the conjunction of probabilistic triples has commutativity and associativity (Definition 2.4). So, the intersection of URDB relations r_1, r_2 and r_3 under the probabilistic conjunction strategy $\otimes$ and every chosen key also has commutativity and associativity. From that, by Definition 4.9, it follows Equations (7) and (8).

From commutativity and associativity of the union, intersection of sets, it yields commutativity and associativity of the union of probabilistic triples (Definition 2.5). Therefore, the union of URDB relations r_1, r_2 and r_3 under the probabilistic disjunction strategy $\oplus$ and every chosen key also has commutativity and associativity. From that, by Definition 4.10, it follows Equations (9) and (10).

For ending this section, we note that the computing complexity of URDB algebraic operations is a polynomial under the size of relations. For instance, 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 (Definitions 4.3 and 4.4), then the cost for the selection of each tuple in a URDB relation (Definition 4.5) also is some constant (i.e., O(1)). It results in the computing time complexity of the selection operation on a URDB relation having n tuples is O(n). Similarly, the computing time complexity of Cartesian product and join operations on two URDB relations having n and m tuples is O(nm). Thus, we can say that the performance of URDB model in computing and manipulating uncertain information is good and can apply it in practice.

5. **Conclusions.** We have introduced a new relational database model and algebra, abbreviated to URDB, as a development following the probabilistic relational database model with multivalued attributes for representing and dealing with uncertain information. In URDB, each relation is a set of tuples whose attributes may take more than one uncertain value represented by an extended probabilistic triple. The uncertain degree of functional dependencies, keys and relations on attribute values as well as the satisfied degree of data queries are computed and determined by using the probabilistic interpretation of binary relations on sets. The URDB algebra with a full set of basic operations is defined and built by using the new combination strategies of extended probabilistic triples. Basic properties of the URDB operations are proposed and proven completely to say that URDB is a sound and coherent model.

Towards applying URDB in practice, we will build a management system for URDB with the familiar querying and manipulating language like SQL that is able to represent and handle uncertain information in the real world.

REFERENCES

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

[2] C. J. Date, *An Introduction to Database Systems*, 8th Edition, Addison Wesley, 2004.

[3] P. Bosc, D. Kraft and F. Petry, Fuzzy sets in database and information systems: Status and opportunities, *Journal of Fuzzy Sets and Systems*, vol.156, no.3, pp.418-426, 2005.

[4] J. Galindo, A. Urrutia and M. Piattini, *Fuzzy Databases: Modeling, Design and Implementation*, Idea Group Inc., 2006.

[5] E. Doumard, O. Pivert, G. Smits and V. Thion, Processing fuzzy relational queries using fuzzy views, *Proc. of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, New Orleans, LA, USA, pp.332-337, 2019.

[6] H. Nguyen, T. U. N. Nguyen and N. D. Le, Fuzzy relational database model and management system for imprecise information, *Journal of Computer Science and Cybernetics*, vol.37, no.2, pp.145-162, 2021.

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

[8] I. I. Ceylan, S. Borgwardt and T. Lukasiewicz, Most probable explanations for probabilistic database queries, *Proc. of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, pp.950-956, 2017.

[9] 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.

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

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

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

[13] S. Zhang and C. Zhang, A probabilistic data model and its semantics, *Journal of Research and Practice in Information Technology*, vol.35, no.4, pp.237-256, 2003.
[14] A. Dekhtyar, R. Ross and V. S. Subrahmanian, Probabilistic temporal databases, I: Algebra, *ACM Transactions on Database Systems*, vol.26, no.1, pp.41-95, 2001.
[15] W. Zhao, A. Dekhtyar and J. Goldsmith, Databases for interval probabilities, *International Journal of Intelligent Systems*, vol.19, no.9, pp.789-815, 2004.
[16] L. V. S. Lakshmanan, N. Leone, R. Ross and V. S. Subrahmanian, Probview: A flexible probabilistic database system, *ACM Transactions on Database Systems*, vol.22, no.3, pp.419-469, 1997.
[17] R. Ross and V. S. Subrahmanian, Aggregate operators in probabilistic databases, *Journal of the ACM*, vol.52, no.1, pp.54-101, 2005.
[18] D. Dey and S. Sarkar, Generalized normal forms for probabilistic relational data, *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.3, pp.485-497, 2002.
[19] D. Barbara, H. Garcia-Molina and D. Porter, The management of probabilistic data, *IEEE Transactions on Knowledge and Data Engineering*, vol.4, no.5, pp.487-502, 1992.
[20] T. Eiter, T. Lukasiewicz and M. Walter, A data model and algebra for probabilistic complex values, *Annals of Mathematics and Artificial Intelligence*, vol.33, pp.205-252, 2001.
[21] T. Eiter, J. J. Lu, T. Lukasiewicz and V. S. Subrahmanian, Probabilistic object bases, *ACM Transactions on Database Systems*, vol.26, no.3, pp.264-312, 2001.
[22] Y. Kornatzky and S. E. Shimony, A probabilistic object-oriented data model, *Data and Knowledge Engineering*, vol.12, pp.143-166, 1994.
[23] S. K. Lee, An extended relational database model for uncertain and imprecise information, *Proc. of the 18th Conference on Very Large Data Bases*, Vancouver, British Columbia, Canada, pp.211-220, 1992.
[24] H. Nguyen, A probabilistic relational database model and algebra, *Journal of Computer Science and Cybernetics*, vol.31, no.4, pp.305-321, 2015.
[25] H. Nguyen, Extending relational database model for uncertain information, *Journal of Computer Science and Cybernetics*, vol.35, no.4, pp.355-372, 2019.
[26] H. Nguyen, T.-N. Nguyen and T.-T.-N. Tran, A probabilistic relational database model with uncertain multivalued attributes, *ICIC Express Letters*, vol.16, no.3, pp.241-248, 2022.
[27] L. Yan and M. Zongmin, A probabilistic nested relational database model with fuzzy probability measures, *Proc. of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, New Orleans, LA, USA, pp.253-257, 2019.
[28] A. V. Vitianingsih, I. Wisnubhadra, S. S. K. Baharin, R. Marco and A. L. Maukar, Classification of pertussis vulnerable area with location analytics using multiple attribute decision making, *International Journal of Innovative Computing, Information and Control*, vol.16, no.6, pp.1943-1957, 2020.
[29] T. Friedman and G. Broeck, Symbolic querying of vector spaces: Probabilistic databases meets relational embeddings, *Proc. of the 36th Conference on Uncertainty in Artificial Intelligence*, Toronto, Canada, vol.124, pp.1268-1277, 2020.
[30] J. Bernad, C. Bobed and E. Mena, Uncertain probabilistic range queries on multidimensional data, *Information Sciences*, vol.537, pp.334-367, 2020.
[31] M. Huang, Z. Hu and L. Wang, Optimal consensus control for heterogeneous nonlinear non-affine multi-agent systems with uncertain control directions, *ICIC Express Letters*, vol.16, no.2, pp.177-185, 2022.

Author Biography

Hoa Nguyen received his Ph.D. degree in Computer Science at Vietnam National University, Ho Chi Minh City, Vietnam, in 2008.

Dr. Nguyen is currently an associate professor at Information Technology Faculty, Saigon University, Vietnam. His research interest includes imprecise and uncertain knowledge representation, fuzzy databases and probabilistic dadabases.