

# Probabilistic XML functional dependencies based on possible world model

Ping Yan<sup>1</sup>, Teng Lv<sup>2\*</sup>, Weimin He<sup>3</sup>

<sup>1</sup>School of Science, Anhui Agricultural University, Hefei 230036, China

<sup>2</sup>Teaching and Research Section of Computer, Army Officer Academy, Hefei 230031, China

<sup>3</sup>Department of Computing and New Media Technologies, University of Wisconsin-Stevens Point, 2100 Main Street, Stevens Point, WI 54481

Received 1 March 2014, www.cmnt.lv

## Abstract

With the increase of uncertain data in many new applications, such as sensor network, data integration, web extraction, etc., uncertainty both in relational databases and XML datasets has attracted more and more research interests in recent years. As functional dependencies (FDs) are critical and necessary to schema design and data rectification in relational databases and XML datasets, it is also significant to study FDs in uncertain XML datasets. This paper first proposed XML functional dependencies (XFDs) of deterministic XML dataset based on tree tuple models. Then two new kinds of functional dependencies based on possible worlds model for probabilistic XML dataset are introduced: probabilistic XML functional dependencies (pXFDs) and probabilistic approximate XML functional dependencies (pAXFDs). pXFDs extend the concept of XFDs of deterministic XML dataset by considering the probability of each possible world of probabilistic XML dataset, and pAXFDs extend the concept of probabilistic XML functional dependencies of probabilistic XML dataset by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset.

*Keywords:* uncertain XML, functional dependency, inference rule, closed set

## 1 Introduction

XML (Extensible Markup Language) has become the de facto standard of data exchange and is widely used in many fields. With the increase in applications such as data integration, web extraction, sensor networks, etc., XML datasets may be obtained from heterogeneous data sources and are not always deterministic. In such cases, XML datasets may contain uncertain data for the same attribute or element due to different data sources, information extraction, approximate query, and data measurement. Although functional dependencies (FDs) of uncertain XML datasets are much more complicated than the counterparts of traditional relational databases and deterministic XML datasets, it is possible and necessary to study them in uncertain XML datasets as shown in the paper.

**Related work.** Although there has been a lot of significant work in functional dependencies for relational databases and XML datasets, none of them can be directly applied to uncertain XML datasets. We analyse the related work in the following three aspects:

(1) For traditional relational databases, functional dependencies are thoroughly studied for several decades [1, 2]. Ref. [3] proposed a concept of functional dependencies in relational databases, which can deal with slight variations of data values. Ref. [4] proposed the conditional functional dependencies to detect and correct

data inconsistency. It is obviously that the techniques of traditional relational databases cannot be directly applied to XML due to the significant difference in structure between XML documents and relational databases.

(2) For traditional deterministic XML datasets, functional dependencies are also thoroughly studied for some years. There are two major approaches to define functional dependencies in XML research community, i.e. path-based approach and sub-tree/sub-graph-based approach. In path-based approach [5-11], XML datasets are represented by a tree structure, and some paths of the tree with their values are used in defining XML functional dependencies. In Sub-tree/Sub-graph-based approach [12, 13], functional dependencies of XML datasets are defined by sub-graph or sub-tree in XML datasets. A sub-graph or a sub-tree is a set of paths of XML datasets. As an improvement over Sub-tree/Sub-graph-based approach functional dependencies, Refs. [14, 15] deal with XML functional dependencies with some constraint condition such that there exists a sub-tree is equal. The above XML functional dependencies cannot deal with uncertainty in XML datasets, which is the research topic of the paper. Ref. [16] proposes an approach to discover a set of minimal XML Conditional Functional Dependencies (XCFDs) from a given XML instance to improve data consistency. The XCFDs extends XML Functional Dependencies (XFDs) by incorporating conditions into XFD specifications. It is

\* *Corresponding author* e-mail LT0410@163.com

easy to see that all the functional dependencies defined above cannot deal with uncertainty in XML datasets.

(3) For uncertain relational databases, Ref. [17] proposes the probabilistic functional dependency for probabilistic relational databases which associated with a likelihood of the traditional functional dependency is satisfied. Ref.[18] proposes some kinds of functional dependencies for probabilistic relational databases, such as Probabilistic Approximate Functional Dependencies (pAFD), Conditional Probabilistic Functional Dependencies (CpFD), and Conditional Probabilistic Approximate Functional Dependencies (CpAFD), which combine approximate, conditional, and approximate/conditional characteristics into traditional functional dependencies to defined corresponding functional dependencies for probabilistic relational databases. Ref. [19] proposes horizontal functional dependencies and vertical functional dependencies for uncertain relational databases, which extends the traditional relational functional dependencies into the uncertain relational databases. Although these work of uncertain relational databases are meaningful and significant, they can not directly applied in uncertain XML datasets, as XML are more complicated in structure then relational databases.

**Contributions.** In this paper, we will extend the concept of traditional deterministic XML functional dependencies (XFDs) to study the FDs of probabilistic XML datasets by considering the probability of each possible world and the degree of truth of tuples in each possible world. The main contributions of the paper are detailed as followings:

(1) We first proposed XML functional dependencies (XFDs) of deterministic XML dataset based on tree tuple models.

(2) Then a new kind of FDs called probabilistic XML functional dependencies (pXFDs) based on possible worlds model for probabilistic XML dataset is introduced, which extends the concept of XFDs of deterministic XML dataset by considering the probability of each possible world of probabilistic XML dataset.. pXFDs, and pAXFDs A sound and complete inference rules are given for the three types of uncertain XML functional dependencies.

(3) Finally another new kind of FDs called probabilistic approximate XML functional dependencies (pAXFDs) is introduced, which extend the concept of pXFDs of probabilistic XML dataset by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset.

(4) We also analyze the relationship among the tree kinds of FDs: XFDs, pXFDs and pAXFDs. pXFDs extend XFDs by considering the probability of each possible world of probabilistic XML dataset, and pAXFDs extend pXFDs by considering the degree of truth of tree tuples in each possible world of probabilistic XML dataset. More generally speaking, pXFDs are more

general than XFDs, and pAXFDs are more general than pXFDs.

**Organizations.** The rest of the paper is organized as following: Section 2 gives an example as our research motivation and demonstration of the concepts throughout the paper. Three types of functional dependencies, including XFDs, pXFDs, and pAXFDs, of probabilistic XML datasets are given in Section 3. Finally, Section 4 concludes the paper and points out the future directions of the paper.

## 2 A Motivating example

Suppose we want to know the people impressions of relationship between a man's salary and his diploma/height. In terms of functional dependencies, if it is the case that [diploma, height]→salary? We design a questionnaire and obtain some data as the following:

TABLE 1 Four interviewees' impressions about the relationship between a man's salary and his diploma/height

Record	Interviewee ID	diploma	height	salary	probability
1	1	High	high	high	0.9
2	1	High	low	high	0.8
3	1	Low	high	low	0.5
4	1	Low	low	low	0.7
5	2	High	high	high	0.9
6	2	High	low	high	0.7
7	2	Low	high	high	0.5
8	2	Low	low	low	0.8
9	3	High	high	high	0.6
10	3	Low	low	low	0.7
11	4	High	low	high	0.7
12	4	Low	high	high	0.5

For Record 1 in Table 1, interviewee 1 thinks (with confidence of 0.9) that if a man's diploma is high and height is high, then his salary is high. We can store Table 1 as an XML file (survey.xml) as following (attribute "Prob" stands for the confidence of an interviewee's answer):

```

<survey>
  <interviewee>
    <ID>1</ID>
    <answer Prob='0.9'>
      <diploma>high</diplomas>
      <height>high</height>
      <salary>high</salary>
    </answer>
    <answer Prob='0.8'>
      <diploma>high</diplomas>
      <height>low</height>
      <salary>high</salary>
    </answer>
    ... ..
  </interviewee>
  <interviewee>
    <ID>2</ID>

```

```

<answer Prob='0.9'>
  <diploma>high</diplomas>
  <height>high</height>
  <salary>high</salary>
</answer>
... ..
</interviewee>
<interviewee>
  <ID>3</ID>
  <answer Prob='0.7'>
    <diploma>high</diplomas>
    <height>high</height>
    <salary>high</salary>
  </answer>
  ... ..
</interviewee>
<interviewee>
  <ID>4</ID>
  <answer Prob='0.7'>
    <diploma>high</diplomas>
    <height>low</height>
    <salary>high</salary>
  </answer>
  ... ..
</interviewee>
</survey>
    
```

We will use the above XML file as demonstration of our proposed concepts throughout the paper.

### 3 Functional dependencies of probabilistic XML dataset

We first give some preliminary definitions such as DTD (Document Type Definition), XML tree, tree tuple, etc.:

**Definition 1 (DTD).** A DTD[20] is defined to be  $D=(E, A, P, R, r)$ , where (1)  $E$  is a finite set of element types; (2)  $A$  is a finite set of attributes; (3)  $P$  is a mapping from  $E$  to element type definitions. For each  $\tau \in E$ ,  $P(\tau)$  is a regular expression  $\alpha$  defined as  $\alpha ::= S | \varepsilon | \tau' | \alpha | \alpha | \alpha | \alpha^*$ , where  $S$  denotes string types,  $\varepsilon$  is the empty sequence,  $\tau' \in E$ , “|”, “”, and “\*” denote union, concatenation and Kleene closure respectively; (4)  $R$  is a mapping from  $E$  to the power set of  $A$ :  $P(A)$ ; (5)  $r \in E$  is called the element type of the root.

A path  $p$  in  $D=(E, A, P, R, r)$  is defined to be  $p = \omega_1 \dots \omega_n$ , where (1)  $\omega_1 \in r$ ; (2)  $\omega_i \in P(\omega_{i-1})$ ,  $i \in [2, n-1]$ ; (3)  $\omega_n \in P(\omega_{n-1})$  if  $\omega_n \in E$  and  $P(\omega_n) \neq \Phi$ , or  $\omega_n = S$  if  $\omega_n \in E$ , and  $P(\omega_n) = \Phi$ , or  $\omega_n \in R(\omega_{n-1})$  if  $\omega_n \in A$ . Let  $paths(D) = \{p | p \in D\}$ .

**Definition 2 (XML tree).** Let  $D=(E, A, P, R, r)$ . An XML tree  $T$  conforming to  $D$  (denoted by  $T \models D$ ) is defined to be  $T=(V, lab, ele, att, val, root)$ , where (1)  $V$  is a finite set of nodes; (2)  $lab$  is a mapping from  $V$  to  $E \cup A$ ; (3)  $ele$  is a partial function from  $V$  to  $V^*$  such that for any  $v \in V$ ,  $ele(v) = [v_1, \dots, v_n]$  if  $lab(v_1) \dots lab(v_n)$  is defined in  $P(lab(v))$ ; (4)  $att$  is a partial function from  $V$  to  $A$  such that for any  $v \in V$ ,  $att(v) = R(lab(v))$  if  $lab(v) \in E$  and  $R(lab(v))$  is defined in  $D$ ; (5)  $val$  is a partial function from  $V$  to  $S$  such that for any  $v \in V$ ,  $val(v)$  is defined if  $P(lab(v)) = S$  or  $lab(v) \in A$ ; (6)  $lab(root) = r$  is called the root of  $T$ .

Given a DTD  $D$  and an XML tree  $T \models D$ , a path  $p$  in  $T$  is defined to be  $p = v_1 \dots v_n$ , where (1)  $v_1 \in root$ ; (2)  $v_i \in ele(v_{i-1})$ ,  $i \in [2, n-1]$ ; (3)  $v_n \in ele(v_{n-1})$  if  $lab(v_n) \in E$ , or  $v_n \in att(v_{n-1})$  if  $lab(v_n) \in A$ , or  $v_n = S$  if  $P(lab(v_{n-1})) = S$ . Let  $paths(T) = \{p | p \in T\}$ .

Fig.1 is a part of an XML tree corresponding to XML file survey.xml in Seciton 2, in which Interview 1 thinks with confidence 0.9 that if a man’s diploma is high and height is high, then his salary is high.

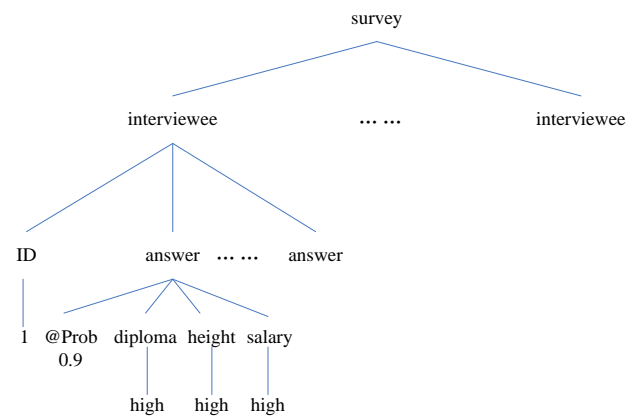


FIGURE 1 A part of an XML tree corresponding to XML file survey.xml

**Definition 3 (Tree tuple).** Given a DTD  $D=(E, A, P, R, r)$  and an XML tree  $T=(V, lab, ele, att, val, root)$  conforming to  $D$ , a tree tuple  $t$  for a node  $v$  in  $T$  is a tree rooted on node  $v$  with all of its decedent nodes.

In XML file survey.xml, there are 4 “interviewee” tree tuples, Fig 1 shows the first “interviewee” tree tuple which is the tree rooted on the first “interviewee” node (ID=1). There are also 12 “answer” tree tuples in the XML file, and Fig 1 shows the first “answer” tree tuple, which is the tree, rooted on the first “answer” node with probability of 0.9.

Based on the above concepts, we give the definition of XML functional dependencies (XML FDs) of deterministic XML dataset, which is based on tree tuple model.

**Definition 4 (XML FDs).** Given a DTD  $D$  and an XML tree  $T \models D$ , an XML Functional Dependency (XFD) has the form  $\{s_h, S_L \rightarrow S_R\}$ , where  $s_h$  is the head path,  $S_L = \{s_{L1}, s_{L2}, \dots, s_{Lm}\}$  is the left path set (i.e. determinant path set), and  $S_R = \{s_{R1}, s_{R2}, \dots, s_{Rn}\}$  is the right path set (i.e. determined path set). For  $\forall s_{Li} \in S_L, s_{Rj} \in S_R$ ,  $s_h, s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$ , where  $i = 1, \dots, m, j = 1, \dots, n$ . If  $T$  satisfies XFD  $\{s_h, S_L \rightarrow S_R\}$  (denoted by  $T \models \{s_h, S_L \rightarrow S_R\}$ ), then it implies that for each tree tuple  $t$  rooted on node  $last(s_h)$ ,  $t(s_h.s_{L1}, s_h.s_{L2}, \dots, s_h.s_{Lm})$  can uniquely determines  $t(s_h.s_{R1}, s_h.s_{R2}, \dots, s_h.s_{Rn})$ , where  $last(s_h)$  denotes the last symbol of head path  $s_h$ .

By considering the probability of each possible world of probabilistic XML dataset, we give the definition of probabilistic XML FDs of deterministic XML dataset based on possible world model by extending the above concept of XFDs as following:

**Definition 5 (Probabilistic XML FDs).** Given a DTD  $D$  and an uncertain XML tree  $T \models D$ , a Probabilistic XML FD (pXFD) has the form  $\{s_h, S_L \rightarrow_{pXFD} S_R\}$ , where  $s_h$  is the head path,  $S_L = \{s_{L1}, s_{L2}, \dots, s_{Lm}\}$  is the left path set (i.e. determinant path set), and  $S_R = \{s_{R1}, s_{R2}, \dots, s_{Rn}\}$  is the right path set (i.e. determined path set). For  $\forall s_{Li} \in S_L, s_{Rj} \in S_R$ ,  $s_h, s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$ , where  $i = 1, \dots, m, j = 1, \dots, n$ .  $T$  satisfies pXFD  $\{s_h, S_L \rightarrow_{pXFD} S_R\}$  (denoted by  $T \models \{s_h, S_L \rightarrow_{pXFD} S_R\}$ ) with a confidence  $c$  if and only if the sum of probability of the fraction of possible worlds in which the corresponding XFD  $\{s_h, S_L \rightarrow S_R\}$  holds is  $c$ ,

$$c = \frac{\sum_{i=1}^n p(PW_i)}{\sum_{j=1}^n p(PW_j)}$$

where  $p(PW_i)$  stands for the probability of possible world  $PW_i$ ,  $\sum p(PW_i)$  stands for the sum of probability of all those possible worlds which satisfy XFD  $\{s_h, S_L \rightarrow S_R\}$ , and  $n$  is number of all possible worlds of  $T$ .

**Example 1.** XML file survey.xml has 64 possible worlds (PW) w.r.t. element “interviewee”, the first possible world is  $PW_1$  with the following 4 tree tuples:

$T_1^1 =$  (survey.interviewee.ID=1,  
survey.interviewee.answer.@Prob=0.9,  
survey.interviewee.answer.diploma=high,  
survey.interviewee.answer.height=high,  
survey.interviewee.answer.salary=high),  
 $T_2^1 =$  (survey.interviewee.ID=2,  
survey.interviewee.answer.@Prob=0.9,  
survey.interviewee.answer.diploma=high,  
survey.interviewee.answer.height=high,  
survey.interviewee.answer.salary=high),

$T_3^1 =$  (survey.interviewee.ID=3,  
survey.interviewee.answer.@Prob=0.6,  
survey.interviewee.answer.diploma=high,  
survey.interviewee.answer.height=high,  
survey.interviewee.answer.salary=high),  
 $T_4^1 =$  (survey.interviewee.ID=4,  
survey.interviewee.answer.@Prob=0.7,  
survey.interviewee.answer.diploma=high,  
survey.interviewee.answer.height=low,  
survey.interviewee.answer.salary=high).

It is easy to see that  $PW_1$  satisfies the XFD with confidence  $0.9*0.9*0.6*0.7=0.3402$ , i.e.,  $PW_1$  contribute to the final confidence of the entire XML file survey.xml with confidence 0.3402. We omit other possible worlds considering the space here and we can see that the number of possible worlds which satisfy XFD  $\{survey.interviewee.answer, [diploma, height] \rightarrow salary\}$  is 48, so XML file survey.xml satisfies pXFD  $\{survey.interviewee.answer, [diploma, height] \rightarrow_{pXFD} salary\}$  with confidence  $c = \frac{\sum_{i=1}^n p(PW_i)}{\sum_{j=1}^n p(PW_j)} = \frac{11.9416}{13.1116} \approx 0.91$ ,

where  $p(PW_i)$  stands for the probability of possible world  $PW_i$  (for example,  $p(PW_1) = 0.9*0.9*0.6*0.7=0.3402$ ),  $\sum p(PW_i)$  stands for the sum of probability of all those possible worlds which satisfy the XFD  $\{survey.interviewee.answer, [diploma, height] \rightarrow salary\}$ , and  $n$  is the number of all possible worlds of survey.xml file. The meaning of the pXFD  $\{survey.interviewee.answer, [diploma, height] \rightarrow_{pXFD} salary\}$  is that the four interviewees in the survey think (with confidence  $c=0.91$ ) that a person’s diploma and height can determine his/her salary.

The relationship between XFDs and pXFDs are given in the following theorem:

**Theorem 1 (Relationship between XFD and pXFDs).** If each possible world satisfies a pXFD with probability of 1 (i.e. the XML dataset are deterministic), then the pXFD is equal to the corresponding XFD.

**Proof.** From the definitions of pXFDs and XFDs, we can see that pXFDs are natural extensions of corresponding traditional deterministic XFDs by considering the probability of each possible world of probabilistic XML dataset. So if all possible world satisfy a pXFD with probability of 1, then the entire XML dataset must satisfy the corresponding XFD.

It should be noted that pXFDs suffer from the same kind of flaws that XFDs do. More specifically, if the XML dataset is dirty, just a tree tuple in a possible world that does not conform to the XFDs can cause the entire possible world to satisfy the XFDs. For example, consider possible world  $PW_{34}$  in XML file survey.xml, which has the following 4 tree tuples:

$T_1^{34} =$  (survey.interviewee.ID=1,  
survey.interviewee.answer.@Prob=0.5,  
survey.interviewee.answer.diploma=low,



survey.interviewee.answer.height=high,  
 survey.interviewee.answer.salary=low),  
 $T_2^{34} =$  (survey.interviewee.ID=2,  
 survey.interviewee.answer.@Prob=0.9,  
 survey.interviewee.answer.diploma=high,  
 survey.interviewee.answer.height=high,  
 survey.interviewee.answer.salary=high),  
 $T_3^{34} =$  (survey.interviewee.ID=3,  
 survey.interviewee.answer.@Prob=0.6,  
 survey.interviewee.answer.diploma=high,  
 survey.interviewee.answer.height=high,  
 survey.interviewee.answer.salary=high),  
 $T_4^{34} =$  (survey.interviewee.ID=4,  
 survey.interviewee.answer.@Prob=0.5,  
 survey.interviewee.answer.diploma=low,  
 survey.interviewee.answer.height=high,  
 survey.interviewee.answer.salary=high).

It is easy to see that  $PW_{34}$  does not satisfy pXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pXFD}$  salary} as tree tuple  $T_1^{34}$  and tree tuple  $T_4^{34}$  are contradict to each other.

To solve the above problem, we propose the concept of probabilistic approximate XML FDs as following:

**Definition 6 (Probabilistic Approximate XML FDs).** Given a DTD  $D$  and an uncertain XML tree  $T \models D$ , a Probabilistic Approximate XML FD (pAXFD) has the form  $\{s_h, S_L \rightarrow_{pAXFD} S_R\}$ , where  $s_h$  is the head path,  $S_L = \{s_{L1}, s_{L2}, \dots, s_{Lm}\}$  is the left path set (determinant path set), and  $S_R = \{s_{R1}, s_{R2}, \dots, s_{Rn}\}$  is the right path set (determined path set). For  $\forall s_{Li} \in S_L, s_{Rj} \in S_R$ ,  $s_h, s_h.s_{Li}, s_h.s_{Rj} \in paths(D)$ .  $T$  satisfies pAXFD  $\{s_h, S_L \rightarrow_{pAXFD} S_R\}$  (denoted by  $T \models \{s_h, S_L \rightarrow_{pAXFD} S_R\}$ ) with a confidence  $c$  if and only if the sum of probability of each possible world multiplied by the maximal fraction of the tree tuples of the possible world, in which the corresponding pXFD  $\{s_h, S_L \rightarrow_{pXFD} S_R\}$  holds is  $c$ ,  $c = \frac{\sum_{i=1}^n (p(PW_i) \times \alpha_i)}{\sum_{j=1}^n p(PW_j)}$ , where  $p(PW_i)$  stands

for the probability of possible world  $PW_i$ , and  $\alpha_i$  stands for the maximal tree tuples fraction in possible world  $PW_i$ , which satisfies the corresponding pXFD  $\{s_h, S_L \rightarrow_{pXFD} S_R\}$ , and  $n$  is the number of all possible worlds of  $T$ .

**Example 2.** As noted before, possible world  $PW_{34}$  in XML file survey.xml does not satisfy pXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pXFD}$  salary} as tree tuple  $T_1^{34}$  and tree tuple  $T_4^{34}$  are contradict to each other. If we just remove any one of them, then  $PW_{34}$  satisfies the pXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pXFD}$  salary}. It is easy to see that for pXFD {survey.interviewee.answer, [diploma,

height]  $\rightarrow_{pXFD}$  salary} to be satisfied by  $PW_{34}$ , the removed minimal fraction of tree tuples is 1/4, so  $PW_{34}$  satisfies the pAXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pAXFD}$  salary} with confidence  $p(PW_{34}) * (1-1/4) = 0.5 * 0.9 * 0.6 * 0.5 * 3/4 = 0.10125$ . For the entire XML file survey.xml, it satisfies the pAXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pAXFD}$  salary} with the following confidence  $c = \frac{\sum_{i=1}^n (p(PW_i) \times \alpha_i)}{\sum_{j=1}^n p(PW_j)} = \frac{12.8191}{13.1116} \approx 0.98$ , where  $p(PW_i)$

stands for the probability of possible world  $PW_i$ , and  $\alpha_i$  stands for the maximal tree tuples fraction in possible world  $PW_i$ , which satisfies the corresponding pXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pXFD}$  salary}, and  $n$  is the number of all possible worlds. The meaning of the pAXFD {survey.interviewee.answer, [diploma, height]  $\rightarrow_{pAXFD}$  salary} is that the four interviewees in the survey approximately think (with confidence  $c=0.98$ ) that a man's diploma and height can determine his salary. We can see that the pAXFD's confidence is generally higher than that of pXFD (the former is 0.98, and the latter is 0.91).

The relationship between pXFDs and pAXFDs are given in the following theorem:

**Theorem 2 (Relationship between pXFDs and pAXFDs).** If all tree tuples of each possible world either satisfy a pXFD or not at all, then the entire probabilistic XML dataset satisfies the corresponding pAXFD with the same confidence. In general, the confidence of a pAXFD is always greater than the confidence of the corresponding pXFD.

**Proof.** (1) From the definitions of pXFDs and pAXFDs, we can see that a pAXFD is a natural extension of the corresponding pXFD by considering the degree of truth of tree tuples in each possible world of a probabilistic XML dataset. For a pAXFD, if each tree tuple of each possible world either satisfies the corresponding pXFD or not at all, then the degree of truth of tree tuples in each possible world is either 1 or 0 w.r.t. the pAXFD. So the pAXFD is equal to the corresponding pXFD with the same confidence. (2) In each possible world that the pXFD holds, the degree of the truth ( $\alpha_i$  in Definition 6) of the corresponding pAXFD is 1. In each possible world that the pXFD does not hold, the degree of the truth of the corresponding pAXFD is always greater than 0. So the final confidence of the pAXFD, which is the weighted sum of the confidences of all possible worlds multiplied by the corresponding degree of truth, is always greater than the confidence of the corresponding pXFDs.

#### 4 Conclusions

This paper studies the functional dependencies (FDs) of probabilistic XML datasets, which extends the notions of

functional dependencies of uncertain relational databases and traditional functional dependencies of XML datasets based on tree-tuple model. Three kinds of functional dependencies such as XFDs, pXFDs, and pAXFDs are given in the paper: XFDs are traditional FDs to capture the relationship between XML element in a deterministic XML dataset; pXFDs are natural extensions of XFDs by considering the probability of each possible world of probabilistic XML dataset; and pAXFDs are natural extensions of pXFDs by considering the degree of truth of each possible world of probabilistic XML dataset, which is very useful in the case that there are noisy or dirty data in probabilistic XML dataset. The relationship among XFDs, pXFDs, and pAXFDs are also studied in the paper. Generally speaking, pXFDs are more general than XFDs, and pAXFDs are more general than pXFDs.

An interesting work in the future is to assess the confidence of pXFDs and pAXFDs. As we know that




both pXFDs and pAXFDs are defined on possible world model and the number of possible world is very large generally, it is a not trivial work to access the confidence of pXFDs and pAXFDs. Another interesting work is to mine the pXFDs and pAXFDs in a given probabilistic XML dataset. Specifically speaking, given a probabilistic XML dataset, find a minimal set of pXFDs or pAXFDs that are equivalent to or more general than any set of pXFDs or pAXFDs that hold over the XML dataset with a confidence higher than a user specified threshold.

### Acknowledgments

The work is supported by Natural Science Foundation of Anhui Province (No.1208085MF110) and Natational Natural Science Foundation of China (No. 11201002).

### References

- [1] Abiteboul S, Hull R, Vianu V 1995 *Foundations of Databases*. Addison-Wesley Boston
- [2] Ullman J D 1988 *Principles of Database and Knowledge-Base Systems 1* Computer Science Press, New York
- [3] Koudas N, Saha A, Srivastava D 2009 Venkatasubramanian, S.: Metric functional dependencies *Proc. of the 25th International Conference on Data Engineering* 1275-8 IEEE Computer Society Press, New York
- [4] Bohanno P, Fan W, Geerts F, Jia X, Kementsietsidis A 2007 Conditional functional dependencies for data cleaning *Proc. of the 23rd International Conference on Data Engineering* 746-55 IEEE Computer Society Press, New York
- [5] Janosi-Rancz K T, Varga V, Nagy T 2010 Detecting XML functional dependencies through formal concept analysis *Proc. of the 14th East European Conference on Advances in Databases and Information Systems* 595-8 Springer, Heidelberg
- [6] Lee M L, Ling T W, Low W L 2002 Designing functional dependencies for XML *Proc. of the 8th International Conference on Extending Database Technology: Advances in Database Technology* 124-41 Springer-Verlag, London
- [7] Liu J, Vincent M, Liu C 2003 Functional dependencies, from relational to XML *Proc. of 5th International Andrei Ershov Memorial Conference* 531-8 Springer, Heidelberg
- [8] Liu J, Vincent M, Liu C 2003 Local XML functional dependencies *Proc. of 5th ACM CIKM International Workshop on Web Information and Data Management* 23-8 ACM, New York
- [9] Vincent M, Liu L 2003 Functional dependencies for XML *Proc. of 5th Asian-Pacific Web Conference* 22-34 Springer, Heidelberg
- [10] Vincent M, Liu J, Liu C 2004 Strong functional dependencies and their application to normal forms in XML *TODS* 29 445-62
- [11] Yan P, Lv T 2006 Functional dependencies in XML documents *Proc. of the 8th Asia Pacific Web Conference Workshop* 29-37 Springer, Heidelberg
- [12] Hartmann S, Link S 2003 More functional dependencies for XML *Proc. of the 7th East European Conference on Advances in Databases and Information Systems* 355-69 Springer, Heidelberg
- [13] Hartmann S, Link S, Trinh T 2010 Solving the implication problem for XML functional dependencies with properties *Proc. of the 18th Workshop on Logic, Language, Information and Computation* 161-75 Springer, Heidelberg
- [14] Lv T, Yan P 2008 Removing XML data redundancies by constraint-tree-based functional dependencies *Proc. of ISECS International Colloquium on Computing, Communication, Control, and Management* 595-9 IEEE Computer Society, Washington, DC
- [15] Lv T, Yan P 2007 XML normal forms based on constraint-tree-based functional dependencies *Proc. of Joint 9th Asia-Pacific Web Conference and 8th International Conference on Web-Age Information Management Workshops* 348-57 Springer, Heidelberg
- [16] Vo L T H, Cao J, Rahayu W 2010 Discovering conditional functional dependencies in XML data *Proc. of the 22nd Australasian Database Conference* 143-52 Australian Computer Society, Sydney
- [17] Wang D Z, Dong L, Sarma A D, Franklin M J, Halevy A Y 2007 Functional dependency generation and applications in pay-as-you-go data integration systems *Proc. of the 12th International Workshop on the Web and Databases*, <http://www.cs.berkeley.edu/~daisyw/webdb09.pdf>
- [18] De S, Kambhampati S *Defining and mining functional dependencies in probabilistic databases* <http://arxiv.org/pdf/1005.4714v2>
- [19] Sarma A D, Ullman J, Widom J Schema design for uncertain databases *Proc. of Alberto Mendelzon Workshop on Foundations of Data Management* <http://ilpubs.stanford.edu:8090/820/>
- [20] Fan W, Libkin L 2002 On XML integrity constraints in the presence of DTDs *JACM* 49 368-406

Authors	
	<p><b>Ping Yan, born in December, 1972, Urumqi, Xinjiang Uygur Autonomous Region, China</b></p> <p><b>Current position, grades:</b> a professor in Anhui Agricultural University, PhD.</p> <p><b>University studies:</b> BSc degree in Applied Mathematics from Xinjiang University (1994), MSc degree in Applied Mathematics from Xinjiang University (1999), Ph.D degree in Applied Mathematics from Fudan University (2002)</p> <p><b>Scientific interest:</b> Her research interest fields include Neural Networks and Data management</p> <p><b>Publications:</b> more than 50 papers published in various journals and referenced conferences</p> <p><b>Experience:</b> She has teaching experience of 20 years, has completed 6 scientific research projects</p>
	<p><b>Teng Lv, born in April, 1975, Datong, Shanxi Province, China</b></p> <p><b>Current position, grades:</b> an associate professor in Army Office Academy, PhD.</p> <p><b>University studies:</b> BSc degree in Computer Science from Artillery Academy (1997), MSc degree in Computer Science from Artillery Academy (2000), Ph.D degree in Computer Science from Fudan University (2003)</p> <p><b>Scientific interest:</b> Her research interest fields include Data management</p> <p><b>Publications:</b> more than 70 papers published in various journals and referenced conferences</p> <p><b>Experience:</b> He has teaching experience of 9 years, has completed 4 scientific research projects</p>
	<p><b>Weimin He, born in December, 1973, Kunmin, Yunnan Province, China</b></p> <p><b>Current position, grades:</b> an assistant professor in UWSP, PhD.</p> <p><b>University studies:</b> BSc degree in Computer Science from Yunnan University (1995), MSc degree in Computer Science from Yunnan University (2000), Ph.D degree in Computer Science from University of Texas at Arlington (2008)</p> <p><b>Scientific interest:</b> Her research interest fields include XML Data Management, Information Retrieval and Peer-to-Peer Computing Data management</p> <p><b>Publications:</b> more than 30 papers published in various journals and referenced conferences</p> <p><b>Experience:</b> He has teaching experience of 9 years, has completed 5 scientific research projects</p>