International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.185-194
http://dx.doi.org/10.14257/ijdta.2016.9.12.17

On Uncertain Probabilistic Data Modeling

Teng Lv[1] Ping Yan[2,*] and Weimin He[3]

[1]*School of Information Engineering, Anhui Xinhua University, Hefei 230088, P.R.China*

[2*]*School of Science, Anhui Agricultural University, Hefei 230036, P.R.China*

[3]*Department of Computing and New Media Technologies, University of Wisconsin-Stevens Point, 2100 Main Street, Stevens Point, WI 54481, USA*

Lt0410@163.com, want2fly2002@163.com, whe@uwsp.edu

Abstract

Uncertainty in data is caused by various reasons including data itself, data mapping, and data policy. For data itself, data are uncertain because of various reasons. For example, data from a sensor network, Internet of Things or Radio Frequency Identification is often inaccurate and uncertain because of devices or environmental factors. For data mapping, integrated data from various heterogonous data sources is commonly uncertain because of uncertain data mapping, data inconsistency, missing data, and dirty data. For data policy, data is modified or hided for policies of data privacy and data confidentiality in an organization. But traditional deterministic data management mainly deals with deterministic data which is precise and certain, and cannot process uncertain data. Modeling uncertain data is a foundation of other technologies for further processing data, such as indexing, querying, searching, mapping, integrating, and mining data, etc. Probabilistic data models of relational databases, XML data and graph data are widely used in many applications and areas today, such as World Wide Web, semantic web, sensor networks, Internet of Things, mobile ad-hoc networks, social networks, traffic networks, biological networks, genome databases, and medical records, etc. This paper presents a survey study of different probabilistic models of uncertain data in relational databases, XML data, and graph data, respectively. The advantages and disadvantages of each kind of probabilistic modes are analyzed and compared. Further open topics of modeling uncertain probabilistic data such as semantic and computation aspects are discussed in the paper. Criteria for modeling uncertain data, such as expressive power, complexity, efficiency, extension are also proposed in the paper.

***Keywords**: data uncertainty; uncertain data model; probabilistic data model; XML; relational database; graph data*

1. Introduction

Data uncertainty is ubiquitous in many fields, such as mobile ad-hoc networks, social networks, traffic networks, biological networks, genome databases, medical records, etc. Data uncertainties are caused by many different reasons. Three major reasons are as follows: First, data itself are uncertain because of various reasons. For example, data from a sensor network, Internet of Things (IoTs) or Radio Frequency Identification (RFID) is often inaccurate and uncertain because of devices or environmental factors. Second, integrated data from various heterogonous data sources is commonly uncertain because of uncertain data mapping, data inconsistency, missing data, and dirty data. Finally, data is modified or hided for policies of data privacy and data confidentiality in an organization.

* To whom correspondence should be addressed.

ISSN: 2005-4270 IJDTA

Traditional data management mainly deals with deterministic data in which data is precise and certain, and cannot process uncertain data.

As data model is the key and foundation for other data management technologies, including indexing, querying, searching, mapping, integrating, and mining data, how to design an efficient and powerful model for uncertain data is necessary and important to other related research topic, such as data integration, data search and query, data quality and evaluation. Researchers proposed many approaches to modeling data uncertainties including rule-based models[1], fuzzy models[2], Dempster-Shafer theory of evidence-based models[3], and probability models [4]: Rule-based models apply an inference engine or semantic reasoner to infer uncertainty and imprecision based on the interaction of input and the rule base; Fuzzy models uses fuzzy technologies and tools such as fuzzy entities, attributes, relationship, aggregation, and constraints to model data uncertainty and imprecision; Dempster-Shafer theory of evidence-based models use Dempster-Shafer theory to represent data uncertainty and imprecision; and Probabilistic models represent data uncertainty by probabilistic theories, which is mostly relied on possible worlds model. As probabilistic models are widely used in many applications and in many different data format, such as structured, semi-structured, unstructured, and graph data, this paper is concentrated on probabilistic models of uncertain data.

Organizations. The rest of the paper is organized as follows. Probabilistic models in relational databases are given in Section 2. XML probabilistic data models are given in Section 3. Graph probabilistic data models are given in Section 4. We conclude the paper and point out the future directions of the topic in Section 5.

2. Probabilistic Relational Models

Probabilistic models in relational databases have been studied for more than two decades, e.g. Refs.[5,6] proposed such methods by incorporating uncertain characteristic in traditional relational models, which are mainly based on the possible worlds model[7]. In probabilistic relational models, a probabilistic database is a representation for a probability distribution over a set of possible worlds, which contain all possible instances of the database. A formal definition of possible worlds and probabilistic databases is defined as following:

Definition 1. Suppose the set of all possible database instances is $I = \{I_1, I_2, I_3, \ldots, I_n\}$, a probabilistic database Pr is a probability distribution on possible database instances I such that $\sum_{i=1}^{n} \Pr(I_i) = 1$, and a possible world PW is a set of all possible database instances such that $\Pr(I_i) > 0$.

According to the uncertain granularities, probabilistic relational models can be classified into tuple-level and attribute-level probabilistic relational models, and a function called probability distribution function (PDF) is used to assign a probability to a tuple or an attribute, respectively.

The simplest of this kind of probabilistic relational models are independent tuple-level probabilistic relational models[8], which assume that each tuple is independent to all other tuples, i.e., a tuple is existed or not does not dependent on all other tuples. As each tuple is assumed to be independent of all others, the probability of a possible world PW is given by

$$\Pr[PW] = \prod_{j \in PW} P_j \prod_{j \notin PW} (1 - P_j)$$

where $j \in PW$ if tuple t_j is in PW, and $j \notin PW$ otherwise.

In some situations, a tuple's existence may dependant on other tuples, i.e. they are not independent to other tuples. This kind of probabilistic relational models can be captured by generation rules [9]. This kind of tuple-level model is dependent tuple-level probabilistic relational models [10,11]. Suppose the m tuples are grouped as k ($k \leqslant m$) generation rules as g_1, g_2, ..., g_k according to dependency of tuples. The probability of each generation rule is given by

$$P(g_l)=\sum_{t_j \in g_l} P(t_j) \text{ where } (l=1,2,\ldots,k).$$

The probability of a possible world PW is given by

$$\Pr[PW] = \prod_{j \in PW} P_j \prod_{j \notin PW, t_j \in g_l} (1 - P(g_l))$$

where $j \in PW$ denotes t_j is in a generation rule g_l and $j \notin PW$ otherwise.

Tuple-level probabilistic relational models cannot deal with finer granularities such as uncertainty associating to attributes of relational tables. To represent a finer granularity of uncertainty, attribute-level probabilistic relational models are used, in which a probability assigns to each attribute to specify the occurrence of an attribute in a possible world. Ref.[12] used a sub-relation of a tuple to store attribute probability, Ref.[13] proposed a model of Probabilistic or-set table based on attribute-level probabilistic relational model, and Ref.[14] represented uncertain attribute values by lineage. Furthermore, Ref.[15] combined attribute-level and tuple-level probabilistic relational models into a hybrid probabilistic retaliation model, which is a probabilistic c-tables by incorporating probability distributions functions (PDF) for the values taken by their variables.

3. Probabilistic XML Models

Semi-structured data such as XML (Extensible Markup Language) models have more flexibility in structure and semantics than relational models. When considering data uncertainty modeling, the flexibilities of XML make the problem more difficult and challenging than that of relational databases.

The first kind of probabilistic XML models assumes that probability dependency only existed in local area, i.e., the probability dependency only exists between parent and child elements and we call them probabilistic XML model with local dependencies. Ref.[16] assigned a probability attribute "Prob" for each edge of a parent and its child element to indicate their local dependency. Also, an attribute "Dist" is used to as a probabilistic distribution function (PDF) to specify "Prob" values' distribution. Two distribution types of "Dist" called "mutually-exclusive" and "independent" are defined to indicate whether its sub-elements "Prob" values are mutually exclusive or independent to each others.

Example 1. The following probabilistic XML data file describes information of universities. Each university (with a probability indicated by attribute "Prob") has a specific university name and a specific president of the university. Each president has name and age both with probabilities. All probabilities are defined by a PDF "DIST". So the file conforms to the above mentioned model (probabilistic XML model with local dependencies):.

```
<universities>
 <university Prob = "0.9">
  <universityName> MY University </universityName>
   <presidentsOfUniversity>
    <Dist type = "mutually-exclusive">
     <Val Prob = "0.5">
      <name>
       <Dist>
```

```
          <Val Prob = "0.4"> Cai Y. </Val>
          <Val Prob = "0.7"> Cai P. </Val>
          <Val Prob = "0.9"> Cai Y.P. </Val>
        </Dist>
      </name>
      <age>
        <Dist type = "mutually-exclusive">
          <Val Prob = "0.6"> 35 </Val>
          <Val Prob = "0.7"> 45 </Val>
          <Val Prob = "0.9"> 55 </Val>
          <Val Prob = "0.3"> 65 </Val>
        </Dist>
      </age>
    </Val>
    <Val Prob = "0.6">
      <name>
        <Dist>
          <Val Prob = "0.5"> Zhang Y. </Val>
          <Val Prob = "0.6"> Zhang Y. </Val>
          <Val Prob = "0.7"> Zhang X.Y </Val>
        </Dist>
      </name>
      <age>
        <Dist type = "mutually-exclusive">
          <Val Prob = "0.9"> 35 </Val>
          <Val Prob = "0.5"> 36 </Val>
          <Val Prob = "0.6"> 39 </Val>
          <Val Prob = "0.2"> 40 </Val>
        </Dist>
      </age>
    </Val>
    <Val>
      ...
    </Val>
  </Dist>
 </presidentsOfUniversity>
</university>
<university>
  ...
</university>
</universities>
```

Suppose to query the president of university with name "Cai Y.P." and age 55, the probability is:

$$\Pr((\text{name} = \text{Cai Y.P.}) \wedge (\text{age} = 55) \wedge \text{presidentsOfUniversity}) = 0.9 \times 0.9 \times 0.5 \times 0.9 = 0.3645.$$

Ref.[17] proposed another probabilistic XML model by incorporating constraints in a probabilistic XML tree model, which used constraints to capture probabilistic dependencies among probabilstic XML data items. Also, constraints can include some aggregate functions such as count( ), max( ), min( ), and ratio( ). As a result, the model can be extended to give a probabilistic interpretation of such constraints.

The second kind of probabilistic XML models is probabilistic XML model with global dependencies[18], which has advantages to represent probabilistic relationship not only between ancestors-descendants (probabilistic XML tree model with local dependencies) but also between all nodes in XML data file. One method to capture such global dependencies is to use a fuzzy tree with probabilistic event variables as probabilistic conditions to nodes in XML data file. The following is such an example:

Example 2. Figure1. is a fuzzy tree with 4 event variables with corresponding probabilities as Table 1.

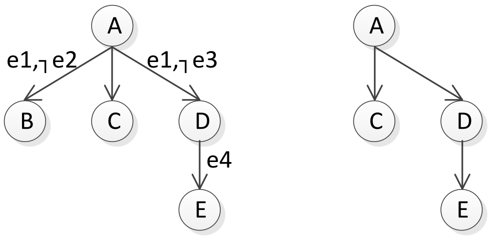

Figure 1. (a) A probabilistic XML data with global dependencies (b) a possible sub-tree T1

Table 1. Event variables and corresponding probabilities

Event variables	probability
e1	0.9
e2	0.8
e3	0.7
e4	0.6

The probability of *T1* is:

$$p(T1) = p(e_1) \times p(\neg e_3) \times p(e_4) = 0.9 \times (1 - 0.7) \times 0.6 = 0.162$$

4. Probabilistic Models in Uncertain Graph Data

Modeling, querying and mining uncertain graphs have become an increasingly important research topic[19-21] recently. Probabilistic graphs are a natural model representation in many applications, such as mobile ad-hoc networks, social networks, traffic networks, biological networks, genome databases, medical records, etc. In uncertain or probabilistic graphs, uncertainty can be categorized by three levels: (1) Edge uncertainty, i.e. the probabilistic of an edge between two nodes or vertexes is existed. (2) Node or vertex uncertainty, i.e. the probabilistic of a node is existed. (3) Attribute value of vertexes or nodes uncertainty, i.e. the probabilistic of an attribute of a given node is existed. Moreover, probabilistic graphs may be undirected or directed. So there are 4 kinds of uncertain graph modes as in Table 2. The most commonly used uncertain graph models are based on possible world models, too.

Table 2. Uncertain Graph Modes

	independent	dependent
undirected	1. undirected independent	3. undirected dependent
directed	2. directed independent	4. directed dependent

4.1. Uncertain Graph Models with Independent Probabilities

For uncertain graph models with independent probabilities, there are three types of uncertainty, such as edge uncertainty, node or vertex uncertainty, and attribute of nodes or vertexes uncertainty.

Edge uncertain graph models deal with edge uncertainty of graph data. In independent edge uncertain graph models, each edge is associated with a probability that indicates the likelihood of its existence [19, 20]. The models assume that the existence of an edge is independent of any other edges. For undirected and directed edge uncertain graph models, the process methods are much similar. The formal definition is given in Definition 2.

Definition 2. Consider an uncertain directed (or undirected) independent edge graph G = (V, E, p_E), where V is the set of vertices, E is the set of edges, p_E: E→(0, 1] is a function that assigns each edge e a probability that indicates the likelihood of e's existence. A possible graph of an edge uncertain graph G is a possible instance of G.

Example 3. Consider the following uncertain directed independent edge graph.

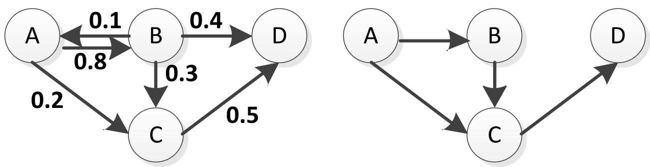

Figure 2. (a) Uncertain Directed Graph G with Probability Associated with Each Edge (b) one Possible Graph G_1 of G

Let G =(V_G, E_G) be the possible graph which is realized by sampling each edge in G according to the probability p(e) and the probability of possible graph G is:

$$P_r[G] = \prod_{e \in E_G} p(e) \prod_{e \in E \setminus E_G} (1 - p(e))$$

We can think of the probabilistic graph G as a world generator process, and each graph in G as a possible world. Figure 2 graph G has 2^6 possible graphs, and the probability of G_1 is:

$Pr[G_1]$ = p(A,B)p(A,C)p(B,C)p(C,D)(1-p(B,A))(1-p(B,D)) = 0.00036.

Example 4. Consider the following uncertain undirected independent edge graph.

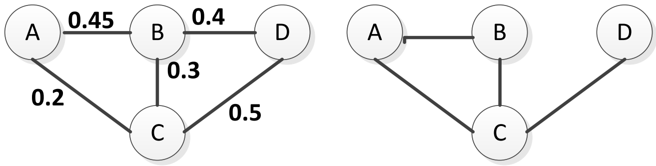

Figure 3. (a) Uncertain Undirected Graph G with Probability Associated to Each Edge (b) one Possible Graph G_2 of G

Figure 3 graph G has 2^5 possible graphs, and the probability of graph G_2 is:

$Pr[G_2]$=p(A,B)p(A,C)p(B,C)p(C,D) (1-p(B,D))=0.0081

Refs.[22,23] adopted the above model by adding a function *w*: E→(0, ∞) to associate each edge a weight *w*. Figure 2 and Figure 3 are specific cases of such model if we assume that each edge has unit-length (unit-weight). A possible graph contains a subset of edges of G, and it has a weight which is the product of the probabilities of all the edges it has.

For node uncertainty, we can extend Definition 2 to include a probability function p_V to deal with node uncertainty as given in Definition 3.

Definition 3. Consider an uncertain directed (or undirected) independent edge and node graph *G* = (V, E, p_V, p_E), where V is the set of vertices, E is the set of edges, p_V: V→(0, 1] is a function that assigns each edge vertex (node) *v* a probability that indicates the likelihood of *v*'s existence, p_E: E→(0, 1] is a function that assigns each edge *e* a probability that indicates the likelihood of *e*'s existence.

Let G =(V_G, E_G) be the possible graph which is sampled each edge in G according to the probability p_E and p_V, so G's probability Pr[G] is:

$$P_r[G] = \prod_{v \in V_G} p(v) \prod_{v \in V \setminus V_G} (1 - p(v)) \prod_{e \in E_G} p(e) \prod_{e \in E \setminus E_G} (1 - p(e))$$

Example 5. Consider the following uncertain undirected independent edge and node graph:

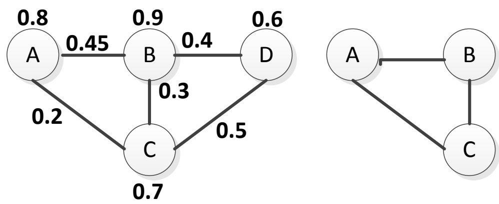

Figure 4. (a) An Uncertain Graph G with Node Probability and Edge Probability (b) A Possible Graph G_3 of G

In Figure 4, the probability of possible graph G_3 is:

$Pr[G_3]$ = p(A)p(B)p(C)(1-p(D))p(A,B)p(A,C)p(B,C)(1-p(B,D))(1-p(C,D)) = 0.00163296

Ref. [24] used such uncertain graph model to study sub-graph queries over large uncertain graphs.

For attribute value of a vertex or node uncertainty, the process methods are much similar to that of node uncertainty and edge uncertainty above mentioned. Suppose each attribute *a* associating with each node has probability function p_A: A→(0, 1] that assigns each attribute *a* in attribute set *A* a probability that indicates the likelihood of *a*'s existence. Let G = (V_G, E_G) be the possible graph which is sampled edge, node, and attribute according to the probability p_E, p_V and p_A in G, so G's probability Pr[G] is:

$$P_r[G] = \prod_{e \in E_G} p(e) \prod_{e \in E \setminus E_G} (1 - p(e)) \prod_{v \in V_G} p(v) \prod_{v \in V \setminus V_G} (1 - p(v)) \prod_{a \in V_G} p(a) \prod_{a \in V \setminus V_G} (1 - p(a))$$

4.2. Uncertain Graph Models with Dependent Probabilities

Although uncertain graph models with independent probabilities can deal with uncertainty with independent probabilities in graph data, which is applicable in many situations, such as social networks, biological networks, etc., they cannot deal with other complicated situations such as uncertainty with dependent in other applications, such as traffic network where an intersection jam may dependent on or expend to its adjacent intersections. Moreover, there may exist dependent relationship between various uncertainties. For example, node *A* maybe dependent on node *B*, edge *e(A,B)* maybe dependent on edge *e(B,C)*, and attribute a_1 of node *A* maybe dependent on attribute a_2 of node *A* in an uncertain graph.

Ref.[25] proposed a probabilistic graph model PEG (probabilistic entity graph), which defines a distribution over possible graphs at the node (entity) level. In PEG, nodes correspond to entities, node labels correspond to attribute values of nodes, and edges correspond to relations between nodes. PGM (probabilistic graphical model)[26] is used to represent probability distribution. PEG model uniformly addresses all the three kinds of uncertainties of uncertain graph, such as node uncertainty, attribute value uncertainty, and edge uncertainty. In PEG, Node uncertainty is modeled by node existence factors, attribute value uncertainty is modeled by node label factors which are probability distributions, and edge uncertainty is modeled by edge existence factors which are also probability distributions.

A PEG model can be extended to represent dependant relationships between edges and node attributes by conditional probabilities. For example, if we want to represent a case of edge existence probabilities dependent on node labels. To achieve this goal, we can replace edge existence probabilities in the PGD by some kind of conditional probabilities containing node existences event.

5. Conclusions

This paper presents a survey study of different kinds of models of uncertain data in relational databases, XML data, and graph data. We mainly discuss and review probabilistic uncertain data models as they not only are widely used in many applications and areas nowadays, but also have better tradeoffs between simplicity and expressive power.

The open problems of modeling uncertain data include both semantic and computation aspects. For semantic aspect, there is no accepted unified model for different uncertain data including relational, XML, and graph data. In real applications, such semantic problems may dependent on specific applications. For computational aspect, algorithms of deterministic data are difficult to deal with the huge computational space of possible worlds, which are usually exponential scale. Another open problem is to propose some criteria [27] for modeling uncertain data, such as expressive power, complexity, efficiency, extension, etc.

Acknowledgements

The work is supported by Colleges Nature Science Research Key Project of Anhui Province (No.KJ2015A325), Anhui Province Quality Engineering (No.2015zy073), Academic Leader Foundation of Anhui Xinhua University (No.2014XXK06), Introduction of Talents Foundation of Anhui Xinhua University, and National Natural Science Foundation of China (No. 11201002).

References

[1] T. Benouhiba and J. M. Nigro, "Uncertainty management in rule based systems application to maneuvers recognition", Proc. of 7th International Conference KES 2003, Lecture Notes in Computer Science, Springer, Oxford, UK, vol. 2773, **(2003)**.
[2] J. Galindo, A. Urrutia, and M. Piattini, "Fuzzy databases: Modeling, design, and implementation", Idea Group Publishing, **(2006)**.
[3] S. K. Lee, "Imprecise and uncertain information in databases: an evidential approach", Proc. of 8th International Conference on Data Engineering, IEEE CS Press, Tempe, Arizona, **(1992)**.
[4] W. Zhang, X. Lin, J. Pei, and Y. Zhang, "Managing uncertain data: probabilistic approaches", Proc. of the Ninth International Conference on Web-Age Information Management, IEEE CS Press, Zhangjiajie, China, **(2008)**.
[5] R. Cavallo and M. Pittarelli, "The theory of probabilistic databases. Proc. of the 13th International Conference on Very Large Data Bases, Brighton, England, **(1987)**.
[6] D. Barbara, H. G. Molina and D. Porter, "IEEE Transactions on Knowledge and Data Engineering", vol. 4, no. 5, **(1992)**.
[7] S. Abiteboul, P. Kanellakis, and G. Grahne, "ACM SIGMOD Record", vol. 16, no. 3, **(1987)**.
[8] N. Fuhr and T. Rolleke, "ACM Transactions on Information Systems", vol. 15, no. 1, **(1997)**.
[9] M. Hua, J. Pei, W. Zhang, and X. Lin, "Efficiently Answering probabilistic threshold top-k queries on uncertain data", Proc. of the 24th IEEE international conference on Data Engineering, IEEE CS Press, Cancún, México, **(2008)**.
[10] A. D. Benjelloun, S. C. Hayworth and J. Widomn, "IEEE Data Engineering Bulletin", vol. 29, no. 1, **(2006)**.
[11] T. S. Jayram, A. McGregor, S. Muthukrishnan, and E. Vee, "Estimating statistical aggregates on probabilistic data streams", Proc. of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Beijing, China, **(2007)**, pp.243-252.
[12] N. Fuhr and T. Rolleke, "A probabilistic NF2 relational algebra for imprecision in databases", http://eprints.kfupm.edu.sa/ 20186/
[13] L. V. S. Lakshmanan, N. Leone, R. Ross, and V. S. Subrahmanian, "ACM Transactions on Database Systems", vol. 22, no. 3, **(1997)**.
[14] D. Sarma, O. Benjelloun, A. Halevy, and J. Widom, "Working models for uncertain data", Proc. of the 22nd International Conference on Data Engineering, IEEE CS Press, Atlanta, Georgia, USA, **(2002)**.
[15] T. J. Green and V. Tannen, "Models for incomplete and probabilistic information", Springer Berlin Heidelberg, vol. 4254, no. 1, **(2010)**.
[16] Nierman and H. V. Jagadish, "ProTDB: probabilistic data in XML", Proc. of the 28th international conference on Very Large Data Base, Hong Kong, China, **(2002)**, pp.646-657.
[17] S. Cohen, S. B. Kimelfeld and Y. Sagiv, "ACM Transactions on Database Systems", vol. 34, no. 3, **(2009)**.
[18] S. Abiteboul and P. Senellart, "Querying and updating probabilistic information in XML", 10th International Conference on Extending Database Technology, Munich, Germany, **(2006)**.
[19] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios, "PVLDB", vol. 3, no. 1, **(2010)**.
[20] Z. Zou, H. Gao, and J. Li, "Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics", Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, USA, **(2010)**.
[21] Z. Zou, J. Li, H. Gao, and S. Zhang, "Finding top-k maximal cliques in an uncertain graph", In International Conference on Data Engineering (ICDE), Long Beach, California, USA, **(2010)**.
[22] R. Jin, L. Liu, B. Ding, and H. Wang, "Distance-constraint reachability computation in uncertain graphs. The 37th International Conference on Very Large Data Bases", 2011, Proceedings of the VLDB Endowment, Seattle, Washington, USA, vol. 4, no. 9, **(2011)**.
[23] M. Potamias, F. Bonchi, A. Gionis and G. Kollios, "K-nearest neighbors in uncertain graphs, The 36th International Conference on Very Large Data Bases", 2010, Proceedings of the VLDB Endowment, Singapore, vol. 3, no. 1, **(2010)**.
[24] Y. Yuan, G. Wang, H. Wang, and L. Chen, "Efficient Subgraph Search over Large Uncertain Graphs. The 37th International Conference on Very Large Data Bases", Proceedings of the VLDB Endowment, Seattle, Washington, USA, vol. 4, no. 11, **(2011)**.

[25] W. E. Moustaf, A. Kimmi, A. Deshpan and L. Getoor, "Subgraph pattern matching over uncertain graphs with identity linkage uncertainty", IEEE 30th International Conference on Data Engineering (ICDE 2014), IL, USA, **(2014)**.
[26] D. Koller and N. Friedman, "Probabilistic graphical models: Principles and techniques", MIT Press, **(2009)**.
[27] P. Parchas, F. Gullo, D. Papadias, and F. Bonchi, "The pursuit of good possible world: extracting representative instances of uncertain graph", International Conference on Management of Data (SIGMOD 2014), Snowbird, UT, USA, **(2014)**.

Authors

Teng Lv, born on April 1975, Datong, Shanxi Province, China
Current position, grades: an associate professor in Anhui Xinhua University, PhD.
University studies: 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)
Scientific interest: His research interest fields include Data management
Publications: more than 70 papers published in various journals and referenced conferences
Experience: He has teaching experience of 11 years, has completed 4 scientific research projects

Ping Yan, born on December 1972, Urumqi, Xinjiang Uygur Autonomous Region, China. Current position, grades: a professor in Anhui Agricultural University, PhD.
University studies: 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)
Scientific interest: Her research interest fields include neural networks and data management
Publications: more than 50 papers published in various journals and referenced conferences
Experience: She has teaching experience of 20 years, has completed 6 scientific research projects

Weimin He, born on December 1973, Kunmin, Yunnan Province, China. Current position, grades: an assistant professor in University of Wisconsin-Stevens Point, PhD.
University studies: 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)
Scientific interest: His research interest fields include XML Data Management, Information Retrieval and Peer-to-Peer Computing Data management
Publications: more than 30 papers published in various journals and referenced conferences
Experience: He has teaching experience of 10 years, has completed 5 scientific research projects

