

Fast Probabilistic Ranking under x -Relation Model

Lijun Chang, Jeffrey Xu Yu and Lu Qin
The Chinese University of Hong Kong, Hong Kong, China
{ljchang,yu,lqin}@se.cuhk.edu.hk

March 10, 2017

Abstract

The probabilistic top- k queries based on the interplay of score and probability, under the possible worlds semantic, become an important research issue that considers both score and uncertainty on the same basis. In the literature, many different probabilistic top- k queries are proposed. Almost all of them need to compute the probability of a tuple t_i to be ranked at the j -th position across the entire set of possible worlds. The cost of such computing is the dominant cost and is known as $O(kn^2)$, where n is the size of dataset. In this paper, we propose a new novel algorithm that computes such probability in $O(kn)$.

1 Introduction

Ranking is an import issue in uncertain data, and has attracted a lot of attentions recently. The probabilistic top- k queries based on the interplay of score and probability, under the possible worlds semantic, were first studied in [14]. In this paper, we show that we can significantly improve the performance for all the probabilistic top- k queries in the literature [16, 17, 6, 7, 8, 18, 3, 11] under the x -Relation model. We achieve it by proposing a new novel algorithm to reduce the dominant cost of computing probabilistic top- k queries to be $O(kn)$, which is known to be $O(kn^2)$, where n is the size of the dataset.

2 x -Relation Model and Probabilistic top-k semantics

In the x -Relation model [1, 17], an x -Relation contains a set of independent x -tuples (called generation rules in [14, 7]). An x -tuple consists of a set of mutually exclusive tuples (or called alternatives) to represent a discrete probability distribution of the possible tuples the x -tuple may take in a randomly instantiated data. In an x -tuple, each alternative t has a score $score(t)$, and a probability $p(t)$ that represents its existence probability over possible instances. In the x -Relation model, the alternatives of x -tuples are assumed to be disjoint. In the following, we denote an x -Relation as \mathcal{X} , an x -tuple as τ , and call an alternative a tuple, denoted as t .

Example 2.1: Fig. 1(a) shows an x -Relation which consists of three x -tuples, $\tau_1 = \{t_1, t_3\}$, $\tau_2 = \{t_2\}$, and $\tau_3 = \{t_4\}$. The x -tuple τ_1 indicates a probability distribution over t_1 and t_3 , with probability $p(t_1) = 0.3$ for its true content to be t_1 , with probability $p(t_3) = 0.5$ for its true content to be t_3 , and with probability $1 - p(t_1) - p(t_3) = 0.2$ for none of t_1 and t_3 to be the true content. \square

In general, an x -Relation, \mathcal{X} , is a probability distribution over a set of possible instances $\{I_1, I_2, \dots\}$. A possible instance, I_j , maintains zero or one alternative for every x -tuple $\tau \in \mathcal{X}$. The probability of an instance I_j , $\Pr(I_j)$, is the probability that x -tuples take certain or none alternatives in I_j , such that $\Pr(I_j) = \prod_{t \in I_j} p(t) \times \prod_{\tau \notin I_j} (1 - \Pr(\tau))$ where $\tau \notin I_j$ means x -tuple τ takes no alternative in I_j and $\Pr(\tau) = \sum_{t \in \tau} p(t)$. The entire set of possible worlds of an x -Relation, \mathcal{X} , denoted as $pwd(\mathcal{X})$, is the set of all the subsets $I_j (\subseteq \mathcal{X})$ with probability greater than 0 ($\Pr(I_j) > 0$).

Example 2.2: Fig. 1(b) shows the total 6 possible worlds for the x -Relation in Fig. 1(a). The possible world $\{t_1, t_2\}$ means that, τ_1 takes the alternative t_1 , τ_2 takes the alternative t_2 , and τ_3

x -tuple	tuple	score	prob
τ_1	t_1	100	0.3
	t_3	80	0.5
τ_2	t_2	90	1.0
τ_3	t_4	70	0.8

(a) x -Relation

Possible world (I)	$\Pr(I)$	top-2
$\{t_2\}$	$(1 - p(t_1) - p(t_3))p(t_2)(1 - p(t_4)) = 0.04$	t_2
$\{t_2, t_4\}$	$(1 - p(t_1) - p(t_3))p(t_2)p(t_4) = 0.16$	t_2, t_4
$\{t_1, t_2\}$	$p(t_1)p(t_2)(1 - p(t_4)) = 0.06$	t_1, t_2
$\{t_1, t_2, t_4\}$	$p(t_1)p(t_2)p(t_4) = 0.24$	t_1, t_2
$\{t_2, t_3\}$	$p(t_3)p(t_2)(1 - p(t_4)) = 0.10$	t_2, t_3
$\{t_2, t_3, t_4\}$	$p(t_3)p(t_2)p(t_4) = 0.40$	t_2, t_3

(b) Possible Worlds

Figure 1: x -Relation Data

takes none. The probability of this possible world becomes $p(t_1)p(t_2)(1 - p(t_4)) = 0.06$. Note that the sum of the probabilities of all the possible worlds is equal to 1. \square

Probabilistic top- k semantics: Several probabilistic top- k semantics have been proposed recently under the x -Relational model including Uncertain Top- k Query (U-Topk) [14, 17], Uncertain k-Ranks Query (U-kRanks) [14, 17], Global-Topk [18], Probabilistic Threshold top- k query (PT- k) [7], and the Probabilistic k top- k query (Pk-topk) [8]. The PT- k and Pk-topk are similar to the Global-Topk. Global Top- k query finds k tuples with the highest top- k probability. PT- k finds all the tuples that have top- k probability above a user-given threshold. Pk-topk finds k tuples with the highest top- k probability in a data stream environment, where every tuple is independent. All the above existing solutions except U-Topk need to compute the probability of a tuple, t_i , to be ranked at the j -th position across the entire set of possible worlds, denoted $p_{i,j}$.

Below, we introduce U-kRanks and Global Top- k with the emphasis on how $p_{i,j}$ is used. Let $p_{i,j}$ be the probability of a tuple t_i to be ranked at the j -th position across the entire set of possible worlds [14, 17].

$$p_{i,j} = \sum_{I \in \text{pwd}(\mathcal{X}), t_i = \Psi_j(I)} \Pr(I) \quad (1)$$

where $\Psi_j(I)$ denote the tuple with the j -th largest score in an instance I of the possible worlds. The answer to a U-kRanks query on an x -Relation \mathcal{X} is a vector (t_1^*, \dots, t_k^*) , where $t_j^* = \arg \max_{t_i} p_{i,j}$ for $j = 1, \dots, k$. Let $tkp(t_i)$ be the top- k probability of a tuple, t_i , which is the marginal probability that t_i is ranked top- k in the possible worlds [18].

$$tkp(t_i) = \sum_{I \in \text{pwd}(\mathcal{X}), t_i \in \text{topk}(I)} \Pr(I) = \sum_{j=1}^k p_{i,j} \quad (2)$$

where $t_i \in \text{topk}(I)$ means that the tuple t_i is ranked as one of the top- k tuples in the instance I . The answer to a Global Top- k query on an x -Relation \mathcal{X} is a set of size k , $\{t_1^*, \dots, t_k^*\}$, which satisfies $tkp(t_j^*) \geq tkp(t)$ for any $j = 1, \dots, k$ and $t \notin \{t_1^*, \dots, t_k^*\}$.

Example 2.3: The U-2Ranks query on Fig. 1(b) returns 2 tuples, t_2 ($\text{score}(t_2) = 90$) and t_3 ($\text{score}(t_3) = 80$), for t_2 is ranked top and t_3 ranked 2nd. The probability for t_2 to be ranked top is $p_{2,1} = 0.04 + 0.16 + 0.1 + 0.4 = 0.7$ and the probability for t_3 to be ranked 2nd is $p_{3,2} = 0.1 + 0.4 = 0.5$. The tuple t_1 has the highest score 100 but with a low probability 0.3, therefore, it is not a result in U-2Ranks. The Global Top-2 query returns a set of 2 tuples

$\{t_2, t_3\}$. Here $tkp(t_2) = 0.04 + 0.16 + 0.06 + 0.24 + 0.10 + 0.40 = 1.0$, because t_2 is ranked as a top-2 tuple in every instance, and $tkp(t_3) = 0.10 + 0.40 = 0.5$, because t_3 is ranked as a top-2 tuple only in two instances. Note that the results of U-kRanks and Global Top-k do not necessarily the same. \square

It is important to note that all these probabilistic ranking queries, namely, U-kRanks, Global-Topk, PT-k, and Pk-topk, need to compute the $p_{i,j}$ values for all $t_i \in \mathcal{X}$ and $j = 1, \dots, k$, and computing $p_{i,j}$ is the dominant cost in such probabilistic ranking queries.

3 $p_{i,j}$ Computing

We discuss $p_{i,j}$ computing for a given k and an x -Relation $\mathcal{X} = \{t_1, \dots, t_n\}$ sorted in the descending score order. For simplicity and without loss of generality, in the following discussions, we further assume there are no tie scores in \mathcal{X} such that $score(t_i) \neq score(t_j)$ for any $t_i \neq t_j$ in \mathcal{X} . Note that all algorithms including our algorithm to be discussed can deal with tie scores with minor modification for computing $p_{i,j}$.

[17] showed that the time complexity of computing $p_{i,j}$ for all $t_i \in \mathcal{X}$ and $j = 1, \dots, k$ is $O(kn^2)$. We introduce it in brief below.

Given an x -Relation $\mathcal{X} = \{t_1, \dots, t_n\}$ sorted in the decreasing score order. Let $\mathcal{X}_i = \{t_1, \dots, t_i\}$ denote a reduced x -Relation on the largest i tuples, together with the projected (exclusive/independent) relationship between tuples. It is obvious that $p_{i,j}$ is the same to be computed either on \mathcal{X} or \mathcal{X}_i , under the x -Relation model. Formally, let $\Pr(\tau|\mathcal{X}_i)$ be the existence probability of an x -tuple τ with respect to \mathcal{X}_i as follows.

$$\Pr(\tau|\mathcal{X}_i) = \sum_{t \in \tau, t \in \mathcal{X}_i} p(t) \quad (3)$$

Then, $\Pr(\tau) = \Pr(\tau|\mathcal{X})$.

We highlight the main idea of computing $p_{i,j}$ in $O(kn^2)$ [17] below. First, consider a special case, where every x -tuple contains only one tuple (single-alternative), or equivalently, all the tuples are independent. Then, $p_{i,j}$ is equal to the probability that a randomly generated possible world from \mathcal{X}_i contains t_i and there are j tuples in total. In other words, $p_{i,j}$ is the sum of the probabilities of the possible worlds that contain t_i and there are exactly $j-1$ tuples taken from the set $\mathcal{X}_{i-1} = \{t_1, \dots, t_{i-1}\}$. Let $r_{i,j}$ denote the probability that a randomly generated possible world from \mathcal{X}_i has exactly j tuples, then $p_{i,j} = p(t_i) \cdot r_{i-1,j-1}$. For the totally independent case, the set of all $r_{i,j}$ values can be computed efficiently by the following dynamic programming equation, in time complexity $O(kn)$.

$$r_{i,j} = \begin{cases} p(t_i) \cdot r_{i-1,j-1} + (1 - p(t_i)) \cdot r_{i-1,j}, & \text{if } i \geq j > 0; \\ (1 - p(t_i)) \cdot r_{i-1,j}, & \text{if } i > j = 0; \\ 1, & \text{if } i = j = 0; \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Second, consider the case where some x -tuples may contain multiple tuples (multi-alternative). The noticeable difference is that $p_{i,j} \neq p(t_i) \cdot r_{i-1,j-1}$ in the multi-alternative case, because an x -tuple contains multiple-alternatives that are mutually exclusive. When it needs to compute $p_{i,j}$ for a tuple t_i , the x -tuple that contains t_i may have other alternatives been computed already. It needs to remember whether an alternative of an x -tuple has already been computed in \mathcal{X}_{i-1} using a set denoted \mathcal{S} . Let $\mathcal{S} = \{\tau_1, \dots, \tau_s\}$ be the set of x -tuples, that have at least one alternative computed in \mathcal{X}_{i-1} already, with probability $\Pr(\tau_l|\mathcal{X}_{i-1})$ for $1 \leq l \leq s$ (Refer to Eq. (3)). When t_i appears and the x -tuple τ_x that contains t_i has already appeared in \mathcal{S} , it computes $p_{i,j}$ as $p_{i,j} = p(t_i) \cdot r'_{s,j-1}$. Here, $r'_{i,j-1}$, for $1 \leq i \leq s$ and $1 \leq j \leq k$, need to be recomputed based on $\mathcal{S} = \{\tau_1, \dots, \tau_s\}$ with $\Pr(\tau_x|\mathcal{X}_{i-1}) = 0$ using Eq. (4), and takes $O(s \cdot k)$ time. In the worst case, it takes $O(i \cdot k)$ to compute $p_{i,j}$ for a specific i . The time complexity to compute $p_{i,j}$ values, for $1 \leq i \leq n$ and $1 \leq j \leq k$, is $O(kn^2)$.

τ_1	$\{t_1(0.3), t_4(0.4)\}$
τ_2	$\{t_2(0.5), t_8(0.2)\}$
τ_3	$\{t_3(0.5), t_6(0.5)\}$
τ_4	$\{t_5(0.6), t_7(0.3)\}$

Table 1: Multi-alternative x -Relation

Example 3.1: Consider an x -Relation, \mathcal{X} , in Table 1 with 4 x -tuples, $\{\tau_1, \tau_2, \tau_3, \tau_4\}$ and 8 tuples $\{t_1, \dots, t_8\}$. Each x -tuple contains two tuples (alternatives). We assume $score(t_i) > score(t_j)$ if $i < j$, and give the probability of each tuple t_i , $p(t_i)$, in the corresponding parentheses. For example, τ_1 has two tuples t_1 and t_4 where $p(t_1) = 0.3$ and $p(t_4) = 0.4$. Let $k = 2$. We show how to compute $p_{i,j}$ for all tuples t_i , for $1 \leq i \leq 8$ and $j = 1, 2$.

Let all 8 tuples in \mathcal{X} be sorted in the decreasing score order, and let \mathcal{S} be the set of x -tuples that have multi-alternatives in \mathcal{X}_{i-1} . Initially, $\mathcal{X}_0 = \emptyset$, $\mathcal{S} = \emptyset$.

First, consider t_1 which is the tuple that has the largest score, and $\mathcal{S} = \emptyset$ implies that t_1 has no preceding alternatives. Because $r'_{0,0} = 1$ and $r'_{0,1} = 0$, thus $p_{1,1} = p(t_1) \cdot r'_{0,0} = 0.3$ and $p_{1,2} = p(t_1) \cdot r'_{0,1} = 0$. $\mathcal{X}_1 = \{t_1\}$. Based on Eq. (3), the current existence probability of τ_1 in \mathcal{X}_1 is $\Pr(\tau_1|\mathcal{X}_1) = p(t_1) = 0.3$. \mathcal{S} is updated to be $\mathcal{S} = \{\tau_1\}$, because the x -tuple τ_1 contains t_1 that has been computed. For simplicity, we use $\mathcal{S} = \{\tau_1(0.3)\}$ to indicate that \mathcal{S} contains τ_1 whose current existence probability is 0.3.

Second, consider the second largest score tuple t_2 , which has no preceding alternatives computed, because the x -tuple τ_2 that contains t_2 does not appear in $\mathcal{S} = \{\tau_1(0.3)\}$. Because $r'_{1,0} = 0.7$ and $r'_{1,1} = 0.3$, thus $p_{2,1} = p(t_2) \cdot r'_{1,0} = 0.35$ and $p_{2,2} = 0.15$. $\mathcal{X}_2 = \{t_1, t_2\}$. Based on Eq. (3), the current existence probability of τ_2 in \mathcal{X}_2 is $\Pr(\tau_2|\mathcal{X}_2) = p(t_2) = 0.5$. $\mathcal{S} = \{\tau_1(0.3), \tau_2(0.5)\}$.

In a similar fashion, the third largest score tuple t_3 is computed which has no preceding alternatives in \mathcal{S} . Because $r'_{2,0} = 0.35$ and $r'_{2,1} = 0.5$, thus $p_{3,1} = 0.5 \cdot 0.35 = 0.175$ and $p_{3,2} = 0.25$. $\mathcal{X}_3 = \{t_1, t_2, t_3\}$. Based on Eq. (3), the current existence probability of τ_3 in \mathcal{X}_3 is $\Pr(\tau_3|\mathcal{X}_3) = p(t_3) = 0.5$. $\mathcal{S} = \{\tau_1(0.3), \tau_2(0.5), \tau_3(0.5)\}$.

Fourth, consider the fourth largest score tuple t_4 . Note that the current $\mathcal{S} = \{\tau_1(0.3), \tau_2(0.5), \tau_3(0.5)\}$. But because tuple t_4 has a preceding alternative t_1 in x -tuple τ_1 which appears in \mathcal{S} already, the existence probability of $\Pr(\tau_1|\mathcal{X}_3) = 0$ is reset. Therefore, \mathcal{S} is updated to be $\mathcal{S} = \{\tau_1(0), \tau_2(0.5), \tau_3(0.5)\}$. In order to compute $r'_{3,0}$ and $r'_{3,1}$, all the $r'_{i,j}$ values, for $i = 1, 2$ and $j = 0, 1$, need to be recomputed as well based on the updated \mathcal{S} . Because $r'_{1,0} = 1$, $r'_{1,1} = 0$, $r'_{2,0} = 0.5$, $r'_{2,1} = 0.5$, $r'_{3,0} = 0.25$, and $r'_{3,1} = 0.5$, thus $p_{4,1} = p(t_4) \cdot r'_{3,0} = 0.1$ and $p_{4,2} = 0.2$. $\mathcal{X}_4 = \{t_1, t_2, t_3, t_4\}$. Based on Eq. (3), the current existence probability of τ_1 in \mathcal{X}_4 is $\Pr(\tau_1|\mathcal{X}_4) = p(t_1) + p(t_4) = 0.3 + 0.4 = 0.7$. Therefore, $\mathcal{S} = \{\tau_1(0.7), \tau_2(0.5), \tau_3(0.5)\}$, which will be used in the next iteration.

The same procedure repeats until all $p_{i,j}$ for all $t_i \in \mathcal{X}$ and $j = 1, 2$ are computed. \square

Note that, between consecutive computations of $p_{i,j}$ and $p_{i+1,j}$, some $r'_{s,j}$ computing cost can be shared [7, 17]. [7] also studied several heuristics to fast compute $p_{i,j}$ but in the worst case it is $O(kn^2)$.

4 A New Novel Algorithm

In this paper, we propose a novel $O(kn)$ algorithm using a newly introduced conditional probability $c_{i,j}$ given below,

$$c_{i,j} = \Pr(\text{Exactly } j \text{ tuples appear in } \{t_1, \dots, t_i\} \mid t_{i+1} \text{ appears}) \quad (5)$$

to fast compute $p_{i,j}$. Consider a general multi-alternative case. Let $\mathcal{X}_i = \{t_1, \dots, t_i\}$ be the set computed already. Now, we consider t_{i+1} , assume t_{i+1} appears. Among the tuples computed already in \mathcal{X}_i , there may exist several tuples in \mathcal{X}_i that are contained in the same x -tuple that

contains t_{i+1} . Those tuples need to be removed in order to compute for t_{i+1} , as we discussed in the previous section by setting the existence probability to be zero. Eq. (5) is the conditional probability of having exactly j tuples in $\mathcal{X}_i = \{t_1, \dots, t_i\}$ after removing those tuples in \mathcal{X}_i that are contained in the same x -tuple that contains t_{i+1} , given t_{i+1} appears. It is interesting to note that

$$\begin{aligned} p_{i,j} &= \Pr(t_i \text{ appears}) \cdot \Pr(\text{Exactly } j-1 \text{ tuples appear in } \{t_1, \dots, t_{i-1}\} \mid t_i \text{ appears}) \\ &= p(t_i) \cdot c_{i-1,j-1} \end{aligned} \quad (6)$$

And the problem becomes how to compute $c_{i,j}$ efficiently. Note that there is no obvious relationship between $c_{i,j}$ and $c_{i-1,j}$ (refer to Eq. (4)). However, we observe that there is a similar relationship between $c_{i,j}$ and $r_{i,j}$. Let τ_x be the x -tuple that contains t_{i+1} . Then, the relationship between $c_{i,j}$ and $r_{i,j}$ becomes as follows,

$$r_{i,j} = \begin{cases} (1 - \Pr(\tau_x | \mathcal{X}_i)) \cdot c_{i,j}, & \text{if } j = 0; \\ (1 - \Pr(\tau_x | \mathcal{X}_i)) \cdot c_{i,j} + \Pr(\tau_x | \mathcal{X}_i) \cdot c_{i,j-1}, & \text{if } j > 0; \end{cases} \quad (7)$$

Lemma 4.1: *Eq. (7) correctly computes $r_{i,j}$, given $c_{i,j}$.* \square

Proof Sketch: Assume that $c_{i,j}$ for $0 \leq j \leq k-1$ are correct as defined, the probability that a randomly generated possible world has exactly j tuples from \mathcal{X}_i is conditioned by the appearance of t_{i+1} . Let τ_x be the x -tuple that has t_{i+1} , and ρ denote $\Pr(\tau_x | \mathcal{X}_i)$. There are two cases.

First, t_{i+1} has no preceding alternative, equivalently $\rho = 0$. Then the two parts in the conditional probability $c_{i,j}$ are independent, $c_{i,j} = \Pr(\text{Exactly } j \text{ tuples appear in } \{t_1, \dots, t_i\})$, where the latter part of the equation is actually $r_{i,j}$. Hence, Eq. (7) correctly computes $r_{i,j}$, given that $c_{i,j}$ are correct.

Second, t_{i+1} has some preceding alternatives, equivalently $\rho > 0$. Assume that $\mathcal{S} = \{\tau_1, \dots, \tau_s, \tau_x\}$ is the set of x -tuples that have alternatives appearing in $\mathcal{X}_i = \{t_1, \dots, t_i\}$, where $\Pr(\tau_l | \mathcal{X}_i) > 0$ for all $\tau_l \in \mathcal{S}$. Then $c_{i,j}$ is the probability that a randomly generated possible world from $\{\tau_1, \dots, \tau_s\}$ ($= \mathcal{S} \setminus \{\tau_x\}$) has exactly j x -tuples, and $r_{i,j}$ is the probability that a randomly generated possible world from $\{\tau_1, \dots, \tau_s, \tau_x\}$ has exactly j x -tuples. Hence, Eq. (7) is correct based on the same idea shown in Eq. (4). \square

Given $c_{i,j}$ we can compute $r_{i,j}$ using Eq. (7). The reverse also holds such that, given $r_{i,j}$, we can compute $c_{i,j}$ correctly by the system of linear equations defined in Eq. (7). A general system of linear equations with n equations and n variables needs time $O(n^3)$. But the system of linear equations defined by Eq. (7) has a special form, there are only two diagonals of the coefficient matrix which are non-zero, so it can be solved in $O(n)$ time [9]. In our problem, there are k linear equations with k variables, it can be solved in time $O(k)$, using $c_{i,0} = r_{i,0}/(1 - \rho)$ and $c_{i,j} = (r_{i,j} - \rho \cdot c_{i,j-1})/(1 - \rho)$ where $\rho = \Pr(\tau_x | \mathcal{X}_i)$, for $1 \leq j \leq k-1$. Note that $0 < \Pr(\tau_x | \mathcal{X}_i) < 1$. In addition, given $c_{i,j}$, $r_{i+1,j}$ can also be computed using Eq. (7), by replacing $\Pr(\tau_x | \mathcal{X}_i)$ with $\Pr(\tau_x | \mathcal{X}_{i+1})$, where τ_x is the x -tuple that contains t_{i+1} .

The algorithm to compute $r_{i,j}$ and $p_{i,j}$ values for a tuple t_i is shown in Algorithm 1. It takes three inputs, namely, the tuple t_i , the $r_{i-1,j}$ values, and a set of x -tuples, $\mathcal{S} = \{\tau_1, \dots, \tau_s\}$, that have been computed with their probability $\Pr(\tau_l) = \Pr(\tau_l | \mathcal{X}_{i-1})$. It first computes ρ (line 1-2). Then, it computes the $c_{i-1,j}$ values by solving a system of linear equations defined by Eq. (7) (line 3-5), and computes the $p_{i,j}$ values (line 6). In line 7-10, it computes the $r_{i,j}$ values using Eq. (7). Finally, it updates the probability $\Pr(\tau_x)$ (line 11-14). Note that, in our algorithm, the only values needed to compute $p_{i,j}$ values are $r_{i-1,j}$ values and $\Pr(\tau_x | \mathcal{X}_{i-1})$.

Theorem 4.1: *Algorithm 1 correctly computes the $p_{i,j}$ values with time complexity of $O(k)$.* \square

Proof Sketch: It is obvious from the discussions above. \square

In order to compute all $p_{i,j}$, we enumerate all $t_i \in \mathcal{X}$, which is sorted in the descending order score, such as $score(t_i) > score(t_j)$ if $i < j$ as given below.

Algorithm 1 CondProb($\mathcal{S}, \mathcal{R}_{i-1}, t_i$)

Input: the probability for x -tuples $\mathcal{S} = \{\tau_1(\text{Pr}(\tau_1)), \dots, \tau_s(\text{Pr}(\tau_s))\}$
 $\mathcal{R}_{i-1} = \{r_{i-1,0}, \dots, r_{i-1,k-1}\}$ and a tuple t_i .

Output: $r_{i,j-1}$ and $p_{i,j}$, for $1 \leq j \leq k$.

```
1: Let  $\tau_x$  be the  $x$ -tuple that has  $t_i$ ;
2:  $\rho \leftarrow \text{Pr}(\tau_x)$  if  $\tau_x(\text{Pr}(\tau_x))$  appears in  $\mathcal{S}$  otherwise 0;
   // compute  $c_{i-1,j}$  and  $p_{i,j}$  for  $0 \leq j \leq k-1$ 
3:  $c_{i-1,0} \leftarrow r_{i-1,0}/(1-\rho)$ ;
4: for  $j \leftarrow 1$  to  $k-1$  do
5:    $c_{i-1,j} \leftarrow (r_{i-1,j} - \rho \cdot c_{i-1,j-1})/(1-\rho)$ ;
6:  $p_{i,j} \leftarrow p(t_i) \cdot c_{i-1,j-1}$ , for  $1 \leq j \leq k$ ;
   // compute  $r_{i,j}$  for  $0 \leq j \leq k-1$ 
7:  $\rho \leftarrow \rho + p(t_i)$ ;
8:  $r_{i,0} \leftarrow (1-\rho) \cdot c_{i-1,0}$ ;
9: for  $j \leftarrow 1$  to  $k-1$  do
10:   $r_{i,j} \leftarrow (1-\rho) \cdot c_{i-1,j} + \rho \cdot c_{i-1,j-1}$ ;
11: if  $\tau_x \notin \mathcal{S}$  then
12:   $\mathcal{S} \leftarrow \mathcal{S} \cup \{\tau_x(\rho)\}$ ;
13: else
14:  update  $\mathcal{S}$  by changing  $\text{Pr}(\tau_x)$  to be  $\rho$ ;
15: return ( $\mathcal{S}, \{r_{i,0}, \dots, r_{i,k-1}\}, \{p_{i,1}, \dots, p_{i,k}\}$ );
```

```
1: Let  $\mathcal{S} = \emptyset$ ;
2: Let  $\mathcal{R}_0 = \{r_{0,0}, r_{0,1}, \dots, r_{0,k-1}\}$  where  $r_{0,j}$ , for  $0 \leq j \leq k-1$ , are computed;
3: for  $i = 1$  to  $n$  do
4:  ( $\mathcal{S}, \mathcal{R}_i, \mathcal{P}_i$ )  $\leftarrow$  CondProb( $\mathcal{S}, \mathcal{R}_{i-1}, t_i$ );
5:  output  $\mathcal{P}_i = \{p_{i,1}, p_{i,2}, \dots, p_{i,k}\}$ ;
```

It is obvious that the time complexity to compute all $p_{i,j}$ is $O(kn)$.

Fig. 2(a) illustrates the existing $O(kn^2)$ approach to compute $r'_{i,j}$ in the stage i based on the stage $i-1$. Note that the stage i is the i -iteration to compute for the i -th largest score tuple in \mathcal{X}_i . On the left side in the stage $i-1$ and the stage i , it indicates that some x -tuple (marked by \bullet) contains several tuples (alternatives). On the other hand, Fig. 2(b) illustrates our $O(kn)$ approach to compute $r_{i,j}$, using $c_{i,j}$, in the stage i based on the stage $i-1$. The shaded parts in Fig. 2(a)(b) indicate the equations needed to compute, and the difference between the two shaded regions confirms the significant cost saving of our approach.

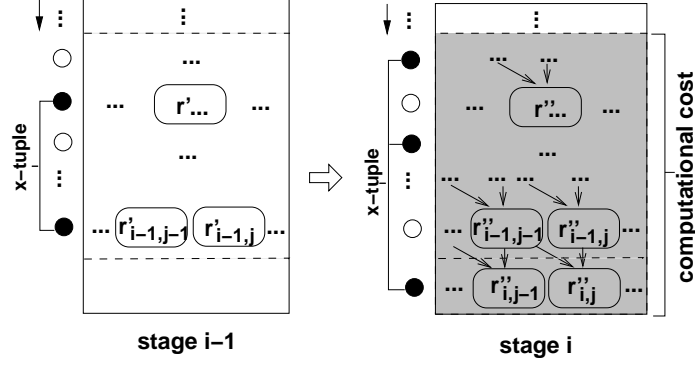
Example 4.1: Consider the example x -Relation in Table 1. We show the steps of our algorithm to compute $p_{i,j}$. Let $k = 2$. We denote the sequence of x -tuples that have been scanned as \mathcal{S} . Initially, $\mathcal{X}_0 = \emptyset$, $\mathcal{S} = \emptyset$, $r_{0,0} = 1$ and $r_{0,1} = 0$.

First, consider t_1 which is the largest score tuple. It has no preceding alternatives, $\text{Pr}(\tau_1) = 0$, $c_{0,0} = 1$ and $c_{0,1} = 0$. Then, $p_{1,1} = p(t_1) \cdot c_{0,0} = 0.3$ and $p_{1,2} = 0$. After computing t_1 , $\mathcal{X}_1 = \{t_1\}$, $\mathcal{S} = \{\tau_1(0.3)\}$, and we have $r_{1,0} = (1 - \text{Pr}(\tau_1)) \cdot c_{0,0} = 0.7$ and $r_{1,1} = (1 - \text{Pr}(\tau_1)) \cdot c_{0,1} + \text{Pr}(\tau_1) \cdot c_{0,0} = 0.3$.

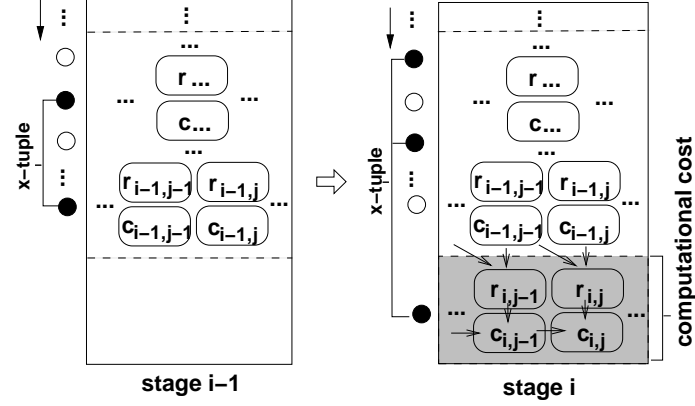
The second largest score tuple t_2 has no preceding alternatives, $\text{Pr}(\tau_2) = 0$, $c_{1,0} = r_{1,0} = 0.7$ and $c_{1,1} = 0.3$. Then, $p_{2,1} = p(t_2) \cdot c_{1,0} = 0.35$ and $p_{2,2} = 0.15$. After computing t_2 , $\mathcal{X}_2 = \{t_1, t_2\}$, $\mathcal{S} = \{\tau_1(0.3), \tau_2(0.5)\}$, and in addition we have $r_{2,0} = 0.35$ and $r_{2,1} = 0.5$.

The third largest score tuple t_3 has no preceding alternatives $\text{Pr}(\tau_3) = 0$, $c_{2,0} = 0.35$ and $c_{2,1} = 0.5$. Then, $p_{3,1} = p(t_3) \cdot c_{2,0} = 0.175$ and $p_{3,2} = 0.25$. After computing t_3 , $\mathcal{X}_3 = \{t_1, t_2, t_3\}$, $\mathcal{S} = \{\tau_1(0.3), \tau_2(0.5), \tau_3(0.5)\}$, and in addition we have $r_{3,0} = 0.175$ and $r_{3,1} = 0.425$.

The fourth largest score tuple t_4 has a preceding alternative t_1 that is contained in x -tuple τ_1 which appears in \mathcal{S} . Therefore, $\rho = \text{Pr}(\tau_1|\mathcal{X}_3) = 0.3$, $c_{3,0} = r_{3,0}/(1 - \rho) = 0.25$,



(a) The Existing $O(kn^2)$ Approach



(b) Our New $O(kn)$ Approach

Figure 2: Computational Cost
 $c_{3,1} = (r_{3,1} - \rho \cdot c_{3,0}) / (1 - \rho) = 0.5$, $p_{4,1} = p(t_4) \cdot c_{3,0} = 0.1$ and $p_{4,2} = 0.2$. After computing t_4 , $\mathcal{X}_4 = \{t_1, t_2, t_3, t_4\}$, $\mathcal{S} = \{\tau_1(0.7), \tau_2(0.5), \tau_3(0.5)\}$, and in addition we have $r_{4,0} = 0.075$ and $r_{4,1} = 0.325$.

The same procedure repeats until all $p_{i,j}$ for all $t_i \in \mathcal{X}$ and $j = 1, 2$ are computed. \square

5 Top-k Generator

Algorithm 1 returns the set of $p_{i,j}$, which can be used to compute the top- k probability of a tuple, e.g. $tkp(t_i) = \sum_{j=1}^k p_{i,j}$. A naive way to get the top- k result is to first compute the top- k probabilities for all tuples, then report the top- k tuples with respect to the top- k probability. In the following, we will first discuss an upper bound, and then propose an early stop condition, which avoids to retrieve all the tuples.

Lemma 5.1: Let $\{t_1, \dots, t_i, \dots\}$ be the order we scan the tuples, or equivalently it is the decreasing score order, and $r_{i,j}$ is defined as above. Then $tkp(t_{i+1}) \leq \sum_{j=1}^k r_{i,j}$, for all $i \geq 1$. This upper bound is also tight for an arbitrary sequence of tuples. \square

Proof Sketch: Let τ_x be the x -tuple that have t_{i+1} , and $p = \Pr(\tau_x | \mathcal{X}_i)$. Note that p may be zero, or equivalently t_{i+1} has no preceding alternative. By Eq. (7), sum up the $r_{i,j}$'s, $\sum_{j=0}^{k-1} r_{i,j} = \sum_{j=0}^{k-1} c_{i,j} - p \cdot c_{i,k-1}$. We have

$$\begin{aligned}
tkp(t_{i+1}) &= \sum_{j=1}^k p_{i+1,j} \\
&= p(t_{i+1}) \cdot \sum_{j=0}^{k-1} c_{i,j} \\
&\leq (1-p) \cdot \sum_{j=0}^{k-1} c_{i,j} \\
&= \sum_{j=0}^{k-1} c_{i,j} - p \cdot \sum_{j=0}^{k-1} c_{i,j} \\
&\leq \sum_{j=0}^{k-1} c_{i,j} - p \cdot c_{i,k-1} \\
&= \sum_{j=0}^{k-1} r_{i,j}
\end{aligned}$$

where the third inequality holds because $\Pr(\tau_x|\mathcal{X}_i) + p(t_{i+1}) \leq 1$, as t_{i+1} is an alternative of x -tuple τ_x . So $tkp(t_{i+1}) \leq \sum_{j=1}^k r_{i,j}$. When $p(t_{i+1}) = 1$, $\Pr(\tau_x|\mathcal{X}_i) = 0$, the above inequalities hold with equality, and therefore $tkp(t_{i+1}) = \sum_{j=1}^k r_{i,j}$. Hence this upper bound is tight. \square

Lemma 5.2: $\sum_{j=0}^{k-1} r_{i,j}$ are in decreasing order, e.g. $\sum_{j=0}^{k-1} r_{i,j} \geq \sum_{j=0}^{k-1} r_{i+1,j}$, for any $i \geq 1$. \square

Proof Sketch: There are two cases, t_{i+1} has preceding alternatives or not.

First, if t_{i+1} does not have preceding alternatives, then $r_{i+1,j}$ can be computed by Eq. (4). Summing up $r_{i+1,j}$, we have $\sum_{j=0}^{k-1} r_{i+1,j} = \sum_{j=0}^{k-1} r_{i,j} - p(t_{i+1})r_{i,k-1} \leq \sum_{j=0}^{k-1} r_{i,j}$. Second, if t_{i+1} has preceding alternatives, assuming t_{i+1} is in the x -tuple τ_x , then $\Pr(\tau_x|\mathcal{X}_i) > 0$. Assume that $\{\tau_1, \dots, \tau_x, \dots, \tau_s\}$ is the set of x -tuples that have alternatives in $\{t_1, \dots, t_i\}$, with probability $\Pr(\tau|\mathcal{X}_i)$. Then $\{\tau_1, \dots, \tau_x, \dots, \tau_s\}$ is also the set of x -tuples that have alternatives in $\{t_1, \dots, t_i, t_{i+1}\}$, and their probability is $\Pr(\tau|\mathcal{X}_{i+1})$, with $\Pr(\tau|\mathcal{X}_{i+1}) = \Pr(\tau|\mathcal{X}_i)$ for all x -tuple τ except τ_x , which has $\Pr(\tau_x|\mathcal{X}_{i+1}) > \Pr(\tau_x|\mathcal{X}_i)$. Let $r_{*,j}$ be the probability that a random generated possible world from $\{\tau_1, \dots, \tau_x, \dots, \tau_s\}/\tau_x$, with probabilities $\Pr(\tau|\mathcal{X}_i)$, has exactly j x -tuples. The relationship between $r_{i,j}$ and $r_{*,j}$, or between $r_{i+1,j}$ and $r_{*,j}$, is the same as Eq. (4) or Eq. (7). Then $\sum_{j=0}^{k-1} r_{i,j} = \sum_{j=0}^{k-1} r_{*,j} - \Pr(\tau_x|\mathcal{X}_i) \cdot r_{*,k-1}$, and $\sum_{j=0}^{k-1} r_{i+1,j} = \sum_{j=0}^{k-1} r_{*,j} - \Pr(\tau_x|\mathcal{X}_{i+1}) \cdot r_{*,k-1}$. So $\sum_{j=0}^{k-1} r_{i+1,j} \leq \sum_{j=0}^{k-1} r_{i,j}$, as $\Pr(\tau_x|\mathcal{X}_{i+1}) > \Pr(\tau_x|\mathcal{X}_i)$. \square

Theorem 5.1: If all the top- k probabilities of the current top- k result, e.g. from the set $\{t_1, \dots, t_i\}$, are greater than or equal to $\sum_{j=0}^{k-1} r_{i,j}$, then we can stop, and guaranty that any potential results in $\{t_{i+1}, \dots, t_N\}$ can not be in the top- k result. \square

With Theorem 5.1, we can develop an algorithm to compute the top- k tuples with respect to their top- k probabilities, which is shown in Algorithm 2. It initializes in line 1-5, and *upBound* denotes the upper bound of the top- k probabilities of the remaining tuples (line 5). While the stop condition is not satisfied (line 6), it retrieves the next largest score tuple (line 7), computes its top- k probability, inserts it into the top- k set (line 8-10), and update the upper bound (line 11). The top- k set is maintained as a min-heap with size of k , *top-k*[k].*tkp* (line 6) is the minimum top- k probability in the min-heap. When inserting a new tuple associate with its top- k probability into min-heap, if its top- k probability is smaller than that at the top of the min-heap, we do not need to insert it. Otherwise, we replace the top tuple of the min-heap with the new tuple and update the heap structure.

Theorem 5.2: Algorithm 2 correctly returns the top- k tuples with highest top- k probabilities. The top- k generator takes time $O(n(k + \log(k)))$, where n is scan depth, or equivalently the number of calls *Next*(). \square

Proof Sketch: The correctness directly follows from the above discussions.

The time complexity of $O(n(k + \log(k)))$ does not take *Next*() into consideration. The initial of line 1-5 takes constant time. Each call of *Prob*() (Algorithm 1) takes $O(k)$ time, based on Theorem 4.1. Line 9, 11 take time $O(k)$. Line 10 takes time $O(\log(k))$, due to the min-heap of size k . Line 6-11 are only executed n times, so the total time complexity is $O(n(k + \log(k)))$. \square

Algorithm 2 Top- k (k)

Input: an integer k , specify the top- k value,**Output:** top- k tuples.

```
1: Let  $\{\tau_1, \dots, \tau_m\}$  be the set of all the  $x$ -tuples;
2: Initialize  $\Pr(\tau_i) \leftarrow 0$ , for  $1 \leq i \leq m$ ;
3:  $top-k \leftarrow \emptyset$ ;
4: Initialize  $r_0 = 1$  and  $r_j = 0$  for  $1 \leq j \leq k-1$ ;
5:  $upBound \leftarrow \sum_{j=0}^{k-1} r_j$ ;
6: while  $top-k[k].tkp < upBound$  do
7:    $t \leftarrow Next()$ ;
8:    $r_j, p_j \leftarrow Prob(\{\Pr(\tau_1), \dots, \Pr(\tau_m)\}, \{r_0, \dots, r_{k-1}\}, t)$ ;
9:    $tkp(t) \leftarrow \sum_{j=1}^k p_j$ ;
10:  Insert  $t$  into  $top-k$ ;
11:   $upBound \leftarrow \sum_{j=0}^{k-1} r_j$ ;
12: return  $top-k$ ;
```

Parameter	Range	Default
$mem-p$	0.1, 0.3, 0.5, 0.7, 0.9	0.5
p	0.1, 0.3, 0.5, 0.7, 0.9	0.3
k	200, 400, 600, 800, 1000	200
$ rule $	5, 10, 15, 20, 25	10
$\#tuple$	20000, 40000, 60000, 80000, 100000	20000
$\#rule$	500, 1000, 1500, 2000, 2500	2000

Table 2: Parameters and Default Values

6 Experiment

We have implemented our algorithm in Visual C++. We compare our CondProb algorithm, denoted CP, for computing $p_{i,j}$, with the heuristics proposed in [7] which are RC (rule-tuple compression only), RC+AR (RC with aggressive reordering), and RC+LR (RC with lazy reordering). The heuristics proposed can improve the efficiency but they are algorithms in $O(kn^2)$, where n is the number of tuples and k is the top- k value. The executable code and data generator used in [7] are downloadable¹. We use exactly the same synthetic dataset as used in [7], which is also included in the package.

The parameters and default values are shown in Table 2. Here, $mem-p$ is the expectation of the membership probability of tuples, p is the threshold specifying the minimum top- k probability of the result tuples returned, k is the top- k value, $|rule|$ is the average number of tuples in a rule (x -tuple), $\#tuple$ is the total number of tuples, and $\#rule$ is the total number of rules (x -tuples).

The experimental results are shown in Fig. 3. In all figures, the shape of the curves for all the four algorithms are all similar, our CP algorithm is 3,000 times faster than RC+LR on average, and 30,000 times faster than RC on average.

7 Related work

Uncertain data has received increasing attention recently, most of them represent the uncertainty as probability values, also called probabilistic data. Many probabilistic data model and systems have been proposed, for example, Trio system [1], MystiQ system [5], MayBMS system [2].

In the literature, several works study computing the top- k results by the interplay of score and probability, based on the possible worlds semantic. U-Topk and U-kRanks queries are first proposed in [14] on a general uncertain data model. [16, 17] improve the performance of the U-Topk and U-kRanks queries using a dynamic programming approach, under an x -Relation model,

¹<http://www.cs.sfu.ca/~jpei/Software/PTKLib.rar>

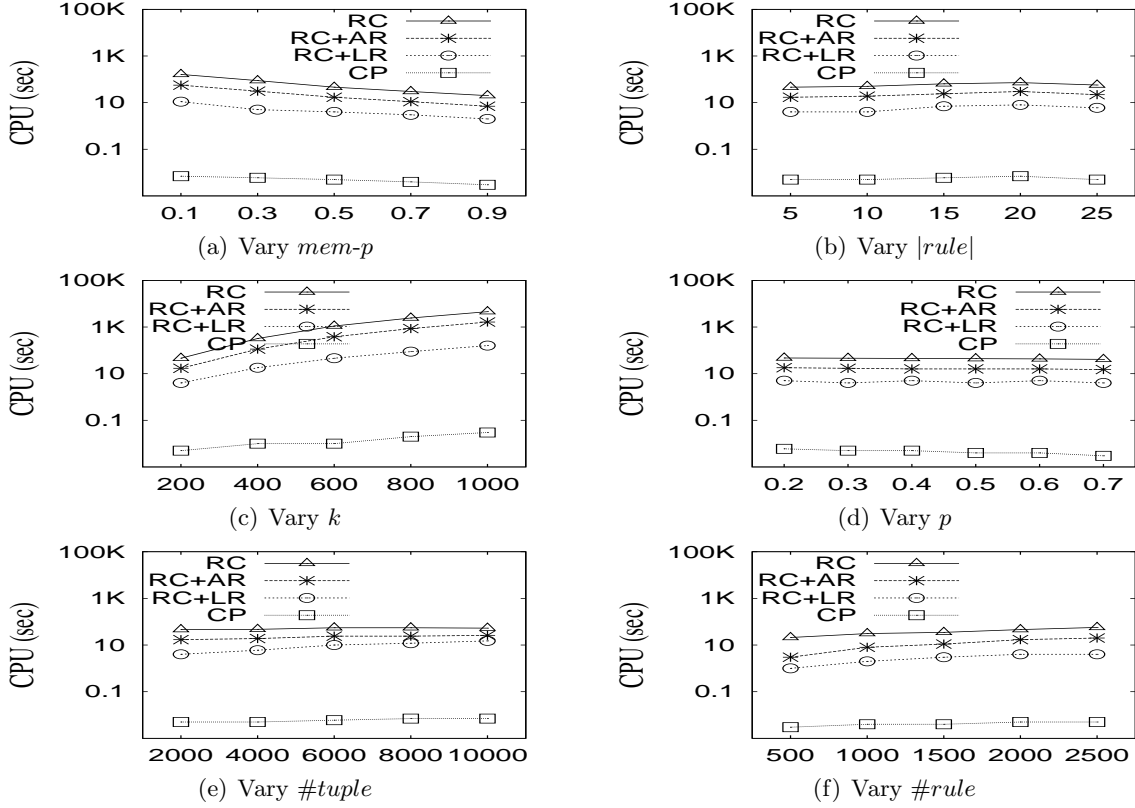


Figure 3: Computing $p_{i,j}$

by utilizing the independent and mutually exclusive relationship between tuples. [6, 7] define the PT- k query, and propose three heuristic approaches to answer the PT- k queries. In [16, 17, 6, 7], to answer a U- k Ranks or PT- k query, they all need to compute $p_{i,j}$, the probability that tuple t_i ranks at the j -th position in possible worlds, for $1 \leq i \leq n$ and $1 \leq j \leq k$, with the time complexity $O(kn^2)$. [8] adapt the U-Topk/U- k Ranks/Global-Topk (Global-Topk [18] is the same as Pk-topk in [8]) queries in a uncertain stream environment under a sliding-window model, and design both space- and time-efficient synopses to continuously monitor the top- k results. But, [8] only consider the single-alternative case, or in other words, all tuples are independent. [3, 11] also need to compute the $p_{i,j}$ values, running the probabilistic ranking in a middleware to answer ranking spatial queries on uncertain spatial data. [15] discusses aggregate queries.

There are also works that find the top- k results based on the probability only. In [13], Re et al. find the k most probable answers for a given general SQL query. In this scenario, each answer has a probability instead of a score, which intuitively represents the confidence of its existence, ranking is only based on probabilities. They use Monte Carlo simulations to get the top- k results efficiently, as in general it is #P-complete to get the existence probability [5]. [12, 10, 4] retrieve k objects from a uncertain spatial database, that have the highest probability to be a skyline point or nearest neighbor.

8 Conclusion

The probabilistic top- k queries based on the interplay of score and probability, under the possible worlds semantic, become an important research issue that considers both score and uncertainty on the same basis. In the literature, many different probabilistic top- k queries are proposed. In the x -Relational model, an x -tuple consists of a set of mutually exclusive tuples to represent a discrete probability distribution of the possible tuples in a randomly instantiated data. Almost all of them need to compute the probability of a tuple t_i to be ranked at the j -th position across the entire set of possible worlds. We call it $p_{i,j}$ computing. The cost of computing $p_{i,j}$ is the dominant cost and is known as $O(kn^2)$, where n is the size of dataset. In this paper, we

proposed a new novel algorithm that computes such probability efficiently based on conditional probability and the system of linear equations. We proved the correctness of our approach, and showed that the time complexity is $O(kn)$. We confirmed the efficiency by comparing our approach with the up-to-date heuristics and found that our approach can be at least 3,000 times faster.

References

- [1] P. Agrawal, O. Benjelloun, A. D. Sarma, C. Hayworth, S. U. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *Proc. of VLDB'06*, 2006.
- [2] L. Antova, T. Jansen, C. Koch, and D. Olteanu. Fast and simple relational processing of uncertain data. In *Proc. of ICDE'08*, 2008.
- [3] T. Bernecker, H.-P. Kriegel, and M. Renz. ProUD: Probabilistic ranking in uncertain databases. In *Proc. of SSDBM'08*, 2008.
- [4] G. Beskales, M. A. Soliman, and I. F. Ilyas. Efficient search for the top-k probable nearest neighbors in uncertain databases. *PVLDB*, 1(1), 2008.
- [5] N. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB J.*, 16(4), 2007.
- [6] M. Hua, J. Pei, W. Zhang, and X. Lin. Efficiently answering probabilistic threshold top-k queries on uncertain data. In *Proc. of ICDE'08*, 2008.
- [7] M. Hua, J. Pei, W. Zhang, and X. Lin. Ranking queries on uncertain data: A probabilistic threshold approach. In *Proc. of SIGMOD'08*, 2008.
- [8] C. Jin, K. Yi, L. Chen, J. X. Yu, and X. Lin. Sliding-window top-k queries on uncertain streams. In *Proc. of VLDB'08*, 2008.
- [9] D. C. Lay. *Linear Algebra and Its Applications (3rd Edition)*. Addison Wesley, July 2002.
- [10] X. Lian and L. Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In *Proc. of SIGMOD'08*, 2008.
- [11] X. Lian and L. Chen. Probabilistic ranked queries in uncertain databases. In *Proc. of EDBT'08*, 2008.
- [12] J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In *Proc. of VLDB'07*, 2007.
- [13] C. Re, N. N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In *Proc. of ICDE'07*, 2007.
- [14] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang. Top-k query processing in uncertain databases. In *Proc. of ICDE'07*, 2007.
- [15] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang. Probabilistic top- and ranking-aggregate queries. *ACM Trans. Database Syst.*, 33(3), 2008.
- [16] K. Yi, F. Li, G. Kollios, and D. Srivastava. Efficient processing of top-k queries in uncertain databases. In *Proc. of ICDE'08*, 2008.
- [17] K. Yi, F. Li, G. Kollios, and D. Srivastava. Efficient processing of top-k queries in uncertain databases with x-Relations. *IEEE Trans. Knowl. Data Eng.*, 20(12), 2008.
- [18] X. Zhang and J. Chomicki. On the semantics and evaluation of top-k queries in probabilistic databases. In *Proc. of DBRank'08*, 2008.