On the Semantics and Evaluation of Top-k Queries in Probabilistic Databases

Xi Zhang Jan Chomicki

SUNY at Buffalo

September 23, 2008

Outline

1. Motivating Examples
2. Semantics of Top-k Queries in Probabilistic Databases
3. Efficient Algorithms for Global-Top$k$ Semantics
4. Global-Top$k$ under General Scoring Functions
5. Related Work
6. Conclusion and Future Work

Outline

1. Motivating Examples
2. Semantics of Top-k Queries in Probabilistic Databases
3. Efficient Algorithms for Global-Top$k$ Semantics
4. Global-Top$k$ under General Scoring Functions
5. Related Work
6. Conclusion and Future Work

Example #1: Smart Environment

- Question "Who were the two visitors in the lab last Sat night?"
- Data
  - Biometric data from sensors
  - Historical statistics

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

Query

A top-k query, where $k = 2$, over the above probabilistic relation.

Example #2: Sensor Network in a Habitat

- Question "What is the temperature of the warmest spot?"
- Sensor reading data
  - At a given time, only one *correct* reading per sensor

Data from a habitat (snapshot)

	Temp.(°F)	Prob
C_1	22	0.6
	10	0.4
C_2	25	0.1
	15	0.6

Query

A top-k query, where $k = 1$, over the above probabilistic relation.

Probabilistic Relation - Definition

A **probabilistic relation** is a triplet $R^p = \langle R, p, \mathcal{C} \rangle$, where

- R is a *support deterministic relation*
- p is a *probability function* $p : R \mapsto (0, 1]$
- $\mathcal{C}$ is a *partition* of R, such that

$$\forall C_i \in \mathcal{C}, \sum_{t \in C_i} p(t) \leq 1$$

Simple v.s. General probabilistic relation

R^p is *simple* iff each *part* contains exactly one tuple, i.e. all tuples are **independent**.

Possible Worlds

Possible Worlds

A *probabilistic relation* represents a set of **possible worlds**, each of which is one possible state of the relation.

Possible Worlds

Possible Worlds

A *probabilistic relation* represents a set of **possible worlds**, each of which is one possible state of the relation.

Determinism

Each **possible world** of a *probabilistic relation* is a *deterministic relation*.

Outline

Top-k Queries in Deterministic Databases

- R is a deterministic relation

Scoring Function s

$$s : R \mapsto \mathbb{R}$$

Ties

- Allow ties in general
- No ties if function s is *injective*

Induced Order over R

For any $t_1, t_2 \in R$

$$t_1 \succ t_2 \text{ iff } s(t_1) > s(t_2)$$

- A *weak order* in general
- A *total order* if function s is *injective*

Top-k Queries in Deterministic Databases (cont.)

Top-k Queries in Deterministic Databases - A Top-k Set

Nondeterministically return a set of k tuples with the highest scores

- *Multiple* such sets in general
- *Unique* if function s is *injective*

Winners v.s. Losers

- A tuple t is a *winner* iff it belongs to the union of all top-k sets
- Otherwise, it is a *loser*

Top-k Queries in Probabilistic Databases

In a probabilistic database...

Need to extend the semantics of deterministic databases

Top-k Queries in Probabilistic Databases

In a probabilistic database...

Need to extend the semantics of deterministic databases

Example

In a top-2 query, which two tuples to return?

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

What is a “Good” Semantics?

- R^p is a probabilistic relation
- S is a top-k answer of R^p

Property #1 - Exact k

When R^p is sufficiently large, $|S| = k$.

Property #2 - Faithfulness

For any t_1, t_2 in R, if

- both the score and the probability of t_1 are higher than those of t_2
- $t_2 \in S$

then $t_1 \in S$.

What is a "Good" Semantics? (Cont.)

Property #3 - Stability

- Raising the score/probability of a winner will not turn it into a loser
- Lowering the score/probability of a loser will not turn it into a winner

Possible Worlds

Smart Lab

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

Possible Worlds:

	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
	Ø	Aidan	Bob	Chris	Aidan Bob	Aidan Chris	Bob Chris	Aidan Bob Chris
Pr	0.042	0.018	0.378	0.028	0.162	0.012	0.252	0.108

$$(Simple)\ Pr(W) = \prod_{t \in W} p(t) \prod_{t \notin W} (1 - p(t))$$

Possible Worlds

Smart Lab

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

Possible Worlds:

	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
	Ø	Aidan	Bob	Chris	Aidan, Bob	Aidan, Chris	Bob, Chris	Aidan, Bob, Chris
Pr	0.042	0.018	0.378	0.028	0.162	0.012	0.252	0.108

$$(Simple)\ Pr(W) = \prod_{t \in W} p(t) \prod_{t \notin W} (1 - p(t))$$

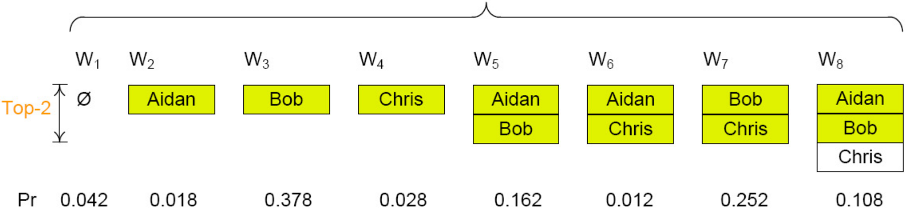
$$(General)\ Pr(W) = \prod_{t \in W} p(t) \prod_{C_i \in \mathcal{C}, C_i \cap W = \emptyset} (1 - \sum_{t \in C_i} p(t))$$

Global-Top$k$ Probability - Example

Smart Lab

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

Possible Worlds:



- $P_{k,s}(Chris) = Pr(W_4) + Pr(W_6) + Pr(W_7) = 0.292$

Global-Top$k$ Probability - Example

Smart Lab

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

Possible Worlds:

	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
Top-2	Ø	Aidan	Bob	Chris	Aidan Bob	Aidan Chris	Bob Chris	Aidan Bob
								Chris
Pr	0.042	0.018	0.378	0.028	0.162	0.012	0.252	0.108

- $P_{k,s}(Bob) = Pr(W_3) + Pr(W_5) + Pr(W_7) + Pr(W_8) = 0.9$
- $P_{k,s}(Aidan) = Pr(W_2) + Pr(W_5) + Pr(W_6) + Pr(W_8) = 0.3$
- $P_{k,s}(Chris) = Pr(W_4) + Pr(W_6) + Pr(W_7) = 0.292$

Global-Top$k$ Probability - Definition

Given

- a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$
- an integer $k \geq 0$
- an *injective* scoring function s over R^p

The **Global-Top$k$ probability** of a tuple t in R, denoted by $P_{k,s}^{R^p}(t)$, is the sum of the probabilities of all possible worlds of R^p whose top-k answer contains t.

$$P_{k,s}^{R^p}(t) = \sum_{\substack{W \in pwd(R^p) \\ t \in top_{k,s}(W)}} Pr(W).$$

Global-Top$k$ Semantics

Global-Top$k$ Semantics

Return a set of k tuples with the highest Global-Top$k$ probability

Global-Top$k$ Semantics

Global-Top$k$ Semantics

Return a set of k tuples with the highest Global-Top$k$ probability

Example: Smart Lab

- $P_{k,s}(Bob) = 0.9$
- $P_{k,s}(Aidan) = 0.3$
- $P_{k,s}(Chris) = 0.292$

Answer $\{Bob, Aidan\}$

Global-Top$k$ Semantics

Global-Top$k$ Semantics

Return a set of k tuples with the highest Global-Top$k$ probability

Example: Smart Lab

- $P_{k,s}(Bob) = 0.9$
- $P_{k,s}(Aidan) = 0.3$
- $P_{k,s}(Chris) = 0.292$

Answer $\{Bob, Aidan\}$

Properties

Global-Top$k$ satisfies *Exact-k* and *Stability* in *simple* and *general* probabilistic relations, and satisfies *Faithfulness* in *simple* probabilistic relations.

Outline

Simple Probabilistic Relations

Given

- a *simple* probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$
- an integer $k \geq 0$
- an *injective* scoring function s over R^p

Assume tuples in R are ordered in the decreasing order of their scores, i.e.

$$R = \{t_1, t_2, \ldots, t_n\}, \text{ and } s(t_1) > s(t_2) > \ldots > s(t_n)$$

Observation #1

For any possible world W, t_i is in the top-k answer of $W \Leftrightarrow$

- W contains t
- W contains **at most** $k - 1$ tuples from $\{t_1, t_2, \ldots, t_{i-1}\}$

Simple Probabilistic Relations - Dynamic Programming

Basic Recurrence

$$q(k,i) = \begin{cases} 0 & k = 0 \\ p(t_i) & 1 \le i \le k \\ (q(k,i-1)\frac{\bar{p}(t_{i-1})}{p(t_{i-1})} + q(k-1,i-1))p(t_i) & \text{otherwise} \end{cases}$$

where $q(k,i) = P_{k,s}(t_i)$ and $\bar{p}(t_{i-1}) = 1 - p(t_{i-1})$.

General Probabilistic Relations

- Basic recurrence no longer holds due to the fact that some tuples are *exclusive*.

Observation #1'

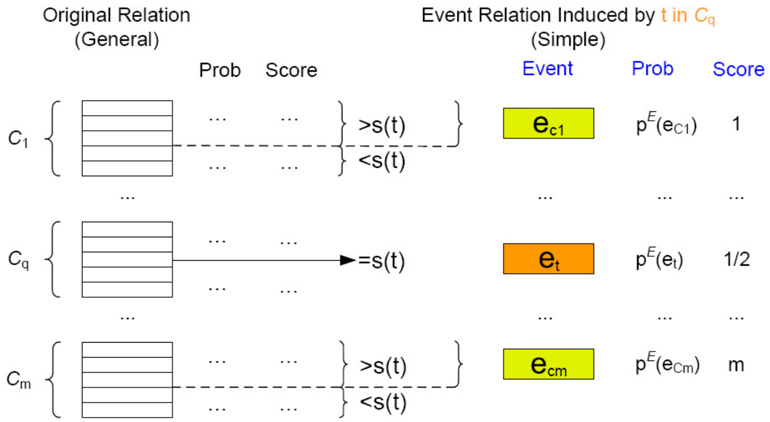
For any possible world W, t_i is in the top-k answer of W $\Leftrightarrow$

- W contains t_i
- W contains **at most** $k-1$ tuples with score higher than that of t_i
- those "better" tuples and t are all from different parts in the partition

Observation #2

All *parts* C_i in the partition $\mathcal{C}$ of R^p are *independent*.

A Reduction to Simple Probabilistic Relations



- Each tuple t in R induces an *event relation*
- event e_t="tuple t is present"
- event e_{C_i}="a tuple from part C_i with score higher than that of t is present", where $C_i \neq C_q$

A Reduction to Simple Probabilistic Relations

Probability p^E of event e_{C_i}

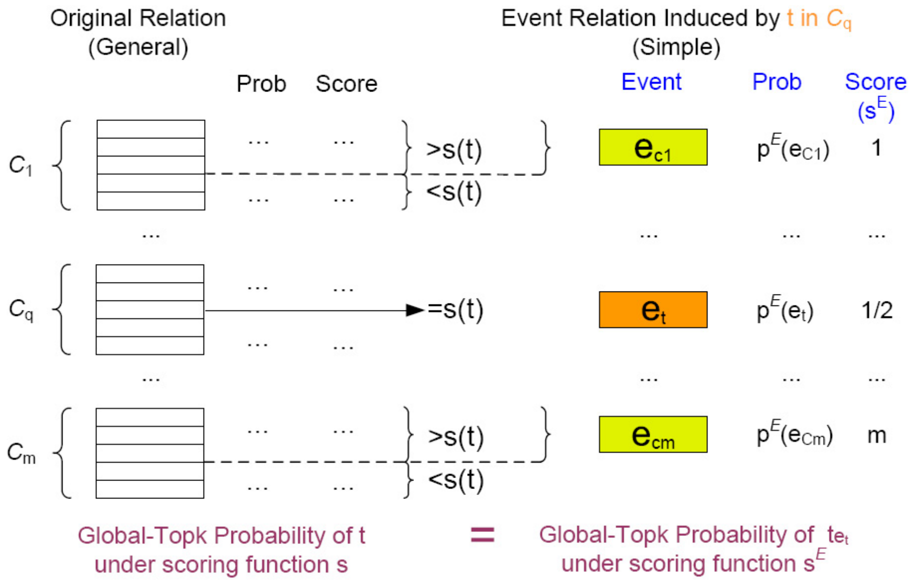
- Rule 1:

$$t \in C_i \text{ then } p^E(e_t) = p(t);$$

- Rule 2:

$$t \notin C_i \text{ then } p^E(e_{C_i}) = \sum_{\substack{t' \in C_i \\ s(t') > s(t)}} p(t')$$

A Reduction to Simple Probabilistic Relations



Example - Sensor Network in a Habitat

Data from a habitat R^p
(General)

	Temp. °F (scoring function s)	Prob
C_1	22	0.6
	10	0.4
C_2	25	0.1
	15	0.6

Event Relation E^p
Induced by t=(Temp: 15) in C_2
(Simple)

Event	Prob	Score s^E
e_{c1}	$p^E(e_{C1}) = 0.6$	1
e_t	$p^E(e_t) = 0.6$	1/2

Global-Topk Probability of t under scoring function s = Global-Topk Probability of te_t under scoring function s^E

- $p^E(e_t) = p(t) = 0.6$
- $p^E(e_{C_1}) = \sum_{t' \in C_1 \wedge s(t') > s(t)} p(t') = p((\text{Temp:}22)) = 0.6$
- $P^{R^p}_{k,s}(t) = P^{E^p}_{k,s^E}(t_{e_t}) = 0.24$

Reduction Theorem

Given

- a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$
- an injective scoring function s over R

For any $t \in R^p$, the Global-Top$k$ probability of t equals the Global-Top$k$ probability of t_{e_t}

$$P^{R^p}_{k,s}(t) = P^{E^p}_{k,s^E}(t_{e_t}).$$

where

- *induced event relation* $E^p = \langle E, p^E, \mathcal{C}^E \rangle$
- *injective scoring function* $s^E : E \to \mathbb{R}, s^E(t_{e_t}) = \frac{1}{2}$ and $s^E(t_{e_{C_i}}) = i$

Global-Top$k$ in a General Probabilistic Relation

We have the following algorithm based on the *Reduction Theorem*

- For each tuple t in R, calculate the Global-Top$k$ probability $P_{k,s}(t)$ via a polynomial reduction to its induced event relation
- Return a set of k tuples with the highest Global-Top$k$ probability

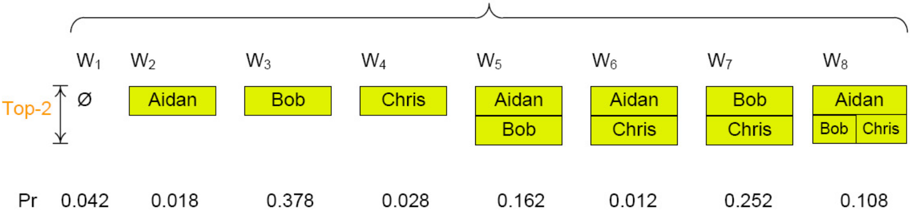
Outline

Global-Top$k$ Probability in the Presence of Ties

Smart Lab

Name	Biometric Score (Face, Voice, ...)	Prob. of Sat Nights
Aidan	0.65	0.3
Bob	0.45	0.9
Chris	0.45	0.4

Possible Worlds:



- $P_{k,s}(Bob) = Pr(W_3) + Pr(W_5) + Pr(W_7) + \frac{1}{2}Pr(W_8) = 0.846$
- $P_{k,s}(Aidan) = Pr(W_2) + Pr(W_5) + Pr(W_6) + Pr(W_8) = 0.3$
- $P_{k,s}(Chris) = Pr(W_4) + Pr(W_6) + Pr(W_7) + \frac{1}{2}Pr(W_8) = 0.346$

Global-Top$k$ Probability under General Scoring Functions

Given

- a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$
- an integer $k \geq 0$
- a (*general*) scoring function s over R^p

Global-Top$k$ probability of a tuple t in R,

$$P_{k,s}^{R^p}(t) = \sum_{\substack{W \in pwd(R^p) \\ t \in top_{k,s}(W)}} \alpha(t, W) Pr(W).$$

Global-Top$k$ Probability under General Scoring Functions

Given

- a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$
- an integer $k \geq 0$
- a (*general*) scoring function s over R^p

Global-Top$k$ probability of a tuple t in R,

$$P^{R^p}_{k,s}(t) = \sum_{\substack{W \in pwd(R^p) \\ t \in top_{k,s}(W)}} \alpha(t, W) Pr(W).$$

Equal Allocation Policy for Ties

Let $a = |\{t' \in W | s(t') > s(t)\}|$ and $b = |\{t' \in W | s(t') = s(t)\}|$

$$\alpha(t, W) = \begin{cases} 1 & \text{if } a < k \text{ and } a + b \leq k \\ \frac{k-a}{b} & \text{if } a < k \text{ and } a + b > k \end{cases}$$

Computing Global-Top$k$ under General Scoring Functions

Challenge

- How to integrate the allocation policy with the Global-Top$k$ algorithm?
- Dynamic Programming alone will *not* work

Computing Global-Top$k$ under General Scoring Functions

Challenge

- How to integrate the allocation policy with the Global-Top$k$ algorithm?
- Dynamic Programming alone will *not* work

Simple Probabilistic Relations

Algorithm = Dynamic Programming + Enumeration

Computing Global-Top$k$ under General Scoring Functions

Challenge

- How to integrate the allocation policy with the Global-Top$k$ algorithm?
- Dynamic Programming alone will *not* work

Simple Probabilistic Relations

Algorithm = Dynamic Programming + Enumeration

General Probabilistic Relations

Algorithm = Dynamic Programming + Enumeration + Reduction

Simple Probabilistic Relations

- Let $A = |\{t'|t' \in R, s(t') > s(t)\}|, B = |\{t'|t' \in R, \wedge s(t') = s(t)\}|$
- Worlds contributing to $P^{Rp}_{k,s}(t) =$

	Worlds with 0 tuples from A and $\leq k$ tuples from B including t
$+$	Worlds with 1 tuples from A and $\leq k-1$ tuples from B including t
$+$	.
$+$	.
$+$	.
$+$	Worlds with k tuples from A and ≤ 0 tuples from B including t

Simple Probabilistic Relations (Cont.)

- For every $0 \leq j \leq k$

$$\sum_{\substack{W:j \text{ tuples from } A \\ \text{and } k-j \text{ tuples from } B \text{ including } t}} \alpha(t, W) Pr(W)$$

$$= \left(\sum_{W:j \text{ tuples from } A} Pr(W) \right) \cdot Prob(e_j) \text{ (by independence of tuples)}$$

where *event* $e_j =$"Global-Top$(k-j)$ set from B includes t"

Simple Probabilistic Relations (Cont.)

- For every $0 \leq j \leq k$

$$\sum_{\substack{W: j \text{ tuples from } A \\ \text{and } k-j \text{ tuples from } B \text{ including } t}} \alpha(t, W) Pr(W)$$

$$= \left(\sum_{W: j \text{ tuples from } A} Pr(W) \right) \cdot Prob(e_j) \text{ (by independence of tuples)}$$

where *event* e_j = "Global-Top$(k-j)$ set from B includes t"

- Optimization: sharing of dynamic programming tables

Complexity

Prob. DB	Injective Scoring Fn	General Scoring Fn
Simple	$O(kn)$	$O(k \max(n, m_{max}^2))$
General	$O(kn^2)$	$O(kn^2)$

where m_{max} is the maximal number of tying tuples in R

Outline

Related Work

- Soliman, Ilyas & Chang, ICDE'07
  - Formulate the problem of top-k queries in probabilistic databases
  - Two semantics: U-Top$k$ and U-kRanks
    - U-Top$k$: return the most probable top-k answer set that belongs to possible worlds
    - U-kRanks: for $i = 1, 2, \ldots, k$, return the most probable i^{th}-ranked tuples across all possible worlds
  - Scoring function: injective
- Yi, Li, Kollios & Srivastava, ICDE'08
  - Significantly improve *time* and *space* for U-Top$k$ and U-kRanks
- Hua, Pei, Zhang & Lin, ICDE'08
  - Independently develop a semantics equivalent to Global-Top$k$ under injective scoring functions

Semantics Comparison

Property Satisfaction

Semantics	Exact-k	Faithfulness	Stability
Global-Top$k$	$\surd$	$\surd/\times$*	$\surd$
U-Top$k$	$\times$	$\surd/\times$*	$\surd$
U-kRanks	$\times$	$\times$	$\times$

*"$\surd$" if the database is *simple*, "$\times$" if the database is *general*.

Complexity (under injective scoring functions)

Semantics	Simple Prob. DB	General Prob. DB
Global-Top$k$	$O(kn)$	$O(kn^2)$
U-Top$k$	$O(n \log k)$	$O(n \log k)$
U-kRanks	$O(kn)$	$O(kn^2)$

Generalize Other Semantics to General Scoring Functions

- U-Top$k$ and U-kRanks (Soliman et al. 2007)
  - Both based on possible worlds
  - Possible to be generalized based on *nondeterminism* and *allocation policy*
  - Current algorithms are *not* directly applicable
- PT-k (Hua et al. 2008)
  - Under general scoring functions, the semantics and the algorithm are not compatible

Outline

Conclusion

- Three intuitive semantic properties for top-k queries in probabilistic databases
- Global-Top$k$ semantics, which satisfies all three properties in *simple* probabilistic databases and two properties in general ones
- Efficient algorithms for Global-Top$k$ semantics in simple/general probabilistic databases under injective scoring functions
- Generalization of Global-Top$k$ semantics and computation to general scoring functions

Future Work

- More complete picture of properties of top-k semantics
- Asymmetry of *score* and *probability*
  - This work: *ordinal* score + *cardinal* probability
  - Open: *cardinal* score + *cardinal* probability
- Consider *preference strength* in the semantics
- Relationship among tuples
  - This work: *independent/exclusive* relationship
  - Open: more *complex* relationship
- Other uncertain database models