Semantics of Ranking Queries for Probabilistic Data and Expected Ranks

Graham Cormode
AT&T Labs

Feifei Li
FSU

Ke Yi
HKUST

Uncertain, uncertain, uncertain ... (Probabilistic, probabilistic, probabilistic ...)

- Uncertain data management has become a really ~~hot~~ important area of active research

Uncertain, uncertain, uncertain ... (Probabilistic, probabilistic, probabilistic ...)

- Uncertain data management has become a really ~~hot~~ important area of active research
- Many types of real-world data are uncertain: sensor readings, multimedia, data integration, ...

The model

- An uncertain database is a probability distribution of certain (deterministic) databases

$\{t_1, t_2, t_3\}$, prob $= 0.1$

$\{t_1, t_3, t_4\}$, prob $= 0.2$

$\{t_2, t_3\}$, prob $= 0.5$

...

$\{t_3, t_4\}$, prob $= 0.05$

The model

- An uncertain database is a probability distribution of certain (deterministic) databases

$\{t_1, t_2, t_3\}$, prob $= 0.1$

$\{t_1, t_3, t_4\}$, prob $= 0.2$

$\{t_2, t_3\}$, prob $= 0.5$

...

$\{t_3, t_4\}$, prob $= 0.05$

possible worlds

The model

- An uncertain database is a probability distribution of certain (deterministic) databases

A succinct uncertain model
(e.g. an x-relation)

$\{t_1, t_2, t_3\}$, prob $= 0.1$

$\{t_1, t_3, t_4\}$, prob $= 0.2$

$\{t_2, t_3\}$, prob $= 0.5$

...

$\{t_3, t_4\}$, prob $= 0.05$

possible worlds

The model

- An uncertain database is a probability distribution of certain (deterministic) databases

A succinct uncertain model
(e.g. an x-relation)

tuples	$p(t)$
t_1	0.4
t_2	0.5
t_3	1
t_4	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

Query algorithm

$\{t_1, t_2, t_3\}$, prob $= 0.1$

$\{t_1, t_3, t_4\}$, prob $= 0.2$

$\{t_2, t_3\}$, prob $= 0.5$

...

$\{t_3, t_4\}$, prob $= 0.05$

possible worlds

Query semantics
(definition)

Query semantics for uncertain data

- A principled approach: probabilistic thresholding [Dalvi, Suciu, 04]

Query semantics for uncertain data

- A principled approach: probabilistic thresholding [Dalvi, Suciu, 04]
  - What's the probability that a tuple appears in the query result?

Query semantics for uncertain data

- A principled approach: probabilistic thresholding [Dalvi, Suciu, 04]
  - What's the probability that a tuple appears in the query result?
  - Set a threshold τ, return all tuples t such that $\Pr[t \text{ is in the query results}] \geq \tau$

Query semantics for uncertain data

- A principled approach: probabilistic thresholding [Dalvi, Suciu, 04]
  - What's the probability that a tuple appears in the query result?
  - Set a threshold τ, return all tuples t such that $\Pr[t \text{ is in the query results}] \geq \tau$
  - Or, rank all the tuples by this probability [Re, Dalvi, Suciu, 07]

Query semantics for uncertain data

- A principled approach: probabilistic thresholding [Dalvi, Suciu, 04]
  - What's the probability that a tuple appears in the query result?
  - Set a threshold τ, return all tuples t such that $\Pr[t \text{ is in the query results}] \geq \tau$
  - Or, rank all the tuples by this probability [Re, Dalvi, Suciu, 07]
- Has been applied to a variety of queries
  - Selection, projection, join [Dalvi, Suciu, 04]
  - Range queries [Tao, Cheng, Xiao, Ngai, Kao, Prabhakar, 05]
  - Frequent items [Zhang, Li, Yi, 08]

Ranking (top-k) queries (with scores)

- Very useful queries: rank by importance, rank by similarity, rank by relevance, k-nearest neighbors

Ranking (top-k) queries (with scores)

- Very useful queries: rank by importance, rank by similarity, rank by relevance, k-nearest neighbors
  - U-top$k$: [Soliman, Ilyas, Chang, 07], [Yi, Li, Srivastava, Kollios, 08]
  - U-kRanks: [Soliman, Ilyas, Chang, 07], [Lian, Chen, 08]
  - PT-k: [Hua, Pei, Zhang, Lin, 08]
  - Global-top$k$: [Zhang, Chomicki, 08]

Ranking (top-k) queries (with scores)

- Very useful queries: rank by importance, rank by similarity, rank by relevance, k-nearest neighbors
  - U-top$k$: [Soliman, Ilyas, Chang, 07], [Yi, Li, Srivastava, Kollios, 08]
  - U-kRanks: [Soliman, Ilyas, Chang, 07], [Lian, Chen, 08]
  - PT-k: [Hua, Pei, Zhang, Lin, 08]
  - Global-top$k$: [Zhang, Chomicki, 08]
  - Expected ranks: [this work]

Uncertain data models (with a score attribute)

- Two possibilities: uncertain tuples and uncertain scores

Uncertain data models (with a score attribute)

- Two possibilities: uncertain tuples and uncertain scores

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Uncertain data models (with a score attribute)

- Two possibilities: uncertain tuples and uncertain scores

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

tuple-level uncertainty

Uncertain data models (with a score attribute)

- Two possibilities: uncertain tuples and uncertain scores

tuples	score
t_1	$\{(100, 0.4), (70, 0.6)\}$
t_2	$\{(92, 0.6), (80, 0.4)\}$
t_3	$\{(85, 1)\}$

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	$0.4 \times 0.6 \times 1 = 0.24$
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	$0.4 \times 0.4 \times 1 = 0.16$
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	$0.6 \times 0.6 \times 1 = 0.36$
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	$0.6 \times 0.4 \times 1 = 0.24$

Uncertain data models (with a score attribute)

- Two possibilities: uncertain tuples and uncertain scores

tuples	score
t_1	$\{(100, 0.4), (70, 0.6)\}$
t_2	$\{(92, 0.6), (80, 0.4)\}$
t_3	$\{(85, 1)\}$

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	$0.4 \times 0.6 \times 1 = 0.24$
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	$0.4 \times 0.4 \times 1 = 0.16$
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	$0.6 \times 0.6 \times 1 = 0.36$
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	$0.6 \times 0.4 \times 1 = 0.24$

attribute-level (score-level) uncertainty

PT-k: [Hua, Pei, Zhang, Lin, 08]

- Following the probabilistic thresholding approach

PT-k: [Hua, Pei, Zhang, Lin, 08]

- Following the probabilistic thresholding approach
- For given k, p, consider $\Pr[t \text{ is in the top-}k \text{ list}]$, and return all tuples with this probabilty $\geq p$

PT-k: [Hua, Pei, Zhang, Lin, 08]

- Following the probabilistic thresholding approach
- For given k, p, consider $\Pr[t \text{ is in the top-}k \text{ list}]$, and return all tuples with this probabilty $\geq p$

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Suppose $k = 2, p = 0.5$

$\Pr[t_1] = 0.4$
$\Pr[t_2] = 0.5$
$\Pr[t_3] = 0.8$
$\Pr[t_4] = 0.3$

Global-top$k$: [Zhang, Chomicki, 08]

Global-top$k$: [Zhang, Chomicki, 08]

- For given k, consider $\Pr[t$ is in the top-k list$]$, and return the k tuples having the highest probabilties

Global-top$k$: [Zhang, Chomicki, 08]

- For given k, consider $\Pr[t$ is in the top-k list$]$, and return the k tuples having the highest probabilties

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Suppose $k = 2$, ~~$p = 0.5$~~

$\Pr[t_1] = 0.4$
$\Pr[t_2] = 0.5$
$\Pr[t_3] = 0.8$
$\Pr[t_4] = 0.3$

U-top$k$: [Soliman, Ilyas, Chang, 07]

U-top$k$: [Soliman, Ilyas, Chang, 07]

- For given k, consider $\Pr[T$ is the top-k list$]$, where T is a k-set, and return the T having the highest probabilty

U-top$k$: [Soliman, Ilyas, Chang, 07]

- For given k, consider $\Pr[T$ is the top-k list$]$, where T is a k-set, and return the T having the highest probabilty

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

Suppose $k = 2$

$\Pr[\{t_1, t_2\}] = 0.2$
$\Pr[\{t_1, t_3\}] = 0.2$
$\Pr[\{t_2, t_3\}] = 0.3$
$\Pr[\{t_3, t_4\}] = 0.3$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

U-kRanks: [Soliman, Ilyas, Chang, 07]

U-kRanks: [Soliman, Ilyas, Chang, 07]

- For any rank i, consider $\Pr[t \text{ is ranked at } i]$, and return the t having the highest probabilty for $i = 1, \ldots, k$, respectively

U-kRanks: [Soliman, Ilyas, Chang, 07]

- For any rank i, consider $\Pr[t \text{ is ranked at } i]$, and return the t having the highest probabilty for $i = 1, \ldots, k$, respectively

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

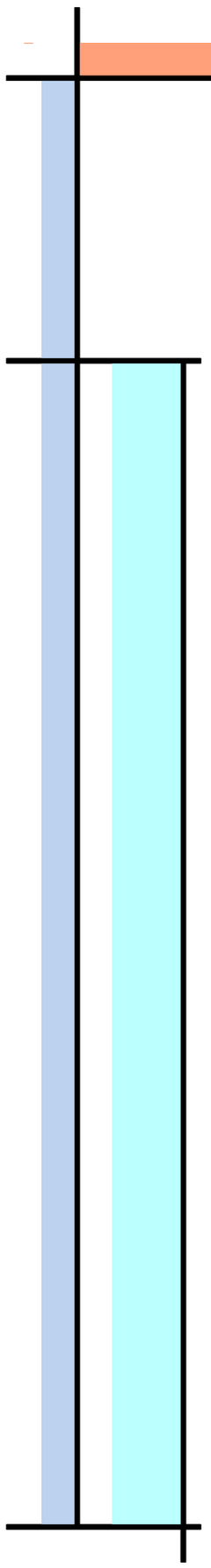
world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

At rank $i = 1$

$\Pr[t_1] = 0.4$
$\Pr[t_2] = 0.3$
$\Pr[t_3] = 0.3$

At rank $i = 2$

$\Pr[t_2] = 0.3$
$\Pr[t_3] = 0.5$
$\Pr[t_4] = 0.3$



In stead of jumping forward and formulate yet another definition, we step back and first think about what properties that a ranking query should satisfy.

Property one: *Value-invariance*

The scores only determine the relative behavior of the tuples: changing the score values without altering the relative ordering should not change the top-k

Property one: *Value-invariance*

The scores only determine the relative behavior of the tuples: changing the score values without altering the relative ordering should not change the top-k

Ranking method	Value-invariant
U-top$k$	✓
U-kRanks	✓
PT-k	✓
Global-top$k$	✓

Property two: *Exact-k*

The top-k list should contain exactly k items.
Also proposed in [Zhang, Chomicki, 08]

Property two: *Exact-k*

The top-k list should contain exactly k items.
Also proposed in [Zhang, Chomicki, 08]

Ranking method	Exact-k
U-top$k$	weak
U-kRanks	✓
PT-k	×
Global-top$k$	✓

Property two: *Exact-k*

The top-k list should contain exactly k items.
Also proposed in [Zhang, Chomicki, 08]

Ranking method	Exact-k
U-top$k$	weak
U-kRanks	✓
PT-k	×
Global-top$k$	✓

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Property two: *Exact-k*

The top-k list should contain exactly k items.
Also proposed in [Zhang, Chomicki, 08]

Ranking method	Exact-k
U-top$k$	weak
U-kRanks	✓
PT-k	×
Global-top$k$	✓

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Suppose $k = 2, p = 0.5$

$$\Pr[t_1] = 0.4$$
$$\Pr[t_2] = 0.5$$
$$\Pr[t_3] = 0.8$$
$$\Pr[t_4] = 0.3$$

Property three: *Unique-rank*

Within the top-k, each reported item should be assigned exactly one position: the same item should not be listed multiple times within the top-k

Property three: *Unique-rank*

Within the top-k, each reported item should be assigned exactly one position: the same item should not be listed multiple times within the top-k

Ranking method	Unique-rank
U-top$k$	✓
U-kRanks	$\times$
PT-k	✓
Global-top$k$	✓

Property three: *Unique-rank*

Within the top-k, each reported item should be assigned exactly one position: the same item should not be listed multiple times within the top-k

Ranking method	Unique-rank
U-top$k$	✓
U-kRanks	×
PT-k	✓
Global-top$k$	✓

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	0.24
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	0.16
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	0.36
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	0.24

Property three: *Unique-rank*

Within the top-k, each reported item should be assigned exactly one position: the same item should not be listed multiple times within the top-k

Ranking method	Unique-rank
U-top$k$	✓
U-kRanks	$\times$
PT-k	✓
Global-top$k$	✓

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	0.24
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	0.16
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	0.36
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	0.24

At rank 1:

$\Pr[t_1] = 0.4$

$\Pr[t_2] = 0.36$

$\Pr[t_3] = 0.24$

At rank 3:

$\Pr[t_1] = 0.6$

$\Pr[t_2] = 0.16$

$\Pr[t_3] = 0.24$

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Ranking method	Unique-rank
U-top$k$	$\times$
U-kRanks	✓
PT-k	weak
Global-top$k$	$\times$

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Ranking method	Unique-rank
U-top$k$	$\times$
U-kRanks	✓
PT-k	weak
Global-top$k$	$\times$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

U-top$k$ with $k = 1$

$\Pr[\{t_1\}] = 0.4$

$\Pr[\{t_2\}] = 0.3$

$\Pr[\{t_3\}] = 0.3$

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Ranking method	Unique-rank
U-top$k$	$\times$
U-kRanks	✓
PT-k	weak
Global-top$k$	$\times$

U-top$k$ with $k = 1$

$\Pr[\{t_1\}] = 0.4$

$\Pr[\{t_2\}] = 0.3$

$\Pr[\{t_3\}] = 0.3$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

U-top$k$ with $k = 2$

$\Pr[\{t_1, t_2\}] = 0.2$

$\Pr[\{t_1, t_3\}] = 0.2$

$\Pr[\{t_2, t_3\}] = 0.3$

$\Pr[\{t_3, t_4\}] = 0.3$

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Ranking method	Unique-rank
U-top$k$	$\times$
U-kRanks	✓
PT-k	weak
Global-top$k$	$\times$

Global-top$k$ with $k = 1$

$\Pr[t_1] = 0.4$

$\Pr[t_2] = 0.3$

$\Pr[t_3] = 0.3$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Property four: *Containment*

The top-$(k+1)$ list should contain all items in the top-k

Ranking method	Unique-rank
U-top$k$	$\times$
U-kRanks	✓
PT-k	weak
Global-top$k$	$\times$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

Global-top$k$ with $k = 1$

$\Pr[t_1] = 0.4$

$\Pr[t_2] = 0.3$

$\Pr[t_3] = 0.3$

Global-top$k$ with $k = 2$

$\Pr[t_1] = 0.4$

$\Pr[t_2] = 0.5$

$\Pr[t_3] = 0.8$

$\Pr[t_4] = 0.3$

Property five: *Stability*

Making an item in the top-k list more likely (higher probability) or more important (higher score) should not remove it from the list.
Also in [Zhang, Chomicki, 08]

In the score-level uncertainty model, replace "higher probability or higher score" with "stochastically greater than"

Ranking method	Unique-rank	
U-top$k$	✓	
U-kRanks	×	[Zhang, Chomicki, 08]
PT-k	✓	
Global-top$k$	✓	

Properties: Summary

Ranking method	*Exact-k*	*Containment*	*Unique-Rank*	*Value-Invariant*	*Stability*
U-top$k$	weak	$\times$	✓	✓	✓
U-kRanks	✓	✓	$\times$	✓	$\times$
PT-k	$\times$	weak	✓	✓	✓
Global-top$k$	✓	$\times$	✓	✓	✓

Our definition: Expected Ranks

- Consider $E[t$'s rank$]$, and return the k tuples having the highest expected ranks

Our definition: Expected Ranks

- Consider $E[t\text{'s rank}]$, and return the k tuples having the highest expected ranks

tuples	score
t_1	$\{(100, 0.4), (70, 0.6)\}$
t_2	$\{(92, 0.6), (80, 0.4)\}$
t_3	$\{(85, 1)\}$

$$E[r(t2)] = 0.24 \times 2 + 0.16 \times 3 + 0.36 \times 1 + 0.24 \times 2 = 1.8$$

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	$0.4 \times 0.6 \times 1 = 0.24$
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	$0.4 \times 0.4 \times 1 = 0.16$
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	$0.6 \times 0.6 \times 1 = 0.36$
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	$0.6 \times 0.4 \times 1 = 0.24$

Our definition: Expected Ranks

- Consider $E[t\text{'s rank}]$, and return the k tuples having the highest expected ranks

tuples	score
t_1	$\{(100, 0.4), (70, 0.6)\}$
t_2	$\{(92, 0.6), (80, 0.4)\}$
t_3	$\{(85, 1)\}$

$$E[r(t1)] = 2.2$$

$$E[r(t2)] = 0.24 \times 2 + 0.16 \times 3 + 0.36 \times 1 + 0.24 \times 2 = 1.8$$

$$E[r(t3)] = 2$$

world W	$\Pr[W]$
$\{t_1 = 100, t_2 = 92, t_3 = 85\}$	$0.4 \times 0.6 \times 1 = 0.24$
$\{t_1 = 100, t_3 = 85, t_2 = 80\}$	$0.4 \times 0.4 \times 1 = 0.16$
$\{t_2 = 92, t_3 = 85, t_1 = 70\}$	$0.6 \times 0.6 \times 1 = 0.36$
$\{t_3 = 85, t_2 = 80, t_1 = 70\}$	$0.6 \times 0.4 \times 1 = 0.24$

Our definition: Expected Ranks

- If a tuple doesn't appear in a world, its rank is considered to be the last one

Our definition: Expected Ranks

- If a tuple doesn't appear in a world, its rank is considered to be the last one

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

$$E[r(t2)] = 0.2\times 2 + 0.2\times 4 + 0.3\times 1 + 0.3\times 3 = 2.4$$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1-p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1-p(t_1))p(t_3)p(t_4) = 0.3$

Our definition: Expected Ranks

- If a tuple doesn't appear in a world, its rank is considered to be the last one

tuples	score	$p(t)$
t_1	100	0.4
t_2	92	0.5
t_3	80	1
t_4	70	0.5

rules	
τ_1	$\{t_1\}$
τ_2	$\{t_2, t_4\}$
τ_3	$\{t_3\}$

world W	$\Pr[W]$
$\{t_1, t_2, t_3\}$	$p(t_1)p(t_2)p(t_3) = 0.2$
$\{t_1, t_3, t_4\}$	$p(t_1)p(t_3)p(t_4) = 0.2$
$\{t_2, t_3\}$	$(1 - p(t_1))p(t_2)p(t_3) = 0.3$
$\{t_3, t_4\}$	$(1 - p(t_1))p(t_3)p(t_4) = 0.3$

$$E[r(t1)] = 2.2$$

$$E[r(t2)] = 0.2 \times 2 + 0.2 \times 4 + 0.3 \times 1 + 0.3 \times 3 = 2.4$$

$$E[r(t3)] = 1.9$$

$$E[r(t3)] = 2.9$$

Properties: Summary

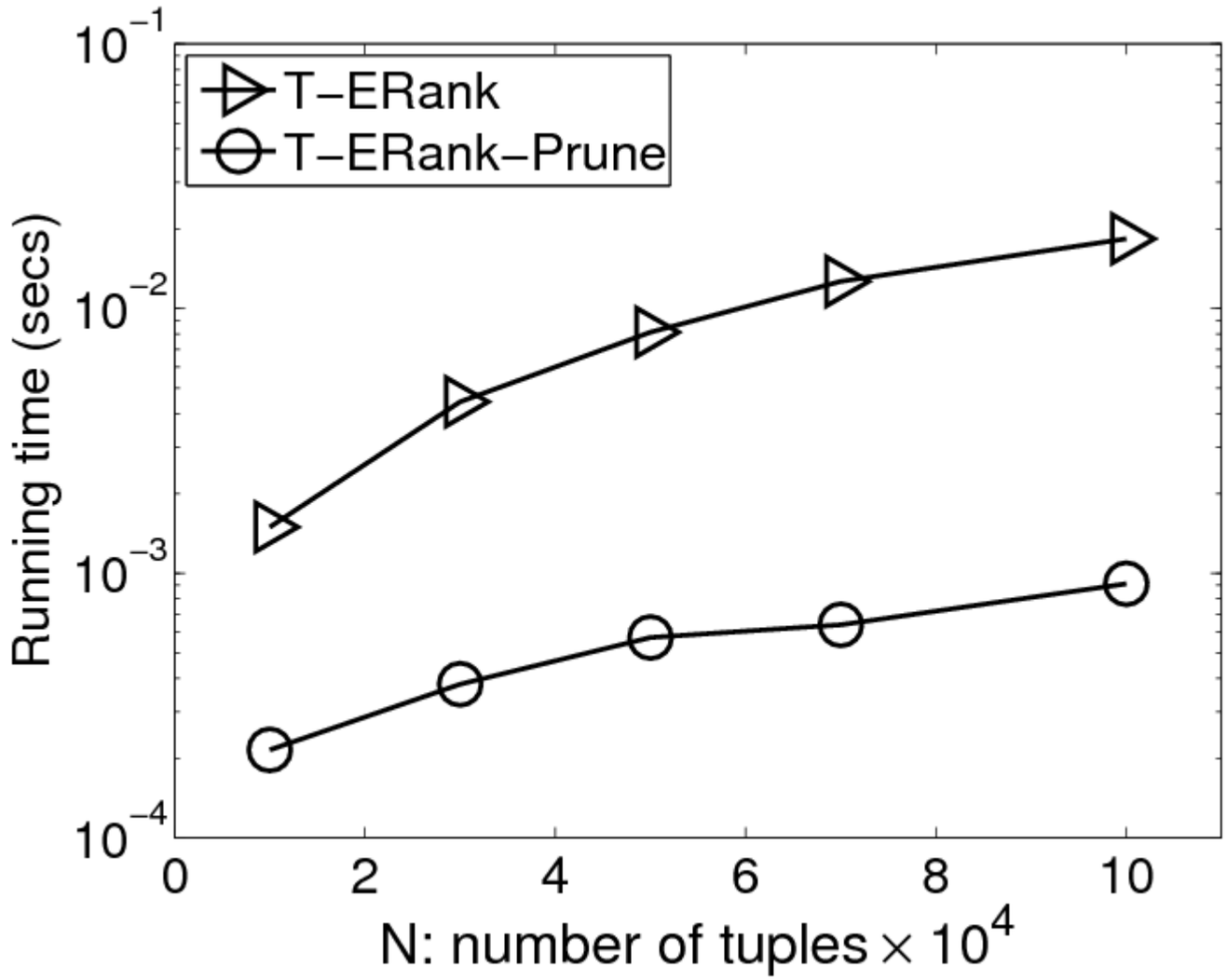
Ranking method	*Exact-k*	*Containment*	*Unique-Rank*	*Value-Invariant*	*Stability*
U-top$k$	weak	$\times$	✓	✓	✓
U-kRanks	✓	✓	$\times$	✓	$\times$
PT-k	$\times$	weak	✓	✓	✓
Global-top$k$	✓	$\times$	✓	✓	✓
Expected rank	✓	✓	✓	✓	✓

Computation

Ranking method	*Exact-k*	*Containment*	*Unique-Rank*	*Value-Invariant*	*Stability*	*Computation*
U-top$k$	weak	$\times$	✓	✓	✓	$n \log n$
U-kRanks	✓	✓	$\times$	✓	$\times$	kn^2
PT-k	$\times$	weak	✓	✓	✓	kn^2
Global-top$k$	✓	$\times$	✓	✓	✓	kn^2
Expected rank	✓	✓	✓	✓	✓	$n \log n$

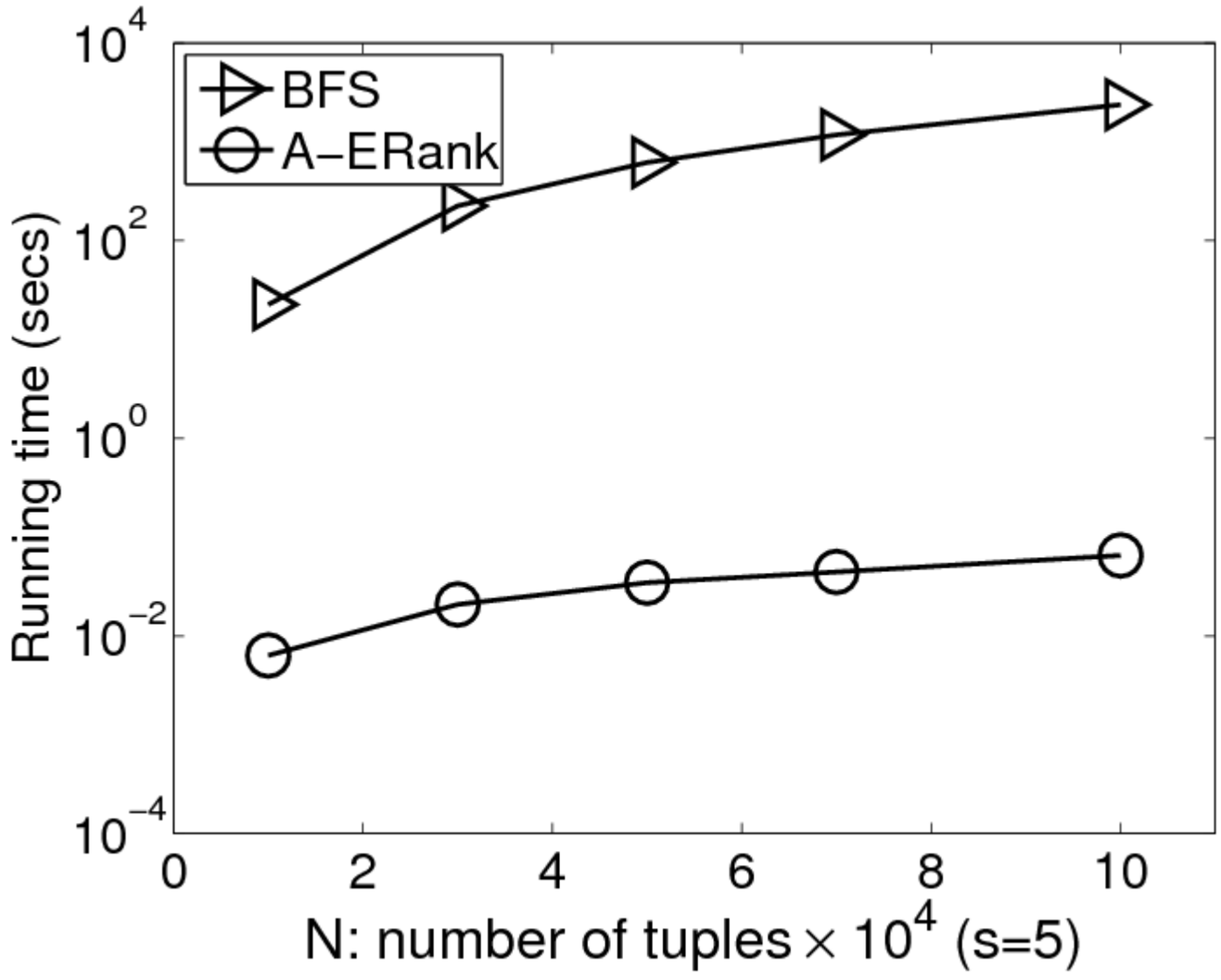
There are also pruning techniques that prune tuples that are not in the top-k

Experimental results: Tuple-level uncertainty



$$k = 100$$

Experimental results: Score-level uncertainty



$$k = 100$$

Other issues

- Why these five properties?

Other issues

- Why these five properties?
  - "Faithfulness" [Zhang, Chomicki, 08]

Other issues

- Why these five properties?
  - "Faithfulness" [Zhang, Chomicki, 08]
- Why expectation of the rank?

Other issues

- Why these five properties?
  - "Faithfulness" [Zhang, Chomicki, 08]
- Why expectation of the rank?
  - How about "median rank"?

The End

THANK YOU

Q and A