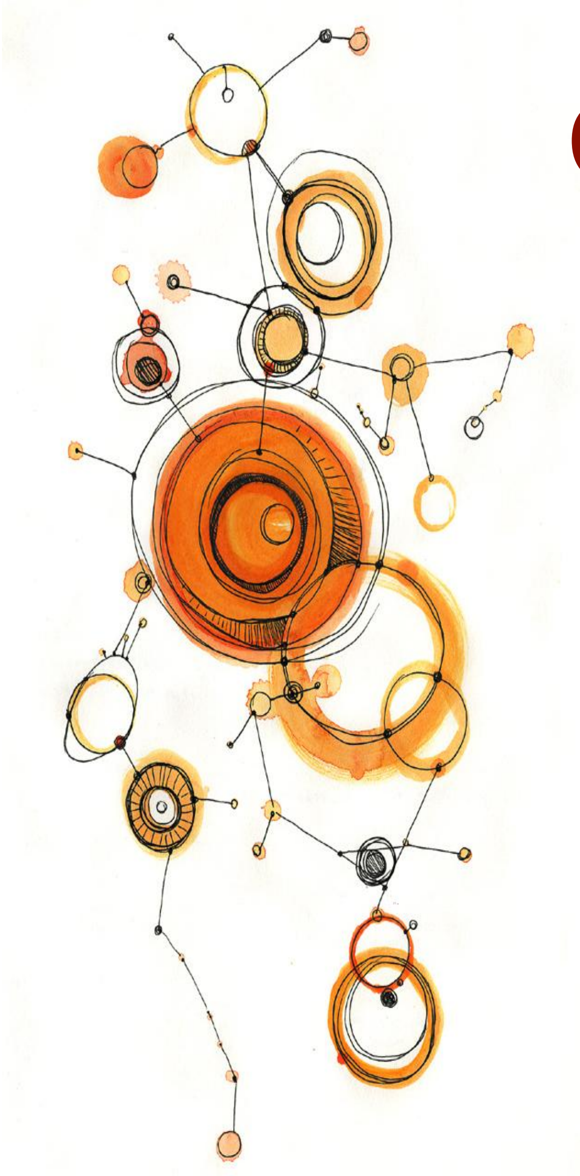


Getting the Best from Uncertain Data: the Correlated Case

Ilaria Bartolini, Paolo Ciaccia and Marco Patella
DEIS
University of Bologna
Italy

SEBD 2012, Venice, Italy, June 24th – 27th, 2012

Talk outline

- Preliminaries on
  - skyline query and
  - probabilistic (uncertain) databases
- Skyline on probabilistic relations including *correlation among tuples*
  - Notion of probabilistic domination
  - How to efficiently check probabilistic domination
- Implementation for different ranking semantics
- Time complexity analysis of algorithms for the resolution of skyline queries
- Conclusion and future work

Skyline of a deterministic relation

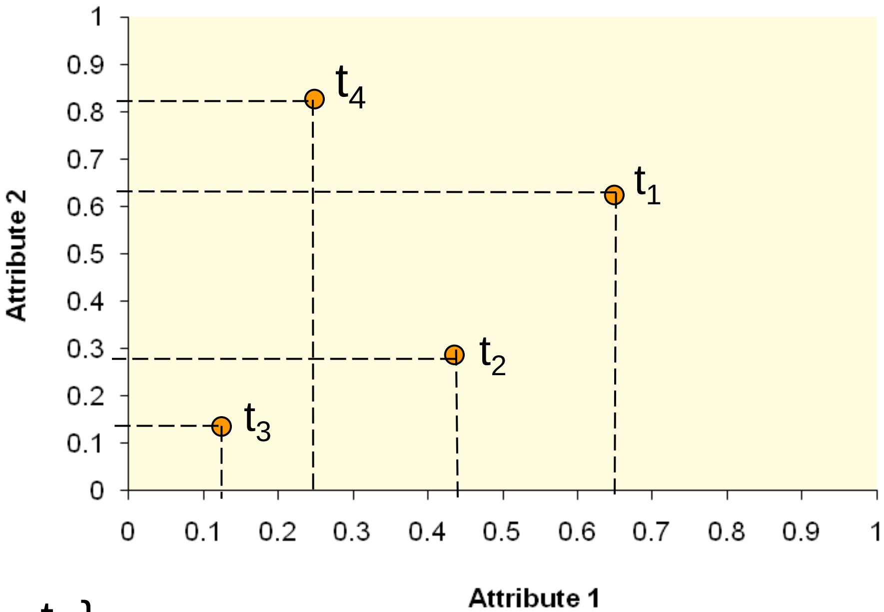
- The Skyline of a deterministic relation R returns all the objects that are not dominated by any other object

  "An object dominates another object if it is equal or better in all its dimensions and better in at least one dimension"

$$SKY(R) = \{u \in R \mid \nexists v \in R : v \succ u\}$$

Traffic-monitoring application

TID	Plate No	Time	Speed
t_1	H-111	11:20	145
t_2	W-266	11:05	137
t_3	X-255	10:50	135
t_4	C-444	10:55	155



Skyline = { t_1, t_4 }

Probabilistic databases

- Uncertain data can be represented through *probabilistic relations*, in which each tuple has also a *probability* (confidence) to appear in an instance of the relation

Traffic-monitoring application (sample of last-hour recording)

TID	Plate No	Radar	Time	Speed	Prob
t_1	X-123	L1	10:53	90	0.2
t_2	X-246	L2	10:50	100	0.15
t_3	X-246	L3	10:40	95	0.1
t_4	X-456	L1	10:32	110	0.1
t_5	X-456	L2	10:30	130	0.3
t_6	X-121	L3	10:30	110	0.2
t_7	X-324	L4	10:30	90	0.5
t_8	X-827	L4	10:20	105	0.35
t_9	X-827	L5	10:15	90	0.4
t_{10}	X-442	L5	10:10	120	0.3
t_{11}	X-442	L2	10:05	140	0.1

- We apply the notion of skyline to the case of probabilistic relations including *correlation among tuples*
  - *x-relation* *model* (i.e., mutual exclusion rules)

Rule: "because radar location, a *same car cannot be detected by two radars within an interval of one hour*"
→ tuples t_2 and t_3 can not be part of a same instance of the relation, for example

Contributions and basic properties

- In (*Bartolini et al., SEBD'11*) we have shown how *skyline queries can be defined for* a probabilistic relation R_p
  - the case of *independent tuples* was analyzed
- In this paper we extend the applicability of skyline queries to the "*correlated* " case (*x-relations*)
  - Our probabilistic domination definition is general
    - does not depend on the *ranking semantics* used
  - Its implementation depends on the specific *ranking semantics*
  - We detail the analysis for 5 commonly used *ranking semantics*

- A *ranking semantics*, given a linear order of (deterministic) tuples and a probabilistic model, produces a new (probabilistic) ranking of tuples
  - Different ranking semantics give with the same input different probabilistic rankings!

Skyline of *Rp* depends on ranking semantics

- It is well known that different ranking semantics yield quite different results for top-*k* queries in probabilistic databases
- This is also the case when such semantics are used for skyline queries
- Continuing our running example:

"A skyline query on the **Time** and **Speed** attributes finds those readings that are the most recent ones and concern high-speed cars"

- In the deterministic case, it is: SKY(R) = { t_1, t_2, t_4, t_5, t_{11}}

Traffic-monitoring application

Ranking semantics	SKY(Rp)
Expected Rank	{t_5, t_7}
Expected Score	{t_1, t_2, t_4, t_5, t_7, t_8, t_{11}}
U-Top1, U-1Ranks, Global-Top1	{t_1, t_5}

Although *t_5 is part of all considered skylines, this is not* the case for other tuples (e.g., *t_7*)

Basic ingredients

- Domination relation $\succ$ between tuples u and v is a *strict partial order*
- A *linear order* $\rhd$ is a strict partial order that is also *connected* (either $u \rhd v$ or $v \rhd u$)
- $\rhd$ is called *linear extension* of $\succ$ *iff* $u \succ v \Rightarrow u \rhd v$
- Any strict partial order $\succ$ equals the intersection of its linear extensions

$$\succ = \bigcap \left\{ \rhd \,\middle|\, \rhd \in Ext(\succ) \right\}$$

- A *probabilistic ranking function* Ψ is a function that, given R_p and a *linear order* on the (deterministic) tuples of R, yields a *probabilistic linear order* $\rhd_p = \psi(\rhd, R_p)$ on the probabilistic tuples of R_p

- The actual *ranking of the probabilistic tuples* is obtained by computing for each tuple *u a value* $\psi_{\rhd}(u)$ *so that*

$$u \rhd_p v \Leftrightarrow \psi_{\rhd}(u) > \psi_{\rhd}(v)$$

Let's put it all together

Probabilistic domination (P-domination for short):

- Given two tuples u and v in R_p, we say that u P-dominates v (i.e., $u \succ_p v$)

$$u \succ_p v \Leftrightarrow u \rhd_p v, \forall \rhd_p = \psi(\rhd, R_p), \rhd \in Ext(\succ)$$

$$\downarrow$$

$$u \succ_p v \Leftrightarrow \psi_{\rhd}(u) > \psi_{\rhd}(v), \forall \rhd \in Ext(\succ)$$

SKY(Rp):

- Consequently, similarly to the deterministic case, the skyline of R_p *is defined as:*

$$SKY(R_p) = \{u \in R \mid \nexists v \in R : v \succ_p u\}$$

where the only difference with deterministic case is that $\succ$ is replaced with $\succ_p$

Our approach

- Goal: check P-domination without materializing any linear extension of $\succ$

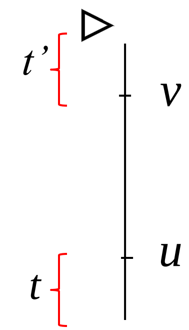
- Note that for $u \succ_p v$ to hold it has to be $\psi_{\triangleright}(u) > \psi_{\triangleright}(v)$ *for all linear extensions* $\triangleright$ *of* $\succ$, that is:

$$u \succ_p v \Leftrightarrow \min_{\triangleright \in Ext(\succ)} \left\{ \frac{\psi_{\triangleright}(u)}{\psi_{\triangleright}(v)} \right\} > 1$$

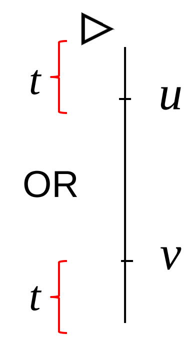
- The key idea is:
  - if we find a linear order $\triangleright$ that is the *most unfavorable one* for *u with respect to v, and* $\psi_{\triangleright}(u) > \psi_{\triangleright}(v)$ holds for this "extremal" order
  - then it will necessarily hold for any other linear extensions of $\succ$

How to check P-domination

- When comparing two tuples u and v, we analyze how other tuples should be arranged in the linear order so as to minimize the ratio $\psi_{\triangleright}(u)/\psi_{\triangleright}(v)$
- Two relevant cases (regardless of the specific probabilistic ranking function Ψ):
  - $u \not\succ v$: if *u does not dominate v,* the *extremal linear order* corresponds to the case where:
    1. $u \triangleright t$ only for those tuples t that u dominates, and
    2. $t' \triangleright v$ *only for those tuples t' that dominate v*



  - $u \succ v$: when *u dominates v,* $u \triangleright v$ has necessarily to hold
    - We just have to determine the order of tuples t which are "*indifferent*" to both u and v
      - Each of such tuples is necessarily sorted either above u (so as to unfavor both u and v) or under v (so as to favor both u and v)

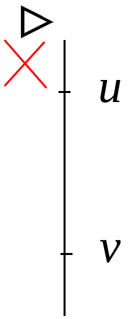


Ranking semantics

- Expected rank (*Cormode et al.*, ICDE'09)
  - $\psi_{\triangleright}(u)$ is the average rank of u
- Expected score (*Cormode et al.*, ICDE'09)
  - $\psi_{\triangleright}(u)$ is the average score of $u = p(u) \times s(u)$
- U-Top*k* (*Soliman et al.*, ICDE'07), U-*k*Ranks (*Soliman et al.*, ICDE'07), Global-Top*k* (*Zhang et al.*, DBRank'08)
  - We are only interested in top-1 results (Skyline is the union of all top-1 results for any linear order)
  - All ranking semantics give the same result for *k*=1
  - $\psi_{\triangleright}(u)$ is the probability of u to be the top-1 tuple

Expected score

- $\psi_{\triangleright}(u)$ is the average score of $u = p(u) \times s(u)$
  - *Does not depend on the tuples that precede* u
  - Does not depend on the correlation model

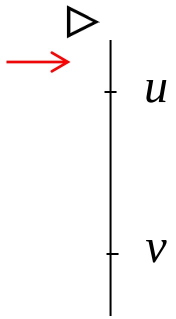


- It is:

$$u \succ_p v \Leftrightarrow u \succ v \wedge \frac{p(u)}{p(v)} \geq 1$$

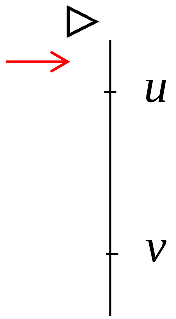
- Note that above result is valid for *any* correlation model
  - Not only for x-relations

U-Top1, U-1Ranks, and Global-Top1



- $\psi_{\triangleright}(u)$ is the probability of u to be the top-1 tuple
  - Thus, $\psi_{\triangleright}(u)$ *only depends on tuples that precede* u
- For x-relations $\psi_{\triangleright}(u)$ is computed as a "*product of sums*"
- It follows that a tuple t that is *indifferent* to both u and v:
  - If t is mutually exclusive to v, then t should be ordered above u
    - t does not influence the probability of v to be the top-1 tuple, because there is no instance containing both t and v
    - We do not favor u
  - Otherwise, t should be ordered under v
    - We do not unfavor v

Expected rank



- $\psi_{\triangleright}(u)$ is the average rank of u
  - Thus, $\psi_{\triangleright}(u)$ *only depends on tuples that precede* u
- For x-relations $\psi_{\triangleright}(u)$ is computed as a "*sum of products*"

- It follows that a tuple t that is *indifferent* to both u and v:
  - If t is mutually exclusive to v, then t should be ordered above u
    - t does not influence the average rank of v,
      because there is no instance containing both t and v
    - We do not favor u
  - If t is mutually exclusive to u, then t should be ordered under v
    - t does not influence the average rank of u,
      because there is no instance containing both t and u
    - We do not unfavor v
  - Otherwise, t could be ordered either above u or under v
    - It depends on the probabilities of u and v
    - Either we do not favor u or we do not unfavor v

Complexity analysis

- Time complexity *for computing the SKY(R_p)*, R_p consists of N tuples
  - In the worst case we should compare every pair of tuples

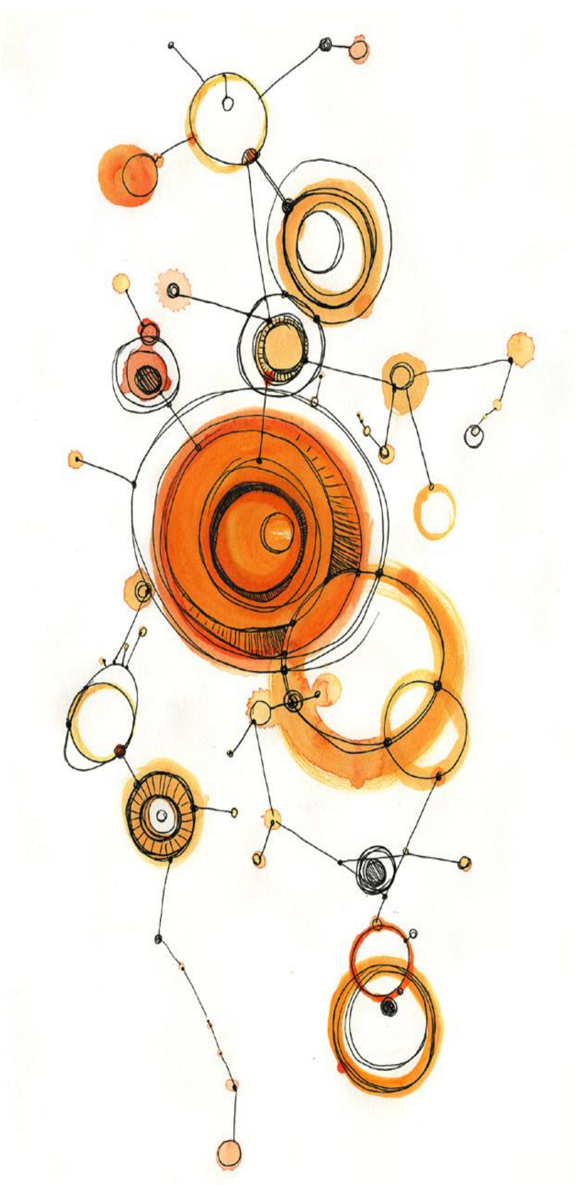
1. Expected score
   - A single condition with $O(1)$ complexity
   - Overall complexity $O(N^2)$
2. U-Top1, U-1Ranks, Global-Top1
   - $u \not\succ v$ with $O(1)$ complexity (with pre-computation)
   - $u \succ v$ with $O(N)$ complexity (possibly all tuples are indifferent to each other)
   - Overall complexity $O(N^3)$
3. Expected rank
   - $u \not\succ v$ with $O(1)$ complexity (with pre-computation)
   - $u \succ v$ with $O(N)$ complexity (possibly all tuples are indifferent to each other)
   - Overall complexity $O(N^3)$

- For 2. and 3., for the case $u \succ v$, we propose an $O(1)$ *sufficient condition* so as to postpone the $O(N)$ check as much as possible

Conclusions and future work

- In this extended abstract we apply the notion of skyline to the case of probabilistic relations
- We elaborated our analysis for the “correlated” case of x-relations, consisting of a set of generation rules specifying the mutual exclusion of tuples
- Extended version of the paper has been accepted for publication on the IEEE TKDE journal

- We argue that there is not a “best” ranking semantics, rather the choice might depend on the specific application at hand and user preferences
  - Understanding the properties of the different semantics is therefore a, both practical and theoretical, interesting research issue
- Moreover, we are also interested in considering other, more expressive, correlation models (e.g., and/xor trees and probabilistic graphical models)



Thank you!

☺