Getting the Best from Uncertain Data: the Correlated Case*

Ilaria Bartolini, Paolo Ciaccia, and Marco Patella

DEIS - Università di Bologna, Italy
{i.bartolini,paolo.ciaccia,marco.patella}@unibo.it

Abstract. In this extended abstract we apply the notion of skyline to the case of probabilistic relations including correlation among tuples. In particular, we consider the relevant case of the x-relation model, consisting of a set of generation rules specifying the mutual exclusion of tuples. We show how our definitions apply to different ranking semantics and analyze the time complexity for the resolution of skyline queries.

1 Introduction

The management of uncertain data has become a very active area of research, due to the huge number of relevant applications it has. Among them, data extraction from the Web, data integration, sensor networks, etc. Consequently, several works have focused on the problem of extending different query types to such scenarios [6, 4]. In [2] we have shown how *skyline queries* can be defined for probabilistic relations, where each tuple has also a probability (confidence) to appear [6, 7]. Although our definition applies to any model of tuple correlation, only the case of independent tuples was analyzed in [2]. In this paper we extend the applicability of skyline queries to the relevant case of *x-relations* [1], where mutual exclusion of tuples is possible. We detail the analysis for the most common ranking semantics (since our skyline is parametric with respect to them) and analyze the time complexity for the resolution of skyline queries.

We adopt the well-known *possible worlds semantics* [4], in which a probabilistic relation R^p represents a set of standard relations, each termed a *possible world*. Compactly, R^p is represented as a triple, $R^p = (R, p, \mathcal{G})$, where R is a relation in the standard sense, i.e., a set of tuples, also called a *deterministic* relation; p is a function that assigns to each tuple $u \in R$ a probability, $p(u) \in (0, 1]$; and $\mathcal{G}$ is a set of *generation rules* that capture correlations among tuples. In the case of the *x-relation* model [1] $\mathcal{G}$ only includes mutual exclusion rules that form a partition of R. In this scenario each $G \in \mathcal{G}$ is also called a *group*, and consists of a set of mutually exclusive tuples (with overall probability ≤ 1). When needed, we will use G_u to denote the group of tuple u. A *possible world* W of R^p is a subset of tuples of R that respect all the rules in $\mathcal{G}$. The set of all possible worlds of R^p is denoted $\mathcal{W}$.

* Extended Abstract

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

Given a (deterministic) relation R whose schema includes a set of numerical attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_d\}$, the *skyline* of R with respect to $\mathcal{A}$, denoted $\textsc{Sky}(R)$, is the set of *undominated* tuples in R. Assuming that on each attribute higher values are preferable, tuple u dominates tuple v, written $u \succ v$, iff it is $u.A_i \geq v.A_i$ for each $A_i \in \mathcal{A}$ and there exists at least one attribute A_j such that $u.A_j > v.A_j$. Thus: $\textsc{Sky}(R) = \{u \in R \mid \nexists\ v \in R : v \succ u\}$. When neither $u \succ v$ nor $v \succ u$ hold, we say that u and v are *indifferent*, written $u \sim v$. A *scoring function* $s()$ on the attributes $\mathcal{A}$ is *monotone* iff $u.A_i \geq v.A_i$ $(i = 1, \ldots, d)$ implies $s(u) \geq s(v)$, and is also *domination-preserving* if $u \succ v$ implies $s(u) > s(v)$.

2 The Skyline of a Probabilistic Relation

In [2] we showed how domination among probabilistic tuples (*P-domination* for short) can be defined by using ingredients drawn from order theory. In particular, $\succ$ is a *strict partial order*, i.e., an irreflexive $(\forall u : u \not\succ u)$ and transitive $(\forall u, v, t : u \succ v \wedge v \succ t \Rightarrow u \succ t)$ binary relation. A *linear* order $\gtrdot$ is a strict partial order that is also *connected*, i.e., for any two distinct tuples u and v, either $u \gtrdot v$ or $v \gtrdot u$. A linear order $\gtrdot$ is called a *linear extension* of $\succ$ iff $u \succ v \Rightarrow u \gtrdot v$, i.e., $\gtrdot$ is *compatible* with $\succ$. It is known that any strict partial order $\succ$ equals the intersection of its linear extensions, i.e., $\succ = \bigcap\{\gtrdot | \gtrdot \in \textsc{Ext}(\succ)\}$, where $\textsc{Ext}(\succ)$ denote the set of all linear extensions of $\succ$.

We then bring in probability by exploiting the concept of *ranking semantics*. Such semantics, which are used for supporting top-k queries on probabilistic data, define a linear order on the probabilistic tuples of R^p when the (deterministic) tuples of R are ordered according to $\gtrdot$ [7, 9, 3]. More precisely, a *probabilistic ranking function* Ψ is a function that, given R^p and a linear order $\gtrdot$ on the tuples of R, yields a *probabilistic linear order* $\gtrdot_p = \Psi(\gtrdot, R^p)$ on the probabilistic tuples of R^p. The actual ranking of the probabilistic tuples is obtained by computing for each tuple u a value $\psi_{\gtrdot}(u)$, so that $u \gtrdot_p v$ iff $\psi_{\gtrdot}(u) > \psi_{\gtrdot}(v)$.

The definition of P-domination is then as follows [2]:

Definition 1 (P-domination). *Let $R^p = (R, p, \mathcal{G})$ be a probabilistic relation, and let $\succ$ be the domination relation on the tuples in R when considering the skyline attributes $\mathcal{A}$. Let Ψ be a probabilistic ranking function on R^p. For any two tuples u and v in R^p, we say that u* P-dominates *v, written $u \succ_p v$, iff for each linear extension $\gtrdot$ of $\succ$, with associated probabilistic linear order $\gtrdot_p = \Psi(\gtrdot, R^p)$, it is $u \gtrdot_p v$, that is:*

$$\boxed{u \succ_p v \Longleftrightarrow u \gtrdot_p v, \ \ \forall \gtrdot_p = \Psi(\gtrdot, R^p), \gtrdot \in \textsc{Ext}(\succ)} \quad (1)$$

Consequently, similarly to the deterministic case, the skyline of R^p is defined as:

$$\boxed{\textsc{Sky}(R^p) = \{u \in R \mid \nexists\ v \in R : v \succ_p u\ \}} \quad (2)$$

3 P-domination with x-relations

In [2] we have put forward the basic principles that allow P-domination to be checked without materializing any linear extension of $\succ$. In particular, for $u \succ_p v$

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

to hold it has to be $\psi_{\gtrdot}(u) > \psi_{\gtrdot}(v)$ for all linear extensions $\gtrdot$ of $\succ$, that is: $u \succ_p v \Leftrightarrow \min_{\gtrdot \in \textsc{Ext}(\succ)} \{\psi_{\gtrdot}(u)/\psi_{\gtrdot}(v)\} > 1$. Consequently, if one finds a linear order that is the most unfavorable one for u with respect to v, and $\psi_{\gtrdot}(u) > \psi_{\gtrdot}(v)$ holds for this "extremal" order, then it will necessarily hold for all other orders compatible with $\succ$. This approach amounts to determine a set of *P-domination rules* that can be checked without the need of materializing any linear extension of $\succ$. Regardless of the specific probabilistic ranking function Ψ, the two relevant cases to consider for P-domination rules are:

$\mathbf{u} \succ \mathbf{v}$: When u dominates v, $u \gtrdot v$ has necessarily to hold; we can thus analyze how other tuples should be arranged in the linear order so as to minimize the ratio $\psi_{\gtrdot}(u)/\psi_{\gtrdot}(v)$.

$\mathbf{u} \not\succ \mathbf{v}$: If u does not dominate v, then extremal linear order corresponds to the case where: 1) $u \gtrdot t$ only for those tuples t that u dominates, and 2) $t' \gtrdot v$ only for those tuples t' that dominate v.

According to these principles, in [2] we have derived specific P-domination rules for the independent case. In the following we concentrate on the x-relation model and derive P-domination rules for a variety of relevant ranking semantics.

3.1 Expected Rank [3]

Given a linear order $\gtrdot$ on the tuples of R, the rank of u in a possible world W is the number of tuples in W that precede u [3]. The expected rank of a tuple u is thus given as the expected value of the rank of u on the set of all possible worlds, each weighted by its probability. This can be expressed as:

$$ER_{\gtrdot}(u) = p(u) \times \sum_{t \gtrdot u, t \notin G_u} p(t) \; + (1 - p(u)) \times \sum_{t \notin G_u} p(t) \; + \sum_{t \in G_u, t \neq u} p(t)$$

where the first term is the expected rank of u in a possible world in which u appears, whereas the other terms account for the contribution of those worlds that do not contain u (where u has rank $|W|$).

Let $H_{G,\gtrdot}(u) = \sum_{t \gtrdot u, t \notin G_u} p(t)$ be the overall probability of those tuples not in G_u that are better than u according to $\gtrdot$. Further, let $P = \sum_{t \in R^p} p(t)$ be the overall probability of all tuples, $P_{u,v} = P - p(u) - p(v)$, and define $P_{G_u} = \sum_{t \in G_u, t \neq u} p(t)$, as the probability of all tuples but u in G_u.

According to Definition 1, and setting $\psi_{\gtrdot}(u) = -ER_{\gtrdot}(u)$, tuple u P-dominates v iff it is $\max_{\gtrdot \in \textsc{Ext}(\succ)}\{ER_{\gtrdot}(u)/ER_{\gtrdot}(v)\} < 1$, i.e., $ER_{\gtrdot}(u) < ER_{\gtrdot}(v)$ for any $\gtrdot \in \textsc{Ext}(\succ)$. This is equivalent to:

$$u \succ_p v \Leftrightarrow \boxed{\frac{p(u)}{p(v)} > \max_{\gtrdot \in \textsc{Ext}(\succ)} \left\{ \frac{P_{u,v} + 1 - H_{G,\gtrdot}(v) - P_{G_v}}{P_{u,v} + 1 - H_{G,\gtrdot}(u) - P_{G_u}} \right\}} \qquad (3)$$

The two cases to consider for checking the validity of Inequality 3 are dealt with as follows. For both here we describe the general case in which u and v belong to different groups first, leaving out for lack of space the (easier) case $G_u = G_v$.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

$\mathbf{u} \succ \mathbf{v}$: The first rule we derive is sufficient (but not necessary) for P-domination to occur, and has the advantage that it can be more efficiently checked than the other we derive for the case $u \succ v$.

Since $u \succ v$, it is $H_{G,\gtrdot}(v) \geq H_{G,\gtrdot}(u) + p(u) - P_{G_v}$, which is obtained by assuming a worst-case scenario for u in which all tuples $t \in G_v$ other than v dominate u. Substituting in Inequality 3 it is obtained:[1]

$$\boxed{u \succ v \ \wedge \ \frac{p(u)}{p(v)} \geq 1 \ \wedge \ \frac{p(u)}{P_{G_u}} \geq 1} \qquad \text{(ER:Rule 1)}$$

In case Rule 1 is not satisfied, a preliminary key observation is that, for any linear order $\gtrdot$ that extends $\succ$, it is $H_{G,\gtrdot}(u) \in [H^-_G(u), H^+_G(u)]$, where the two bounds are respectively defined as:

$$H^-_G(u) = \sum_{t \succ u, t \notin G_u} p(t) \qquad H^+_G(u) = \sum_{u \not\succ t, t \notin G_u} p(t)$$

Clearly, $H^-_G(u)$ is the best possible case for u, since only those tuples from other groups that dominate u are also better than u according to $\gtrdot$, whereas the worst case for u arises when u is better only of those tuples of other groups that it dominates.

Now, the ratio in the right-hand side of Inequality 3 can be maximized by adequately arranging those tuples t that are indifferent to either u or v (or both). Since $u \succ v$, and $\succ$ is transitive, only three cases can occur:

$\mathbf{t} \sim \mathbf{u}, \mathbf{t} \succ \mathbf{v}$: In order to favor v with respect to u it has necessarily to be $t \gtrdot u$ whenever $t \notin G_u$. Let $IB(u,v) = \sum_{t \sim u, t \succ v, t \notin G_u} p(t)$ stand for the mass of probability of such tuples.

$\mathbf{t} \sim \mathbf{v}, \mathbf{u} \succ \mathbf{t}$: For the same reason it has to be $v \gtrdot t$.

$\mathbf{t} \sim \mathbf{u}, \mathbf{t} \sim \mathbf{v}$: To analyze this case, first observe that setting $t \gtrdot u \gtrdot v$ when $t \in G_v$ would penalize only u. Let $II_{G_v}(u,v) = \sum_{t \sim u, t \sim v, t \in G_v} p(t)$ be the total mass of probability of such tuples. At this point the ratio in Inequality 3 would be written as:

$$\frac{N(u,v)}{D(u,v)} \stackrel{\text{def}}{=} \frac{P_{u,v} + 1 - H^-_G(v) - P_{G_v}}{P_{u,v} + 1 - H^-_G(u) - IB(u,v) - II_{G_v}(u,v) - P_{G_u}} \tag{4}$$

Consider now those tuples $t \notin G_u \cup G_v$, whose total mass of probability is $\Delta(u,v) = \sum_{t \sim u, t \sim v, t \notin G_u, t \notin G_v} p(t)$. The two alternatives to consider are either $t \gtrdot u \gtrdot v$ or $u \gtrdot v \gtrdot t$. The choice actually depends on the value of $N(u,v)/D(u,v)$: If $N(u,v) < D(u,v)$, then setting $t \gtrdot u \gtrdot v$ would lead to the ratio $(N(u,v) - \Delta(u,v))/(D(u,v) - \Delta(u,v))$, which is *lower* than $N(u,v)/D(u,v)$. On the other hand, when $N(u,v) > D(u,v)$, then $(N(u,v) - \Delta(u,v))/(D(u,v) - \Delta(u,v)) > N(u,v)/D(u,v)$, thus the case $t \gtrdot u \gtrdot v$ is the one to consider. Finally, when $N(u,v) = D(u,v)$ Inequality 3 degenerates to $p(u) > p(v)$.

[1] Note that in case neither inequality holds strictly in Rule 1 it is still safe to say that $u \succ_p v$, since we assume a domination-preserving tie-breaking rule.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

Putting all together, the 2nd P-domination rule is:

$$\boxed{u \succ v \;\wedge\; \frac{p(u)}{p(v)} \geq \begin{cases} \dfrac{N(u,v)}{D(u,v)} & \text{if } \dfrac{N(u,v)}{D(u,v)} \leq 1 \\ \dfrac{N(u,v) - \Delta(u,v)}{D(u,v) - \Delta(u,v)} & \text{otherwise} \end{cases}} \qquad \text{(ER:Rule 2)}$$

u $\not\succ$ v: When u does not dominate v, P-domination can still occur. In this case it is immediate to see that the ratio in Inequality 3 is maximized by setting $H_{G,\geqslant}(v) = H_G^-(v)$ and $H_{G,\geqslant}(u) = H_G^+(u)$, thus:

$$\boxed{u \not\succ v \;\wedge\; \frac{p(u)}{p(v)} > \frac{P_{u,v} + 1 - H_G^-(v) - P_{G_v}}{P_{u,v} + 1 - H_G^+(u) - P_{G_u}}} \qquad \text{(ER:Rule 3)}$$

3.2 Expected Score [3]

In [3], the case where tuples are ranked by their *expected score*, $ES(u) = s(u) \times p(u)$, is also considered. Here, the scoring function $s()$ is assumed to be domination-preserving, i.e., $u \succ v$ implies $s(u) > s(v)$. For this ranking semantics (the only one that explicitly depends on tuples' scores) it is immediate to derive that:

$$\boxed{u \succ v \;\wedge\; \frac{p(u)}{p(v)} \geq 1} \qquad \text{(ES:Rule 1)}$$

This holds since, for any scoring function $s()$, it has to be $s(u) \times p(u) > s(v) \times p(v)$, i.e., $p(u)/p(v) > s(v)/s(u)$. If $u \not\succ v$, then for any given value of the ratio $p(u)/p(v)$ the right-hand side can be made arbitrarily large and P-domination does not hold. If $u \succ v$, it is $s(v)/s(u) = 1 - \epsilon$, with $\epsilon > 0$ arbitrarily small, from which the result follows.

It is remarkable that, although ranking by expected scores is highly dependent on actual score values, the skyline of R^p is, even in this case, insensitive to attribute values. A major pitfall of expected score ranking is that it completely ignores correlation among tuples, and this obviously propagates to skylines. Further, since the above is the *only* P-domination rule, the skyline obtained from the expected score semantics will be always larger than the skyline in the deterministic case (since $u \succ v$ is necessary for P-domination to hold).

3.3 U-Topk [7], U-kRanks [7], and Global-Topk [9]

The semantics we consider in this section do not define a ranking in the proper sense of term, since ranking varies with the value of k. Therefore, our results do apparently not apply to them. However, P-domination can still be defined by considering the case of top-1 queries. This holds since the skyline is the union of top-1 answers, thus we can limit the analysis to the case $k = 1$, and report as $\textsc{Sky}(R^p)$ those tuples that, for at least one linear order $\geqslant$ compatible with $\succ$, are the top-ranked ones.

For these semantics, which for lack of space we cannot describe here, we can provide a unified analysis, due to the following observation (see also [8]):

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

Observation 1 *For any probabilistic relation R^p, whose tuples are linearly ordered according to $\gtrdot$, the result of a top-1 query computed according to the U-Top1, U-1Ranks, and Global-Top1 semantics is the same, and is given by the tuple u that maximizes the* top-1 probability*, which for x-relations is:*

$$\Pr_{\gtrdot}^{top1}(u) = p(u) \times \prod_{G \neq G_u} \left(1 - \sum_{t \in G, t \gtrdot u} p(t)\right) \tag{5}$$

Thus, for such semantics one can set $\psi_{\gtrdot}(u) = \Pr_{\gtrdot}^{top1}(u)$. Now, by defining $Q_{G,\gtrdot}(u) = \prod_{G \neq G_u} \left(1 - \sum_{t \gtrdot u, t \in G} p(t)\right)$, it is obtained:

$$u \succ_p v \Leftrightarrow \boxed{\frac{p(u)}{p(v)} > \max_{\gtrdot \in \text{EXT}(\succ)} \left\{ \frac{Q_{G,\gtrdot}(v)}{Q_{G,\gtrdot}(u)} \right\}} \tag{6}$$

u ≻ v: To derive the first rule we can assume a worst case scenario for u where all tuples of G_v other than v dominate u, obtaining $Q_{G,\gtrdot}(v) \leq Q_{G,\gtrdot}(u) \frac{1-p(u)}{1-P_{G_v}}$. This, substituted into Equation 6, gives us the following rule:

$$\boxed{u \succ v \ \wedge \ \frac{p(u)}{p(v)} \geq \frac{1 - p(u)}{1 - P_{G_v}}} \tag{Top1:Rule 1}$$

If Rule 1 does not hold, we can still bound $Q_{G,\gtrdot}(u)$ by observing that the best possible order for u is the one where u is preceded by only those tuples that dominate u, while the worst order is the one that puts u after all tuples not dominated by u. Thus, it is $Q_{G,\gtrdot}(u) \in [Q_G^-(u), Q_G^+(u)]$, where the bounds are defined as follows:

$$Q_G^-(u) = \prod_{G \neq G_u} \left(1 - \sum_{u \not\succ t, t \in G} p(t)\right) \qquad Q_G^+(u) = \prod_{G \neq G_u} \left(1 - \sum_{t \succ u, t \in G} p(t)\right)$$

As in Section 3.1, in order to maximize the ratio in the right side of Equation 6 one should only arrange those tuples t that are indifferent to u or v (or both). If u is to be penalized, we should maximize the number of tuples preceding it. To this end, the tuples $t \notin G_u, t \sim u, t \succ v$ and $t \in G_v, t \sim u, t \sim v$ can be all ordered so as to precede u, since they either precede also v or do not contribute to its top-1 probability. Consider now those tuples $t \notin G_v : t \sim u, t \sim v$. The two alternatives to consider are either $t \gtrdot u \gtrdot v$ or $u \gtrdot v \gtrdot t$. To this end we first note that, since $u \succ v$, it is $\sum_{t \gtrdot u} p(t) \leq \sum_{t \gtrdot v} p(t)$, thus having $t \gtrdot u \gtrdot v$ would actually lower the ratio $(1 - \sum_{\substack{t \gtrdot u \\ t \notin G_u \cup G_v}} p(t))/(1 - \sum_{\substack{t \gtrdot v \\ t \notin G_u \cup G_v}} p(t))$, leading to an overall lower value of the right side of Equation 6. By defining $B\_IB(u,v)$ as:

$$B\_IB(u,v) \stackrel{\text{def}}{=} \left(1 - \sum_{\substack{(t \succ u \vee (t \sim u \wedge t \succ v) \vee \\ (t \sim u \wedge t \sim v)), t \in G_v}} p(t)\right) \times \prod_{\substack{G \neq G_u \\ G \neq G_v}} \left(1 - \sum_{\substack{(t \succ u \vee (t \sim u \wedge t \succ v)) \\ t \in G}} p(t)\right)$$

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, Copyright (c) 2012 - Edizioni Libreria Progetto and the authors

the second P-domination rule is thus obtained as:

$$\boxed{u \succ v \;\wedge\; \frac{p(u)}{p(v)} \geq \frac{Q_G^+(v)}{B\_IB(u,v)}} \qquad \text{(Top1:Rule 2)}$$

u $\not\succ$ v: It is easy to see that, since now it could happen that $v \succ u$, it is:

$$\boxed{u \not\succ v \;\wedge\; \frac{p(u)}{p(v)} > \frac{Q_G^+(v)}{Q_G^-(u)}} \qquad \text{(Top1:Rule 3)}$$

4 Complexity Analysis

In this section we analyze, for the alterative ranking semantics covered in Section 3, the time complexity for computing the skyline of a probabilistic relation R^p consisting of N tuples. We first cover the simpler case of the expected scores semantics. In this scenario, the only P-domination rule (ES:Rule 1) is a constant time, $\mathcal{O}(1)$, operation. Consequently, the skyline of R^p can be computed in $\mathcal{O}(N^2)$ time.

Turning to consider the expected rank semantics, it is important to first characterize the time complexity of the three P-domination rules described in Section 3.1:

Rule 1: In the worst case, the evaluation over all pairs of tuples requires $\mathcal{O}(N^2)$ time. This clearly holds when the probability of each group is known in advance, but it remains true even if group probabilities need to be determined (since this can be done before starting to compare tuples).

Rule 3: A single check on a pair of tuples apparently has a complexity of $\mathcal{O}(N)$, because of $H_G^-(v)$ and $H_G^+(u)$; however, such values do not depend on the specific pair of tuples being compared, and thus can be pre-computed for each tuple, which collectively requires $\mathcal{O}(N^2)$ time. Thus, a single evaluation is a constant time, $\mathcal{O}(1)$, operation and evaluation on all pairs of tuples requires $\mathcal{O}(N^2)$ time.

Rule 2: A single evaluation requires $\mathcal{O}(N)$ time, because of $IB(u,v)$, $II_{G_v}(u,v)$, and, possibly, $\Delta(u,v)$. Unless one can afford a $\mathcal{O}(N^2)$ storage overhead for book-keeping (which we do not consider here), such quantities cannot be computed in advance because they depend on both u and v.

Therefore, because of Rule 2, computing the skyline of R^p under the expected rank semantics requires $\mathcal{O}(N^3)$ time in the worst case. This result also applies to the three semantics in Section 3.3 by replacing, in the above arguments, $H_G^-(v)$ and $H_G^+(u)$ with $Q_G^-(u)$ and $Q_G^+(u)$ in Rule 3, and $IB(u,v)$, $II_{G_v}(u,v)$, $\Delta(u,v)$ with $B\_IB(u,v)$ in Rule 2. Since Rule 2 is the most expensive one, it is obvious that efficient algorithms for computing the skyline of a probabilistic relation should postpone Rule 2 as much as possible, so as to check it the minimum number of times.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

5 Final Discussion

It is known [5], that different ranking semantics yield quite different results for top-k queries. We show through an example that this is also the case when such semantics are used for skyline queries. Consider a traffic-monitoring application that collects data by means of a set of radars, a sample of last-hour recording being shown in Figure 1. Because of radar locations, a same car cannot be detected by two radars within an interval of one hour. Each radar reading has associated a Prob value, representing the confidence one has in the reading.

TID	Plate No	Radar	Time	Speed	Prob
t_1	X-123	L1	10:53	90	0.2
t_2	W-246	L2	10:50	100	0.15
t_3	W-246	L3	10:40	95	0.1
t_4	Z-456	L1	10:32	110	0.1
t_5	Z-456	L2	10:30	130	0.3
t_6	H-121	L3	10:30	110	0.2
t_7	Y-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}	C-442	L5	10:10	120	0.3
t_{11}	C-442	L2	10:05	140	0.1

Ranking semantics	$\text{SKY}(R^p)$
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	$\{t_1, t_5\}$

Fig. 1. A probabilistic relation (left) and skylines (rights)

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 $\text{SKY}(R) = \{t_1, t_2, t_4, t_5, t_{11}\}$, and Figure 1 (right) shows the skylines of R^p for the different ranking semantics. Although t_5 is part of all skylines, this is not the case for other tuples (e.g., t_7). As in [5], 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.

References

1. Agrawal, P., Benjelloun, O., Sarma, A.D., Hayworth, C., Nabar, S.U., Sugihara, T., and Widom, J.: Trio: A System for Data, Uncertainty, and Lineage. In: VLDB 2006. pp. 1151–1154. Seoul, Korea (Sep 2006)
2. Bartolini, I., Ciaccia, P., Patella, M.: Getting the Best from Uncertain Data. In: SEBD 2011. pp. 9–20. Maratea, Italy (Jun 2011)
3. Cormode, G., Li, F., Yi, K.: Semantics of Ranking Queries for Probabilistic Data and Expected Ranks. In: ICDE 2009. pp. 305–316. Shanghai, China (Apr 2009)
4. Dalvi, N.N. and Suciu, D.: Efficient Query Evaluation on Probabilistic Databases. In: VLDB 2004. pp. 864–875. Toronto, ON (Aug 2004)
5. Li, J., Saha, B., Deshpande, A.: A Unified Approach to Ranking in Probabilistic Databases. In: VLDB 2009. pp. 502–513. Lyon, France (Aug 2009)
6. Sarma, A.D., Benjelloun, O., Halevy, A.Y., Widom, J.: Working Models for Uncertain Data. In: ICDE 2006. Atlanta, GA (Apr 2006)
7. Soliman, M.A., Ilyas, I.F., Chang, K.C.C.: Top-k Query Processing in Uncertain Databases. In: ICDE 2007. pp. 896–905. Istanbul, Turkey (Apr 2007)
8. Yan, D. and Ng, W.: Robust Ranking of Uncertain Data. In: DASFAA 2011. pp. 254–268. Hong Kong, China (Apr 2011)
9. Zhang, X., Chomicki, J.: On the Semantics and Evaluation of Top-k Queries in Probabilistic Databases. In: DBRank 2008. pp. 556–563. Cancun, Mexico (Apr 2008)

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,