A

Domination in the Probabilistic World: Computing Skylines for Arbitrary Correlations and Ranking Semantics

ILARIA BARTOLINI, PAOLO CIACCIA, and MARCO PATELLA, Università di Bologna

In a probabilistic database, deciding if a tuple u is *better* than another tuple v has not a univocal solution, rather it depends on the specific *probabilistic ranking semantics* (PRS) one wants to adopt so as to combine together tuples' scores and probabilities.

In deterministic databases it is known that skyline queries are a remarkable alternative to (top-k) ranking queries, because they remove from the user the burden of specifying a scoring function that combines values of different attributes into a single score. The skyline of a deterministic relation R is the set of *undominated* tuples in R – tuple u dominates tuple v iff on all the attributes of interest u is better than or equal to v and strictly better on at least one attribute. Domination is equivalent to having $s(u) \geq s(v)$ for *all* monotone scoring functions $s()$.

The skyline of a probabilistic relation R^p can be similarly defined as the set of *P-undominated* tuples in R^p, where now u P-dominates v iff, whatever monotone scoring function one would use to combine the skyline attributes, u is reputed better than v by the PRS at hand. This definition, which is applicable to arbitrary ranking semantics and probabilistic correlation models, is parametric in the adopted PRS, thus it ensures that ranking and skyline queries will always return consistent results.

In this paper we provide an overall view of the problem of computing the skyline of a probabilistic relation. We show how, under mild conditions that indeed hold for all known PRS's, checking P-domination can be cast into an optimization problem, whose complexity we characterize for a variety of combinations of ranking semantics and correlation models. For each analyzed case we also provide specific *P-domination rules*, which are exploited by the algorithm we detail for the case where the probabilistic model is known to the query processor. We also consider the case in which the probability of tuple events can only be obtained through an oracle, and describe another skyline algorithm for this loosely-integrated scenario. Our experimental evaluation of P-domination rules and skyline algorithms confirms the theoretical analysis.

Categories and Subject Descriptors: H.2.4 [**Database Management Systems**]: Query processing

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Skyline queries, Probabilistic database, Ranking semantics

ACM Reference Format:
Ilaria Bartolini, Paolo Ciaccia, and Marco Patella, 2014. Domination in the Probabilistic World: Computing Skylines for Arbitrary Correlations and Ranking Semantics. *ACM Trans. Datab. Syst.* 39, 2, Article A (February 2014), 45 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Since their introduction to the database community in 2001 [Börzsönyi et al. 2001], skyline queries have gained full relevance as a valid support in multi-criteria decision analysis, due to their ability to extract the "most interesting" tuples from a relation R. Given a set $\mathcal{A}$ of numerical attributes of interest, a tuple $t \in R$ is a skyline tuple iff it

Authors' address: Alma Mater Studiorum – Università di Bologna, DISI, Viale Risorgimento, 2 - 40136, Bologna, Italy; email: {i.bartolini, paolo.ciaccia, marco.patella}@unibo.it.

DOI: http://dx.doi.org/10.1145/0000000.0000000

is *undominated*, i.e., there is no other tuple $t' \in R$ that is as good as t on all attributes in $\mathcal{A}$ and strictly better than t on at least one attribute. Skyline queries are a remarkable alternative to top-k ranking queries (which return the k best tuples according to a numerical scoring function), since they require no parameters to be specified and are insensitive to attributes' scales. Further, since the skyline consists of all the top-1 tuples that would result from using *all* scoring functions that are monotone in the skyline attributes $\mathcal{A}$, it can provide users with an overall view of all potential best-ranked objects.

Due to the needs of many emerging applications dealing with inherently *uncertain data*, such as sensor networks [Chong and Kumar 2003; Balazinska et al. 2007], data integration and cleaning [Dong et al. 2007], record lineage [Agrawal et al. 2006; Benjelloun et al. 2008], spatio-temporal and scientific data management [Emrich et al. 2012], to name a few, the issue of extending ranking and skyline queries to such scenarios has been recently investigated.

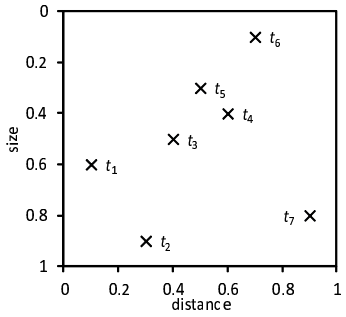
Based on the commonly adopted *possible worlds semantics*, for which a probabilistic relation R^p represents a set of possible worlds, each world W being a subset of R^p, works on top-k queries have first faced the basic problem of providing an adequate *probabilistic ranking semantics* (PRS) able to combine together tuples' scores and probabilities [Soliman et al. 2008; Zhang and Chomicki 2008; Cormode et al. 2009; Li et al. 2009]. Although at a first glance all the proposed PRS's seem to consider a "good match" a tuple t that has both a high score and a high probability, a deeper analysis reveals significant differences among such PRS's. Further, subsequent studies have argued that no single PRS appears to be suitable for all kind of applications, rather the choice should depend on the specific case at hand [Zhang and Chomicki 2009; Li et al. 2011].

The study of skyline queries for uncertain data has been pioneered by Pei et al. [2007], who also introduce the concept of *skyline probability*. The skyline probability, $\Pr_{\text{SKY}}(t)$, of a tuple t is the probability that t is undominated, that is, $\Pr_{\text{SKY}}(t)$ equals the cumulative probability of all the possible worlds W in which t belongs to the skyline of W. The *p-skyline* of a relation R^p is then defined as the set of those tuples t such that $\Pr_{\text{SKY}}(t) \geq p$, where p is a user-defined probability threshold.

Example 1.1. The International Ice Patrol (IIP) Iceberg Sightings Dataset[1] collects information on iceberg activities in the North Atlantic Ocean. The mission is to monitor iceberg danger, plotting and predicting iceberg drift, and broadcasting all known ice to prevent icebergs threatening. Each of the 180,127 records of the IIP database represents a sighting of a single iceberg and contains, among other attributes, the iceberg location (in terms of latitude and longitude), the size of the iceberg, and the reported confidence of the sighting (probability value). One might be interested in determining the most dangerous icebergs, e.g., to minimize the risk of a collision to a relevant man-made installation, like an oil platform. Figure 1(a) shows a sample of 7 points drawn from the IIP dataset, represented in the normalized plane (distance to a target point, size). The deterministic skyline would include those icebergs which are the largest and the closest to the target point. These are t_1 and t_2 in the example. When also the sighting confidence is taken into account, and assuming independence of observations, the skyline probabilities are as in Figure 1(b). Notice that, even if t_2 dominates t_7, it is $\Pr_{\text{SKY}}(t_7) > \Pr_{\text{SKY}}(t_2)$.

The p-skyline approach, although useful to discover tuples with a high skyline probability, has some intrinsic limitations. First, it forces the user to specify a threshold, for which a meaningful value can only be obtained if the distribution of skyline prob-

[1] http://nsidc.org/data/g00807.html.



(a)

t	$p(t)$	$\Pr_{\text{SKY}}(t)$
t_1	0.4	0.4
t_2	0.25	0.25
t_3	0.5	0.225
t_4	0.4	0.09
t_5	0.6	0.135
t_6	0.4	0.022
t_7	0.35	0.263

(b)

Fig. 1. (a) Some tuples from the IIP database; (b) Skyline probabilities.

abilities is known in advance, which is rarely the case. Further, both Atallah et al. [2011] and Zhang et al. [2012] have recently argued, albeit from different points of view, that even tuples with a not-so-high skyline probability might be of interest to users. A second limitation comes from the observation that p-skylines are completely detached from the PRS one wants to use for ranking queries. Thus, it can be the case that the top-1 tuple resulting from a given PRS after applying a scoring function to the skyline attributes has a low skyline probability and, vice versa, that a tuple with a high skyline probability is ranked low by that PRS. This mismatch between the two query types indeed represents a major departure from the deterministic case.

To obviate the limits of p-skylines, Bartolini et al. [2013] have recently introduced the concept of *P-domination*, i.e., domination among probabilistic tuples, and defined the skyline of R^p as the set of those tuples that are P-undominated: For a given PRS, a tuple u P-dominates another tuple v if, for any monotone scoring function, u is ranked higher (i.e., better) than v by that PRS. This definition extends to uncertain data the intuitive idea that (P-)domination amounts to being *always* better, i.e., no matter how different attributes are weighted. The key features of P-domination are that it is applicable to any probabilistic model and is parametric in the PRS one wants to adopt. The latter guarantees that, as in the deterministic case, the resulting skyline will contain all the top-1 tuples resulting from applying such PRS.

Example 1.1. (cont.) Assume the user decides to commit herself to the *expected rank* PRS [Cormode et al. 2009], whose definition is given in Section 5. Depending on the scoring function used to combine the distance and size attributes, the top-1 result obtained from this PRS would be either t_1 or t_3. These two tuples are also the only P-undominated ones, i.e., the skyline based on P-domination is $\{t_1, t_3\}$. Notice that t_3 is ranked only 4th by the p-skyline model. Further, t_3 P-dominates t_7 (whose skyline probability is second only to that of t_1), that is, the expected rank PRS will always rank t_3 better than t_7, regardless of how attribute values are combined together.

1.1. Contributions and Paper Outline

Starting from the definition of P-domination, in this paper we tackle the problem of studying how to compute the skyline of a probabilistic relation for arbitrary correlation models and ranking semantics. To this end, we begin by formalizing P-domination as an optimization problem, called INDARRANGE, which is made possible provided the ranking semantics satisfies two basic properties, which we prove are indeed shared by all commonly used PRS's. We study the complexity of the INDARRANGE problem under a variety of combinations of ranking semantics and probabilistic models, including the

most general case in which the probability of events can only be obtained through an oracle. By considering specific probabilistic models we are able to understand if, and in case how, the query processor can take advantage of this knowledge for computing more efficiently the query result. This *tight integration* between query processing and probabilities can result in effective ad-hoc strategies that would not be possible otherwise. On the other hand, a *loose architecture*, in which the query processing algorithms ignore what the probabilistic model is, allows for greater flexibility and for the analysis of general-purpose solutions, as already observed in [Soliman et al. 2008].

This paper is a prosecution of our work on skyline queries for probabilistic databases, whose first results have been published in [Bartolini et al. 2013]. The concept of P-domination has been introduced in [Bartolini et al. 2013], in which basic formal results on the properties of the resulting skyline are also proved. Besides that, in that paper we focused on specific algorithmic issues, such as sorting and indexing tuples, peculiar to the expected rank semantics, thus ignoring the problems arising from arbitrary correlation models and ranking semantics.

Summarizing, this paper provides the following major contributions:

(1) We show how the problem of checking P-domination can always be reduced to solving an optimization problem (IndArrange), regardless of the model of tuple correlation and provided the adopted ranking semantics satisfies two basic properties. In order to understand how the complexity of IndArrange depends on aspects peculiar to the problem of checking P-domination, rather than also on the difficulty of computing probabilities of arbitrary tuple events and probabilistic scores, we introduce a variant of IndArrange, called IndArrange^P, in which an oracle able to compute any probability and score in $\mathcal{O}(1)$ time exists.
(2) We provide a set of major results that apply to the most commonly adopted semantics for ranking probabilistic tuples and *regardless of the probabilistic database model*: a) For the expected rank (ER) semantics we show that the IndArrange^P problem has complexity $\mathcal{O}(\text{poly}(M))$, where M is the number of tuples that are indifferent to both tuples under comparison. From a practical point of view this implies that, if probabilities can be efficiently computed, then this is also the case for the skyline; b) For the U-Topk, U-kRanks, and Global-Topk semantics the result we provide has a different flavor: There are correlation models for which probabilities can be efficiently computed, yet IndArrange^P is NP-hard. This suggests that for such semantics a careful choice of the probabilistic model is required; c) Finally, we show that the NP-hardness result also applies to a relevant class of parameterized ranking functions introduced in [Li et al. 2011], namely PRF^e.
(3) For specific combinations of ranking semantics and probabilistic models we derive ad-hoc *P-domination rules*, i.e., sufficient and possibly necessary conditions that guarantee that P-domination occurs. Intuitively, tightly coupling the skyline logic with the probabilistic model allows for a more effective, albeit less general, approach, which in some cases can avoid solving at all the IndArrange problem. With the aim of providing a more general characterization of how ranking semantics can influence the problem complexity, we also provide results for the "easy" case (that generalize the observations done for the ER semantics), and for a "hard" case, in which testing P-domination is difficult even when no correlation is present among tuples.
(4) We propose general-purpose algorithms for the tight and the loose integration scenarios.
(5) We extensively test above algorithms so as to understand the impact of different data characteristics, ranking semantics, and probabilistic models, on the execution costs of IndArrange, P-domination tests, and skyline computation. We also

demonstrate how the skyline changes when a different PRS and/or correlation model is adopted. The experimental results we obtain confirm the theoretical analysis on PRS's and correlation models.

The rest of the paper is organized as follows. Section 2 provides background material on probabilistic databases and skyline queries, and reviews related works on skyline queries in probabilistic databases. In Section 3 we briefly review the notion of P-domination and introduce some basic properties of ranking semantics that allow for reducing P-domination to an optimization problem. Section 4 describes the two scenarios we consider, namely loose and tight integration between the query processor and the module in charge of computing probabilities, and details skyline algorithms for both. Section 5 enters into the details of known ranking semantics, for each of them providing general results valid for all probabilistic models as well as specific rules for selected cases. In Section 6 we present extensive experimental results, and in Section 7 we draw our conclusions.

2. BACKGROUND AND RELATED WORK

2.1. Uncertain Relations and Probabilistic Models

We adopt the well-known uncertain data model based on *possible worlds semantics*, in which uncertainty is represented through probabilistic relations [Dalvi and Suciu 2004]. A probabilistic relation, denoted R^p, represents a set of standard relations, each called a *possible world*. We consider a tuple-level uncertainty model, in which each tuple t in R^p has an associated *existence probability*, $p(t) \in (0, 1]$, that represents the probability that t belongs to the database. Sometimes, it is useful to view R^p as a pair, $R^p = (R, p)$, where R is the deterministic part of R^p, i.e., a relation in the standard sense in which tuple probabilities are ignored.

Tuples in R^p can be subject to a variety of *correlation constraints* $\mathcal{C}$ (e.g., mutual exclusion/coexistence of two tuples), whose specific form actually depends on the adopted probabilistic model. Given $\mathcal{C}$, a possible world W of R^p is a subset of tuples of R that respects all the constraints dictated by $\mathcal{C}$, and $\Pr_{\mathcal{C}}(W)$ is its probability, i.e., the probability of the complex event "all and only the tuples in W exist". This is a conjunction of simple events of type "tuple t exists", simply denoted as t, or "tuple t does not exist", denoted as $\neg t$. We denote by $\mathcal{W}$ the set of all possible worlds that can be formed from R^p according to $\mathcal{C}$.

It is well known [Das Sarma et al. 2006] that there is an intrinsic tension among the expressiveness of the probabilistic model (i.e., which possible worlds it can represent) and its complexity (both in terms of query answering and ease of use).

The case of *independent* tuples, $\mathcal{C} = \emptyset$, is the simplest to consider, since the probability of any complex event can be completely factorized into a product of tuple probabilities. For instance, the probability of a possible world W can be simply expressed as $\Pr(W) = \prod_{t \in W} p(t) \prod_{t \notin W} (1 - p(t))$.

In the widely adopted *x-relation* model [Agrawal et al. 2006], $\mathcal{C}$ only includes mutual exclusion rules that form a partition of R (i.e., mutual exclusion is a transitive relation). Let G, also called an x-tuple (or group), denote a maximal set of mutually exclusive tuples (with overall probability ≤ 1). The intuition about this model is that each group corresponds to an *uncertain object*, whose (discrete) distribution is represented by the tuples in that group. The probabilities of complex events can be computed by observing that if t_1 and t_2 belong to the same group, then $\Pr(t_1 \wedge t_2) = 0$ and $\Pr(t_1 \wedge \neg t_2) = \Pr(t_1) \stackrel{\text{def}}{=} p(t_1)$. Therefore, the probability of a possible world W can be

written as:

$$\Pr(W) = \prod_{\substack{G_i \cap W = \{t_{j_i}\} \\ G_i \in \mathcal{G}}} p(t_{j_i}) \prod_{\substack{G_i \cap W = \emptyset \\ G_i \in \mathcal{G}}} \left(1 - \sum_{t \in G_i} p(t)\right)$$

where $\mathcal{G}$ is the set of all groups in R^p.

Common to the above models is the observation that the existence probability of each tuple is given in advance, and only probabilities of complex events need to be computed. This no longer holds for more complex models, such as *Bayesian networks* [Pearl 1988] and *Markov networks* [Cowell et al. 1999], in which correlations among tuples are captured by the topology of a graph. In both cases computing probabilities of arbitrary events is a complex task (#P-hard), with best-known algorithms based on a *junction tree* representation of the graph having complexity $\mathcal{O}(N2^{tw})$, where N is the number of tuples in R^p and tw is the *treewidth* of the junction tree.

2.2. Deterministic Skyline and Scoring Functions

The *skyline* of a (deterministic) relation R with respect to a set of numerical attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_d\}$ is the set of those tuples in R that are *undominated* on $\mathcal{A}$. Without loss of generality, assume that on each attribute higher values are preferable. Then, 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, $\textsc{Sky}(R) = \{u \in R \mid \nexists\, v \in R : v \succ u\}$. We say that u and v are *indifferent* to each other, written $u \sim v$, when neither $u \succ v$ nor $v \succ u$ hold. Clearly, all tuples in $\textsc{Sky}(R)$ are pairwise indifferent. It is well known that $\succ$ is irreflexive ($\forall u : u \not\succ u$) and transitive ($\forall u, v, t : u \succ v \wedge v \succ t \Rightarrow u \succ t$), thus it is a *strict partial order*.

A *linear order* $\succcurlyeq$ is a strict partial order that is also *connected*, i.e., for any two distinct tuples u and v, either $u \succcurlyeq v$ or $v \succcurlyeq u$. A linear order $\succcurlyeq$ is called a *linear extension* of $\succ$ iff $u \succ v \Rightarrow u \succcurlyeq v$, that is, $\succcurlyeq$ is *compatible* with $\succ$. The *rank* of tuple u in the linear order $\succcurlyeq$, denoted $\text{rank}^{\succcurlyeq}(u)$, is the number of tuples preceding u in $\succcurlyeq$: $\text{rank}^{\succcurlyeq}(u) = |\{t \mid t \succcurlyeq u\}|$.

A *scoring function* $s()$ on the attributes $\mathcal{A}$ is *monotone* iff $u.A_i \geq v.A_i$ ($i = 1, \ldots, d$) implies $s(u) \geq s(v)$. Notice that by ordering tuples with a monotone function and then breaking ties by respecting possible domination relationships yields a linear extension $\succcurlyeq$ of $\succ$. Finally, observe that for any two tuples u and v that differ on at least one skyline attribute, it is $u \succ v$ iff for all monotone scoring functions $s()$ it is $s(u) \geq s(v)$. Hereafter, we always implicitly assume that all tuples are distinct on $\mathcal{A}$.[2]

2.3. Related Work

The research on skyline queries on deterministic databases has spanned different facets of the problem, ranging from the development of efficient algorithms (see, among others, [Börzsönyi et al. 2001; Chomicki et al. 2003; Papadias et al. 2005; Godfrey et al. 2005; Bartolini et al. 2008; Zhang et al. 2009]), to extensions to a variety of advanced scenarios, among which: Data streams [Tao and Papadias 2006], distributed [Trimponias et al. 2013] and parallel [Afrati et al. 2012] environments, subspaces [Yuan et al. 2005], and low-cardinality domains [Morse et al. 2007]. Chomicki et al. [2013] survey these and other skyline-related topics.

Skyline queries on uncertain data have been first studied in [Pei et al. 2007], where the concept of *skyline probability* is introduced. The probabilistic model assumed

[2]Considering duplicate $\mathcal{A}$-values has no practical influence on the problems we study, yet it would unnecessarily lengthen the analysis.

in [Pei et al. 2007] is a simplified case of the x-relation model, in that all the m_i tuples in a same group G_i have the same probability, $1/m_i$. The skyline probability of a tuple u is the probability that u exists times the probability that no tuple in other groups and dominating u exists, that is:

$$\Pr_{\text{SKY}}(u) = p(u) \times \prod_{\substack{G_i \in \mathcal{G} \\ G_i \neq G(u)}} \left(1 - \sum_{\substack{t \in G_i \\ t \succ u}} p(t)\right) = \frac{1}{|G(u)|} \times \prod_{\substack{G_i \in \mathcal{G} \\ G_i \neq G(u)}} \left(1 - \sum_{\substack{t \in G_i \\ t \succ u}} \frac{1}{m_i}\right)$$

where $G(u)$ denotes the group of u. The skyline probability of a group G_i is then defined as $\Pr_{\text{SKY}}(G_i) = \sum_{t \in G_i} \Pr_{\text{SKY}}(t)$. Finally, given a user-defined probability threshold p, the *p-skyline* of a relation R^p is the set of those groups with skyline probability $\geq p$. Notice that although Pei et al. [2007] consider a *group-level* p-skyline, i.e., the result is a set of groups, it also makes sense to consider a *tuple-level* p-skyline, in which the result consists of tuples with high skyline probability. In this case it is evident that the concept of skyline probability can be applied to any correlation model.

As already observed in the Introduction, there is no apparent relationship between skyline probabilities and the semantics one wants to adopt for supporting ranking queries. While in the deterministic case one is guaranteed that the best-ranked tuple according to an arbitrary monotone scoring function is part of the skyline, this relevant connection is lost in the probabilistic case, since there is no guarantee that the best tuple according to a certain ranking semantics has also a high skyline probability. The essence of the problem is that the thresholding approach inherent in the concept of p-skyline completely ignores the actual *utility* that single tuples and/or groups may have for a specific user, i.e., how a user might want to weigh the different relevance of skyline attributes.

Based on the above arguments, Atallah et al. [2011] have studied the problem of efficiently computing *all* skyline probabilities, arguing that even objects with a not-so-high probability might be of interest to some user. The subquadratic-time algorithms they develop to this purpose cannot however solve the problem of providing users with a limited set of relevant results, which is what the skyline represents in the deterministic case.

Based on concepts of stochastic domination, that has its roots in economics and decision theory, Lin et al. [2011] and Zhang et al. [2012] have proposed the *stochastic skyline* operator. Given a set of uncertain objects, the *lskyline* contains all the top-1 results when one considers the expected value of multiplicative monotone scoring functions, where expectation is taken over the samples of the uncertain object. More precisely, given monotone functions $f_1, \ldots, f_d$, consider for a tuple u_i the value of $f(u_i) \stackrel{\text{def}}{=} \prod_{j=1}^{d} f_j(u_i.A_j)$. For an uncertain object G, therefore it is $E[f(G)] = \sum_{u_i \in G} f(u_i)p(u_i)$.

Since checking stochastic domination in this scenario is NP-complete, in [Zhang et al. 2012] an alternative skyline definition is introduced, called *gskyline*, for which domination can be computed in polynomial time using a max-flow algorithm. Further, the gskyline always includes the lskyline (and both are incomparable to the p-skyline), since the limitation to multiplicative functions is removed. Although neither lskylines nor gskylines require a threshold parameter to be applied, they are defined only if R^p represents a set of disjoint uncertain objects, thus they cannot be applied to arbitrary probabilistic models. Further, even if both gskyline and lskyline are somehow related to scoring functions, these are defined at the group-level and cannot therefore be used to rank single tuples.

In Table I we summarize the major features of the skyline based on P-domination that we analyze in this paper, and contrast them with p-skyline, gskyline, and lskyline.

Table I. Main features of P-domination-based skyline versus other proposals. (*) For models other than x-relation, the p-skyline is necessarily a set of tuples.

	p-skyline	gskyline, lskyline	P-domination skyline
Parameter-free (no threshold)	No	Yes	Yes
Arbitrary correlation models	Yes(*)	No	Yes
Consistency with ranking queries	No	No	Yes
Skyline result (tuples vs. groups)	Both(*)	Groups	Tuples

3. THE P-DOMINATION PROBLEM

In this section we first briefly review the basic definitions concerning P-domination and the skyline of a probabilistic relation, after that we enter into the details of how checking P-domination can be reduced to solving an optimization problem. Preliminary to our discussion is the concept of *probabilistic ranking semantics*, PRS for short, which is a way to rank the tuples of a probabilistic relation R^p when they are linearly ordered.

Definition 3.1 (*Probabilistic Ranking Semantics – PRS*).
Let R^p be a probabilistic relation whose tuples obey the correlation constraints $\mathcal{C}$, and let $\succcurlyeq$ be a linear ordering of the tuples of R. A *probabilistic ranking semantics* (PRS) Ψ is a function that takes as input R^p, $\mathcal{C}$, and $\succcurlyeq$, and assigns to each tuple u a value $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u)$, called the *probabilistic score* of u. The resulting order based on Ps values is called a *probabilistic order*.

The above definition is a generalization of the ideas underlying commonly adopted semantics for ranking probabilistic tuples, which will be dealt with in Section 5. The intuition about PRS's is that ranking a set of probabilistic tuples is based on a criterion that, in the general case, looks at tuple probabilities, at how tuples are ordered by considering attribute values ($\succcurlyeq$), and at how they are correlated ($\mathcal{C}$). Note that the above definition excludes those semantics that explicitly depend on *score values*, rather than just on the relative ordering of tuples that a scoring function induces on the tuples in R. An extensive discussion on why score values should not be considered for ranking probabilistic tuples can be found in [Jestes et al. 2011], where this property is called *value-invariance*. In Section 5.4.2 we analyze the only known semantics that is *not* value-invariant, namely *expected score* [Cormode et al. 2009], whereas a discussion on what considering scores implies on the results we derive can be found in Appendix B.

As argued in [Li et al. 2009] and [Zhang and Chomicki 2009], different application domains can have peculiar requirements about the properties that a PRS should satisfy, thus no clear winner is likely to exist. For this reason, rather than binding the definition of skyline to a specific PRS, we adopt a *parametric* approach, in which *any* PRS can be used.

Definition 3.2 (*P-domination and Skyline*).
Let R^p be a probabilistic relation, whose tuples are correlated by $\mathcal{C}$, and let $\succ$ be the domination relation on the tuples in R when considering the skyline attributes $\mathcal{A}$. Let Ψ be a PRS. 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 $\succcurlyeq$ of $\succ$ it is $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) \geq \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$, with the inequality being strict for at least one linear extension.

The skyline of R^p (based on Ψ and $\mathcal{C}$) is consequently defined as:

$$\boxed{\textsc{Sky}(R^p) = \{u \in R^p \mid \nexists\, v \in R^p : v \succ_p u\}} \tag{1}$$

i.e., as the set of P-undominated tuples.[3]

[3]To simplify the notation, the dependency of $\succ_p$ and $\textsc{Sky}(R^p)$ on Ψ and $\mathcal{C}$ is understood.

Example 3.3. Let $R = \{t_1, t_2, t_3\}$, with $t_1 \succ t_2$. There are 3 linear orders compatible with $\succ$, denoted $\succcurlyeq_a, \succcurlyeq_b$, and $\succcurlyeq_c$, as shown in Figure 2. Assuming that for each of them a PRS Ψ computes probabilistic scores as in the figure, we have that both t_1 and t_2 P-dominates t_3. For instance, it is $t_1 \succ_p t_3$ since on each linear order the probabilistic score of t_1 is higher than that of t_3. On the other hand, neither $t_1 \succ_p t_2$ nor $t_2 \succ_p t_1$ hold, since $\mathsf{Ps}^{\succcurlyeq_b}_{\Psi,\mathcal{C}}(t_1) > \mathsf{Ps}^{\succcurlyeq_b}_{\Psi,\mathcal{C}}(t_2)$ whereas $\mathsf{Ps}^{\succcurlyeq_c}_{\Psi,\mathcal{C}}(t_2) > \mathsf{Ps}^{\succcurlyeq_c}_{\Psi,\mathcal{C}}(t_1)$. It follows that $\text{SKY}(R^p) = \{t_1, t_2\}$.

$\mathsf{Ps}^{\succcurlyeq_i}_{\Psi,\mathcal{C}}(t_j)$	t_1	t_2	t_3
$\succcurlyeq_a : t_3 \succcurlyeq_a t_1 \succcurlyeq_a t_2$	2.5	2.2	2.2
$\succcurlyeq_b : t_1 \succcurlyeq_b t_3 \succcurlyeq_b t_2$	3.0	2.1	1.2
$\succcurlyeq_c : t_1 \succcurlyeq_c t_2 \succcurlyeq_c t_3$	3.2	4.0	0.5

Fig. 2. Example of P-domination: Numbers represent probabilistic scores, from which both $t_1 \succ_p t_3$ and $t_2 \succ_p t_3$ are derived.

P-domination can be seen as a combination of two major concepts: Deterministic domination, $\succ$, and probabilistic ranking semantics, Ψ. Indeed, it is the case that $u \succ_p v$ iff, whatever monotone scoring function one applies to linearly order the tuples, u will be ranked no worse than v by Ψ.

As proved in [Bartolini et al. 2013], the P-domination relation is a strict partial order, regardless of which ranking semantics Ψ is adopted. Further, if tuples are ordered using a monotone scoring function, and u is the top-1 tuple when the PRS Ψ is used, then u is guaranteed to be part of $\text{SKY}(R^p)$. Thus, as in the deterministic case, the skyline based on P-domination is able to reveal those tuples which are not manifestly inferior to some other tuples and, consequently, provides a concise view of which are, in a probabilistic sense, all the potential best objects.

Definition 3.2 is not practical for checking P-domination, since enumerating all linear extensions of $\succ$ can incur a prohibitively high cost.[4] To obviate this, we first introduce the following key concept:

Definition 3.4 $((u, v)$*-Adversarial Order*$)$.
For given R^p, $\mathcal{C}$, and Ψ, a linear order $\succcurlyeq$ that extends $\succ$ is called a (u, v)-adversarial order if it minimizes the difference $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$.

Clearly, since Definition 3.2 implies that $u \succ_p v$ holds iff it is:[5]

$$\min_{\succcurlyeq} \left\{ \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v) \right\} \geq 0 \tag{2}$$

where each $\succcurlyeq$ extends $\succ$, it follows that analyzing only (u, v)-adversarial orders, i.e., those orders that maximally favor tuple v with respect to tuple u, is all what is needed to check P-domination.

Example 3.5. In Figure 2, the linear order $\succcurlyeq_a$ is a (t_2, t_3)-adversarial order, whereas $\succcurlyeq_c$ is a (t_1, t_2)-adversarial order. Since in the first case the difference is non-negative it is $t_2 \succ_p t_3$, whereas $t_1 \not\succ_p t_2$ since $3.2 - 4.0 < 0$.

[4]The number of linear extensions can be exponential in the number of tuples.
[5]Definition 3.2 also requires that strict inequality holds for at least one linear order $\succcurlyeq$. We always implicitly assume this additional condition, thus avoiding to repeat it each time.

3.1. Properties of Probabilistic Ranking Semantics

In this section we describe how, given two basic properties of PRS's, which are indeed enjoyed by all commonly used ranking semantics, it is possible to search for a (u, v)-adversarial order, thus checking if u P-dominates v, without enumerating linear extensions.

3.1.1. The Case $u \succ v$. We first consider the case $u \succ v$, for which it is guaranteed that $u \succcurlyeq v$ as well. Let t be any tuple in R^p other than u and v. We analyze how t can be related to u and v in terms of the domination/indifference relations and determine how t should be placed relatively to u and v in a (u, v)-adversarial order.

When neither u nor v is indifferent to t, any order $\succcurlyeq$ compatible with $\succ$ will necessarily exhibit one of the following three patterns:

$$t \succ u \succ v \implies t \succcurlyeq u \succcurlyeq v;$$
$$u \succ v \succ t \implies u \succcurlyeq v \succcurlyeq t;$$
$$u \succ t \succ v \implies u \succcurlyeq t \succcurlyeq v.$$

Now, consider two tuples t and t', both dominating u and such that $t \sim t'$. How should they relatively be ordered in a (u, v)-adversarial order? It is plain to see that, if the probabilistic score of u (and v as well) depends on how t and t' are relatively arranged, then there will be no alternative to that of evaluating $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u)$ and $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$ for *all* orders $\succcurlyeq$ that differ in how they rank pairwise indifferent tuples that dominate both u and v. The same observation applies to the other two patterns above, which leads us to introduce the following property:

PROPERTY 1 (SET-DEPENDENCY).
Let R^p be a probabilistic relation, $\succcurlyeq$ a linear ordering of the tuples in R, and Ψ a PRS*. Let u be a tuple of R^p, and let $\mathrm{Up}^{\succcurlyeq}(u) = \{t \in R \,|\, t \succcurlyeq u\}$ be the set of tuples preceding (i.e., better than) u in $\succcurlyeq$. We say that Ψ is* set-dependent *if, given any other linear order $\succcurlyeq'$ with $\mathrm{Up}^{\succcurlyeq'}(t) = \mathrm{Up}^{\succcurlyeq}(t)$, it is $\mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(t) = \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(t)$ for all tuples t and correlation constraints $\mathcal{C}$.*

The set-dependency property captures the intuition that the probabilistic score of a tuple only depends on its probability and attributes' values and on which tuples precede it, but not on their specific ranks. As it will be shown in the following, all known PRS's are indeed set-dependent.

Let us now consider the two cases in which a tuple t is indifferent to only one of u and v, namely $u \succ t, v \sim t$ or $u \sim t, t \succ v$. The following property is the key to decide how to deal with such cases.

PROPERTY 2 (RANK-MONOTONICITY).
Let R^p be a probabilistic relation and Ψ a PRS*. We say that Ψ is* rank-monotone *if, for any tuple u in R^p and any pair of orders $\succcurlyeq$ and $\succcurlyeq'$ such that: 1) $\mathrm{Up}^{\succcurlyeq}(u) \subset \mathrm{Up}^{\succcurlyeq'}(u)$, and 2) $t \succcurlyeq t'$ implies $t \succcurlyeq' t'$ for all $t, t' \in \mathrm{Up}^{\succcurlyeq}(u)$, it is $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) \geq \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(u)$ for all correlation constraints $\mathcal{C}$.*

The rank-monotonicity property guarantees that worsening the rank of a tuple cannot increase its probability score, which is a reasonable behavior indeed shared by all known PRS's. If a PRS Ψ is both set-dependent and rank-monotone, then we concisely say that Ψ is *set-monotone*,[6] in which case it is enough to have $\mathrm{Up}^{\succcurlyeq}(u) \subset \mathrm{Up}^{\succcurlyeq'}(u)$ to infer that $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) \geq \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(u)$.

[6]Set-monotonicity is quite similar to what others have called *stability* in the context of top-k queries [Zhang and Chomicki 2008].

While set-dependency simplifies the search of a (u, v)-adversarial order by guaranteeing that the exact ranking of tuples that are ordered in a certain way with respect to u and v is not relevant, rank-monotonicity is used to determine how to arrange those tuples t that are indifferent to only one of u and v.

LEMMA 3.6. *Let Ψ be a rank-monotone* PRS. *If $u \succ v$, there exists a (u, v)-adversarial order $\succcurlyeq$ such that, for any tuple t in R^p:*

(1) *if $u \succ t$ and $v \sim t$, then $u \succcurlyeq v \succcurlyeq t$;*
(2) *if $u \sim t$ and $t \succ v$, then $t \succcurlyeq u \succcurlyeq v$.*

PROOF. We only prove case (1) above, case (2) requiring similar arguments.

Intuitively, ordering t after v aims to avoid decreasing, due to rank-monotonicity, the probabilistic score of v. Consider any order $\succcurlyeq'$ such that $u \succcurlyeq' t \succcurlyeq' v$. We claim that, for each such $\succcurlyeq'$, there exists an order $\succcurlyeq$ for which it is $u \succcurlyeq v \succcurlyeq t$, and such that $\mathrm{Up}^{\succcurlyeq}(u) = \mathrm{Up}^{\succcurlyeq'}(u)$, $\mathrm{Up}^{\succcurlyeq}(v) \subset \mathrm{Up}^{\succcurlyeq'}(v)$, and the relative ordering of any pair of tuples t, t' in $\mathrm{Up}^{\succcurlyeq}(v)$ is the same in both orders. From this, due to rank-monotonicity, the result follows since $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v) \leq \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(v)$.

Given $\succcurlyeq'$, we can generate $\succcurlyeq$ with the required properties by just moving t after v, and doing the same for all tuples t' such that $t \succ t'$ and $t' \succcurlyeq' v$ (note that it is necessarily $t' \sim v$, since $t \sim v$). Since this process leaves $\mathrm{Up}^{\succcurlyeq'}(u)$ unchanged, whereas it removes tuples from $\mathrm{Up}^{\succcurlyeq'}(v)$, the result is proved. □

If a PRS is *not* rank-monotone, it is possible that a (u, v)-adversarial order violates Lemma 3.6. This is shown through a simple example that makes use of a rather strange ranking semantics, called EVEN.

Example 3.7. Regardless of the correlation model, a fact hereafter denoted by $\mathcal{C} = \bigstar$, the EVEN PRS assigns to tuple u the probabilistic score:

$$\mathsf{Ps}^{\succcurlyeq}_{\text{EVEN},\bigstar}(u) = \begin{cases} p(u)(N - \mathrm{rank}^{\succcurlyeq}(u)) & \text{if } \mathrm{rank}^{\succcurlyeq}(u) \text{ is even} \\ p(u)(N - \mathrm{rank}^{\succcurlyeq}(u))/3 & \text{otherwise} \end{cases}$$

Recall that $\mathrm{rank}^{\succcurlyeq}(u) = |\{t \mid t \succcurlyeq u\}|$ is the rank of u in the linear order $\succcurlyeq$ and N is the cardinality of R^p. Notice that EVEN is set-dependent but *not* rank-monotone, since increasing (i.e., worsening) the rank of a tuple can lead to increase its probabilistic score as well (in particular, when passing from an odd to an even rank).

Consider case (1) in Lemma 3.6. We show that given two orders $\succcurlyeq$ and $\succcurlyeq'$, such that $u \succcurlyeq v \succcurlyeq t$ and $u \succcurlyeq' t \succcurlyeq' v$, the first one is not necessarily a (u, v)-adversarial order, thus contradicting Lemma 3.6. The relation R^p we consider to this purpose has $N = 3$ tuples: u, v, and t, with probabilities 0.25, 0.9, and 0.5, respectively. By hypothesis, it is $u \succ v$ and $u \succ t$, thus the orders $\succcurlyeq$ and $\succcurlyeq'$ as described above are the only ones compatible with $\succ$. Since $\mathrm{rank}^{\succcurlyeq}(u) = \mathrm{rank}^{\succcurlyeq'}(u) = 0$, the score of u will be $0.25(3-0) = 0.75$ in both cases. On the other hand, it is $\mathsf{Ps}^{\succcurlyeq}_{\text{EVEN},\bigstar}(v) = 0.9(3-1)/3 = 0.6$, since v has rank 1 in $\succcurlyeq$, whereas $\mathsf{Ps}^{\succcurlyeq'}_{\text{EVEN},\bigstar}(v) = 0.9(3-2) = 0.9$ when v has rank 2 in $\succcurlyeq'$. It follows that $\succcurlyeq$ is not a (u, v)-adversarial order.

The only remaining case to be analyzed when $u \succ v$ involves those tuples t that are indifferent to both u and v. Unlike the two cases covered by Lemma 3.6, now the choice of how to arrange such tuples in a (u, v)-adversarial order is partially undetermined, since any of such tuples can be ordered either before u (thus lowering the probabilistic scores of both u and v) or after v. The third possible case, namely t ordered *between* u and v, is excluded by the following lemma.

LEMMA 3.8. *Let Ψ be a rank-monotone* PRS. *If $u \succ v$ and t is a tuple such that $u \sim t$ and $v \sim t$, then, for each order $\succcurlyeq'$ with $u \succcurlyeq' t \succcurlyeq' v$ there exists a (u, v)-adversarial order $\succcurlyeq$ for which it is either $t \succcurlyeq u \succcurlyeq v$ or $u \succcurlyeq v \succcurlyeq t$.*

PROOF. The proof is based on arguments that closely follow those in the proof of Lemma 3.6. For an order $\succcurlyeq'$ such that $u \succcurlyeq' t \succcurlyeq' v$ we can modify $\succcurlyeq'$ in one of the following ways:

(1) Move t before u, together with all tuples t' such that $t' \succ t$ and $u \succcurlyeq' t'$ (here it is $t' \sim u$, since $t \sim u$). Since this process leaves $\mathrm{Up}^{\succcurlyeq'}(v)$ unchanged, whereas it adds tuples to $\mathrm{Up}^{\succcurlyeq'}(u)$, the difference between the probabilistic scores of u and v cannot increase due to this transformation.
(2) Move t after v, together with all tuples t' such that $t \succ t'$ and $t' \succcurlyeq' v$ (here it is $t' \sim v$, since $t \sim v$). Since this process leaves $\mathrm{Up}^{\succcurlyeq'}(u)$ unchanged, whereas it removes tuples from $\mathrm{Up}^{\succcurlyeq'}(v)$, also in this case it will be $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v) \leq \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq'}_{\Psi,\mathcal{C}}(v)$.

□

Although Lemma 3.8 partially cuts down the search space for a (u, v)-adversarial order, it leaves open the possibility that in such orders there will be some tuples t, indifferent to both u and v, that will precede both while others will follow both. More precisely, let $\mathrm{IND}(u, v) = \{t_1, t_2, \ldots, t_M\}$ be the set of tuples indifferent to both u and v. If t_i and t_j are in $\mathrm{IND}(u, v)$, and it is the case that $t_i \succ t_j$, then only the following cases can occur in an order $\succcurlyeq$ that extends $\succ$ (thus, also in a (u, v)-adversarial order):

(1) $t_i \succcurlyeq t_j \succcurlyeq u \succcurlyeq v$;
(2) $t_i \succcurlyeq u \succcurlyeq v \succcurlyeq t_j$;
(3) $u \succcurlyeq v \succcurlyeq t_i \succcurlyeq t_j$.

In particular, it cannot be the case that t_j precedes u whereas t_i follows v. More in general, the domination relationships among the tuples in $\mathrm{IND}(u, v)$ force the subset of tuples in $\mathrm{IND}(u, v)$ that will be ordered before u to be an *upper set* of $\mathrm{IND}(u, v)$. This is formalized by the following:

THEOREM 3.9 (THE INDARRANGE PROBLEM).
Let u and v be two tuples in R^p such that $u \succ v$, and let $\mathrm{IND}(u, v) = \{t_1, t_2, \ldots, t_M\}$ be the set of tuples indifferent to both u and v. For each $t_i \in \mathrm{IND}(u, v)$, define a binary variable y_i, where $y_i = 1$ if tuple t_i is arranged before u, whereas $y_i = 0$ if t follows v. Assume that the PRS *Ψ is set-monotone and that all tuples indifferent to only one of tuples u and v are arranged according to Lemma 3.6. Further, let $\succcurlyeq_{\mathbf{Y}}$ be an order corresponding to a specific $\mathbf{Y} = \langle y_1, \ldots, y_M \rangle$ vector. If $\succcurlyeq_{\mathbf{Y}}$ is a solution to the following* INDARRANGE *problem:*

$$\begin{array}{lll} \textit{minimize} & \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{\Psi,\mathcal{C}}(v) & \\ \textit{subject to} & y_i \in \{0, 1\} & i \in [1..M] \\ & y_i \geq y_j \textit{ if } t_i \succ t_j & i, j \in [1..M] \end{array} \tag{3}$$

then $\succcurlyeq_{\mathbf{Y}}$ is a (u, v)-adversarial order.

PROOF. Since Ψ is set-dependent, all orders corresponding to a specific $\mathbf{Y}$ vector will yield the same probabilistic scores for both u and v. Since $\succcurlyeq_{\mathbf{Y}}$ minimizes their difference by respecting all domination relationships among the tuples in $\mathrm{IND}(u, v)$ (it is $y_i \geq y_j$ if $t_i \succ t_j$) the result follows. □

Theorem 3.9 allows the structure of a (u, v)-adversarial order to be precisely characterized as follows, in which the relevant sets of tuples are made explicit (see also Figure 3):

— Tuples preceding u: These influence the probabilistic scores of both u and v:
  — $\text{BTR}(u, v)$: the tuples better than u (thus, also better than v);
  — $\text{INDBTR}(u, v)$: the tuples indifferent to u and better than v (see Lemma 3.6);
  — $\text{IND\_1}(u, v)$: the tuples in $\text{IND}(u, v)$ for which solving INDARRANGE yields $y_i = 1$;
— Tuples following u and preceding v: These only influence the probabilistic score of v:
  — $\text{WRSBTR}(u, v)$: the tuples worse than u and better than v;
— Tuples following v: These have no influence on the probabilistic scores of u and v:
  — $\text{WRS}(u, v)$: the tuples worse than v (thus, also worse than u);
  — $\text{WRSIND}(u, v)$: the tuples worse than u and indifferent to v (see Lemma 3.6);
  — $\text{IND\_0}(u, v)$: the tuples in $\text{IND}(u, v)$ for which solving INDARRANGE yields $y_i = 0$.

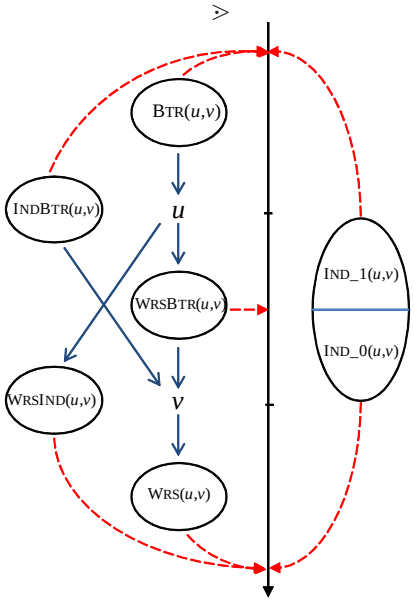


Fig. 3. The structure of a (u, v)-adversarial order. Solid arrows represent domination among tuples while dashed arrows indicate where sets of tuples should be arranged wrt u and v. Better ranked tuples are shown higher in the figure.

The relevance of Theorem 3.9 is twofold. First, it shows that the structure of a (u, v)-adversarial order depends neither on the ranking semantics nor on the probabilistic model, a fact which at first sight might not have been so obvious. Second, it makes evident what is the only part that can influence the complexity of checking P-domination. Indeed, the complexity of the INDARRANGE problem, which ultimately determines that of P-domination, is dependent on three factors:

(1) The complexity of computing probabilities of arbitrary tuple events. It is clear that if probabilistic inference is in itself a complex problem, this propagates to skyline computation, as it happens for other query types [Dalvi et al. 2011].
(2) The cost of computing probabilistic scores, i.e., the complexity of the PRS Ψ.

(3) Assuming that both probabilities and probabilistic scores can be easily computed, how does the complexity depend on the number M of tuples in $\text{IND}(u, v)$?

Since according to Theorem 3.9 the latter is the only part peculiar to the problem of checking P-domination, for the only purpose of complexity analysis we find it convenient to introduce the following variant of the INDARRANGE problem:

Definition 3.10 (*The* INDARRANGE^P *problem*).
The INDARRANGE^P problem is the same as INDARRANGE, with the only variant that the probability of an arbitrary tuple event and the probabilistic score of an arbitrary tuple can be obtained from an oracle in constant time.

Let $M_{\max}$ be the maximum cardinality of $\text{IND}(u, v)$ taken over all pairs of tuples u and v in R^p such that $u \succ v$. From Theorem 3.9 and the above definition, it then follows that the time complexity of solving INDARRANGE^P is a function of $M_{\max}$ only.

Notice that, even with the assumption that computing probabilistic scores is easy (as it is indeed the case for known PRS's), the complexity of INDARRANGE^P largely depends on the adopted PRS, as it will be demonstrated in Section 5. In particular, in that section we will exhibit scenarios for which the complexity of INDARRANGE^P varies from constant-time to exponential-time, even when tuples are independent.

3.1.2. The Case $u \sim v$. When $u \sim v$ the minimization of the difference $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$ can be performed by splitting the problem into the two sub-problems of minimizing $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u)$ and maximizing $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$, which can be independently solved.

To this end, we can exploit an obvious consequence of the set-monotonicity property, for which the following preliminary definition is needed.

Definition 3.11 (*Bounds on the Up set*).
Let R^p be a probabilistic relation. For any tuple u in R^p, the sets $\text{Up}^-(u)$ and $\text{Up}^+(u)$ are defined as:

$$\text{Up}^-(u) = \{t \in R \,|\, t \succ u\} \tag{4}$$

$$\text{Up}^+(u) = \{t \in R \,|\, u \not\succ t, t \neq u\} \tag{5}$$

The two following lemmas directly follow from the above definition:

LEMMA 3.12. *For any linear extension $\succcurlyeq$ of $\succ$, the set $\text{Up}^{\succcurlyeq}(u)$ of the tuples preceding u in $\succcurlyeq$ satisfies $\text{Up}^-(u) \subseteq \text{Up}^{\succcurlyeq}(u) \subseteq \text{Up}^+(u)$.*

LEMMA 3.13. *Let Ψ be a* PRS. *For any tuple u in R^p and any correlation constraints $\mathcal{C}$, let $\mathsf{Ps}^-_{\Psi,\mathcal{C}}(u)$ (respectively $\mathsf{Ps}^+_{\Psi,\mathcal{C}}(u)$) be the minimum (resp. maximum) value that the probabilistic score of u can attain by considering all linear orders $\succcurlyeq$ that extend $\succ$, i.e., $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) \in [\mathsf{Ps}^-_{\Psi,\mathcal{C}}(u), \mathsf{Ps}^+_{\Psi,\mathcal{C}}(u)]$. If Ψ is set-monotone, then:*

- $\mathsf{Ps}^-_{\Psi,\mathcal{C}}(u)$ *is obtained when* $\text{Up}^{\succcurlyeq}(u) = \text{Up}^+(u)$;
- $\mathsf{Ps}^+_{\Psi,\mathcal{C}}(u)$ *is obtained when* $\text{Up}^{\succcurlyeq}(u) = \text{Up}^-(u)$.

THEOREM 3.14.
Let u and v be two tuples in R^p such that $u \sim v$. If the probabilistic ranking semantics Ψ is set-monotone and $\succcurlyeq$ is a linear order that extends $\succ$ such that, for all tuples $t \in R^p$: (i) $t \succcurlyeq v$ iff $t \succ v$, and (ii) $u \succcurlyeq t$ iff $u \succ t$, then $\succcurlyeq$ is a (u, v)-adversarial order. It follows that $u \succ_p v$ iff $\mathsf{Ps}^-_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^+_{\Psi,\mathcal{C}}(v) \geq 0$.

PROOF. The result directly follows from Lemma 3.13. □

Notice that, since both $\mathsf{Ps}^{-}_{\Psi,\mathcal{C}}(u)$ and $\mathsf{Ps}^{+}_{\Psi,\mathcal{C}}(u)$ only depend on u, it is advisable to *precompute* and store these bounds for all tuples, rather than recomputing them each time they are needed. Doing so, a single P-domination check in the case $u \sim v$ would thus require only $\mathcal{O}(1)$ time.

4. SCENARIOS FOR SKYLINE COMPUTATION

In this section we provide details on the two operating scenarios we consider, namely loose and tight integration, and describe corresponding skyline algorithms. For the tight integration scenario, in which the skyline algorithm is aware of the underlying correlation model, we also describe how P-domination can be checked by means of a set of *P-domination rules*, whose form depends on both the ranking semantics and the probabilistic model.

4.1. Loose Integration

In the loose integration scenario, a *probability engine* (also called *rule engine* in [Soliman et al. 2007]) is in charge of computing the probability of arbitrary tuple events, and all the details of the adopted probabilistic model $\mathcal{C}$ are therefore hidden to the query processor. As in [Soliman et al. 2007] we assume that the interface of the probability engine just includes a method that, given a combination of tuple events E, returns its probability $\Pr_{\mathcal{C}}(E)$.

4.2. Tight Integration

When a skyline algorithm is aware of the underlying correlation model, we can exploit the results in Section 3 so as to check P-domination by means of a set of *rules*, which avoid enumerating the linear extensions of $\succ$. Although the specific form of such rules depends on the specific combination of ranking semantics and probabilistic model (as it will be shown in Section 5), their applicability stays invariant, and can be concisely described as follows:

— **Rule1:** This is a rule that applies to the case $u \succ v$ and does not require to solve the INDARRANGE problem at all. If this rule succeeds then we know that $u \succ_p v$, whereas the test is inconclusive if the rule fails. In other terms (success of) Rule1 is sufficient but not necessary for P-domination to occur. The basic idea of Rule1 is to determine an easy-to-compute lower bound of the difference $\mathsf{Ps}^{\geq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\geq}_{\Psi,\mathcal{C}}(v)$, and then check if this bound is ≥ 0.
— **Rule2:** This is the necessary and sufficient P-domination rule for the case $u \succ v$. In general, this rule relies on the solution of the INDARRANGE problem, thus on the subsets of tuples $\text{IND\_1}(u, v)$ and $\text{IND\_0}(u, v)$ of $\text{IND}(u, v)$ that are arranged by the INDARRANGE solution before u and after v, respectively.
— **Rule3:** This rule is the only one for the case $u \not\succ v$. Notice that, because of Theorem 3.14 in Section 3, Rule3 has always the form $\mathsf{Ps}^{-}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{+}_{\Psi,\mathcal{C}}(v) \geq 0$. For this reason, in Section 5 we will show show how bounds on probabilistic scores can be computed for each considered combination of ranking semantics and probabilistic model.

4.3. Skyline Algorithms

From a performance point of view, one is interested not only in the time required to perform a single P-domination test, but also in how many of such tests are actually required to compute the skyline. It is quite intuitive that, by focusing only on a specific probabilistic model *and* ranking semantics, one can design an ad-hoc skyline algorithm, in which particular optimization techniques are applied (see [Bartolini et al. 2013] for an example in which both sorting and indexing are considered). However, since in this paper our aim is to compare the different combinations of ranking se-

mantics and probabilistic models (and integration scenarios as well) on a fair ground, a different and more neutral approach is needed. To this end, we consider a simple general-purpose approach for computing the skyline. In particular, since the precomputation of bounds is the key to speeding up evaluation of P-domination when $u \not\succ v$, both algorithms (one per integration scenario) present two different phases:

(1) In the first phase, bounds are precomputed. In the tight integration scenario, the cheap Rule1 is also applied so as to early discard some tuples and avoid computing bounds for them;
(2) in the second phase, the actual skyline is computed by applying all the necessary P-domination tests.

This approach is then specialized as follows, depending on the specific scenario at hand.

Loose integration. In the loose integration scenario, the skyline algorithm is unaware of the underlying correlation model, thus there is no analogue of Rule1. This implies that the INDARRANGE problem has to be solved for each considered pair of tuples u and v such that $u \succ v$.

In the first phase of the algorithm we request to the probability engine the probabilities of all the tuple events that are needed to compute lower and upper bounds on probabilistic scores for *all* the tuples in R^p. Then, since the most difficult case to be dealt with is the one when $u \succ v$, in the second phase we exploit a 2-scans approach, which first tests P-domination in the case $u \not\succ v$ using the precomputed bounds, and then solves the INDARRANGE problem for all the currently P-undominated tuples. In Algorithm 1, which depicts this strategy, we denote as INDARRANGE(u, v) the value of the INDARRANGE solution (i.e., the minimal difference of the probabilistic scores of u and v).

Algorithm 1 Skyline algorithm for the loose integration scenario

Input: probabilistic relation R^p, PRS ψ
Output: SKY(R^p)

SKY$(R^p) \leftarrow R^p$
for all tuples $u \in R^p$ **do** ▷ 1st phase
    invoke the probability engine to obtain all the probabilities needed to compute $\mathsf{Ps}^-_{\psi,\mathcal{C}}(u)$ and $\mathsf{Ps}^+_{\psi,\mathcal{C}}(u)$
for all tuples $u \in$ SKY(R^p) **do** ▷ 2nd phase – 1st scan
    for all tuples $v \in$ SKY(R^p), $v \neq u$, $v \sim u$ **do**
        if $\mathsf{Ps}^-_{\Psi,\mathcal{C}}(u) \geq \mathsf{Ps}^+_{\Psi,\mathcal{C}}(v)$ **then** SKY$(R^p) \leftarrow$ SKY$(R^p) \setminus \{v\}$
        else if $\mathsf{Ps}^-_{\Psi,\mathcal{C}}(v) \geq \mathsf{Ps}^+_{\Psi,\mathcal{C}}(u)$ **then** SKY$(R^p) \leftarrow$ SKY$(R^p) \setminus \{u\}$, **break**
for all tuples $u \in$ SKY(R^p) **do** ▷ 2nd phase – 2nd scan
    for all tuples $v \in$ SKY(R^p), $v \not\sim u$ **do**
        if $u \succ v$ and INDARRANGE$(u, v) \geq 0$ (i.e., $u \succ_p v$) **then**
            SKY$(R^p) \leftarrow$ SKY$(R^p) \setminus \{v\}$
        else if $v \succ u$ and INDARRANGE$(v, u) \geq 0$ (i.e., $v \succ_p u$) **then**
            SKY$(R^p) \leftarrow$ SKY$(R^p) \setminus \{u\}$, **break**

Tight integration. When the algorithm is aware of the underlying correlation model, the first phase can exploit the cheap Rule1 to avoid computing bounds for those tuples that are discovered to be P-dominated. As to the second phase, we still use a 2-scans approach on the tuples that are not P-dominated using Rule1 (Algorithm 2).

However, unlike Algorithm 1, now we can exploit "correlation model"-specific rules, both for the case $u \not\succ v$ (Rule3) and $u \succ v$ (Rule2).

Algorithm 2 Skyline algorithm for the tight integration scenario

Input: probabilistic relation R^p, PRS ψ, correlation model $\mathcal{C}$
Output: $\text{SKY}(R^p)$

1: $\text{SKY}(R^p) \leftarrow R^p$
2: **for all** tuples $u \in \text{SKY}(R^p)$ **do** ▷ 1st phase
3: **for all** tuples $v \in \text{SKY}(R^p)$, $v \not\sim u$ **do**
4: **if** $u \succ v$ and $u \succ_p v$ [Rule1] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{v\}$
5: **else if** $v \succ u$ and $v \succ_p u$ [Rule1] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{u\}$, **break**
6: **if** $u \in \text{SKY}(R^p)$ **then** compute $\text{Ps}^-_{\psi,\mathcal{C}}(u)$ and $\text{Ps}^+_{\psi,\mathcal{C}}(u)$
7: **for all** tuples $u \in \text{SKY}(R^p)$ **do** ▷ 2nd phase – 1st scan
8: **for all** tuples $v \in \text{SKY}(R^p)$, $v \neq u$, $v \sim u$ **do**
9: **if** $u \succ_p v$ [Rule3] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{v\}$
10: **else if** $v \succ_p u$ [Rule3] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{u\}$, **break**
11: **for all** tuples $u \in \text{SKY}(R^p)$ **do** ▷ 2nd phase – 2nd scan
12: **for all** tuples $v \in \text{SKY}(R^p)$, $v \not\sim u$ **do**
13: **if** $u \succ v$ and $u \succ_p v$ [Rule2] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{v\}$
14: **else if** $v \succ u$ and $v \succ_p u$ [Rule2] **then** $\text{SKY}(R^p) \leftarrow \text{SKY}(R^p) \setminus \{u\}$, **break**

5. RULES FOR SPECIFIC SEMANTICS AND CORRELATION MODELS

In this section we analyze known probabilistic ranking semantics, for each of which we provide two types of results. First, we study how P-domination can be checked for the given PRS *regardless* of the specific probabilistic model. Besides providing us with general information on the complexity of INDARRANGE^P for a certain PRS, this is the scenario to consider when the query processor and the engine computing probabilities are only loosely integrated (Algorithm 1). Second, we derive P-domination rules for specific probabilistic models, in particular for the independent and the x-relation cases, to be used in Algorithm 2.

In order to avoid cluttering the text with too many equations, detailed derivations of P-domination rules are collected in the Electronic Appendix that is accessible in the ACM Digital Library.

5.1. Expected Rank

The *expected rank* (ER) of a tuple u is defined as [Cormode et al. 2009]:

$$ER^{\succcurlyeq}_{\mathcal{C}}(u) = \sum_{W \in \mathcal{W}} \Pr_{\mathcal{C}}(W) \times \text{rank}^{\succcurlyeq}_W(u) \tag{6}$$

where $\text{rank}^{\succcurlyeq}_W(u)$ is the rank of u in the possible world W, given the linear order $\succcurlyeq$. This can be conveniently rewritten as:

$$ER^{\succcurlyeq}_{\mathcal{C}}(u) = \sum_{\substack{W \in \mathcal{W} \\ u \in W}} \Pr_{\mathcal{C}}(W) \times \text{rank}^{\succcurlyeq}_W(u) + \sum_{\substack{W \in \mathcal{W} \\ u \notin W}} \Pr_{\mathcal{C}}(W) \times |W| \tag{7}$$

where the first term is the expected rank of u in the possible worlds in which u appears, whereas the second term accounts for the contribution of those worlds that do not contain u.

We first derive the following basic result, which holds for arbitrary probabilistic models:

LEMMA 5.1. *For any set of correlation constraints $\mathcal{C}$, the expected rank of a tuple u can be expressed as a sum of binary probabilistic events, i.e.:*

$$ER^{\succcurlyeq}_{\mathcal{C}}(u) = \sum_{t \succcurlyeq u} \Pr_{\mathcal{C}}(u \wedge t) + \sum_{t} \Pr_{\mathcal{C}}(\neg u \wedge t) \tag{8}$$

PROOF. We rewrite Equation 6 as follows, by exploiting the Iverson bracket notation:[7]

$$\begin{aligned} ER^{\succcurlyeq}_{\mathcal{C}}(u) &= \sum_{\substack{W \in \mathcal{W} \\ u \in W}} \Pr_{\mathcal{C}}(W) \times |\{t \in W, t \succcurlyeq u\}| + \sum_{\substack{W \in \mathcal{W} \\ u \notin W}} \Pr_{\mathcal{C}}(W) \times |W| \\ &= \sum_{\substack{W \in \mathcal{W} \\ u \in W}} \Pr_{\mathcal{C}}(W) \times \sum_{t} [t \in W \wedge t \succcurlyeq u] + \sum_{\substack{W \in \mathcal{W} \\ u \notin W}} \Pr_{\mathcal{C}}(W) \times \sum_{t} [t \in W] \\ &= \sum_{W \in \mathcal{W}} \Pr_{\mathcal{C}}(W) \times \sum_{t} [u \in W \wedge t \in W \wedge t \succcurlyeq u] + \sum_{W \in \mathcal{W}} \Pr_{\mathcal{C}}(W) \times \sum_{t} [u \notin W \wedge t \in W] \\ &= \sum_{t \succcurlyeq u} \Pr_{\mathcal{C}}(u \wedge t) + \sum_{t} \Pr_{\mathcal{C}}(\neg u \wedge t) \end{aligned}$$

□

In order to use the expected rank as a PRS we first need to complement probabilistic scores, since $ER^{\succcurlyeq}_{\mathcal{C}}(u)$ is to be minimized. To this end, we maintain that $\mathsf{Ps}^{\succcurlyeq}_{ER,\mathcal{C}}(u) \equiv N - ER^{\succcurlyeq}_{\mathcal{C}}(u)$, where N is the number of tuples in R^p. Clearly, this choice guarantees that $\mathsf{Ps}^{\succcurlyeq}_{ER,\mathcal{C}}(u) > 0$.

LEMMA 5.2. *The expected rank* PRS *is set-monotone.*

PROOF. Clearly, ER is set-dependent, since from Equation 7 it is apparent that the specific ordering of the tuples preceding u has no influence on $\mathsf{Ps}^{\succcurlyeq}_{ER,\mathcal{C}}(u)$, and the probability of a possible world, $\Pr_{\mathcal{C}}(W)$, is clearly independent of tuples' ordering. Regarding rank-monotonicity, we observe that, when $\mathrm{Up}^{\succcurlyeq}(u) \subset \mathrm{Up}^{\succcurlyeq'}(u)$, it is $\mathrm{rank}^{\succcurlyeq}_{W}(u) \leq \mathrm{rank}^{\succcurlyeq'}_{W}(u)$ in any possible world W containing u. The result then follows since the second term of Equation 7 is independent of the particular linear order. □

Due to Lemma 5.1, the function to be minimized by the INDARRANGE problem can be written as (we omit the intermediate algebraic steps):

$$\begin{aligned} &\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{ER,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{ER,\mathcal{C}}(v) = \\ &\qquad \sum_{t \succ v} \Pr_{\mathcal{C}}(v \wedge t) - \sum_{t \succ u} \Pr_{\mathcal{C}}(u \wedge t) - \sum_{t \in \textsc{IndBtr}(u,v)} \Pr_{\mathcal{C}}(u \wedge t) \\ &\quad + \sum_{t} \left(\Pr_{\mathcal{C}}(\neg v \wedge t) - \Pr_{\mathcal{C}}(\neg u \wedge t)\right) + \sum_{t_i \in \textsc{Ind}(u,v)} y_i \left(\Pr_{\mathcal{C}}(v \wedge t_i) - \Pr_{\mathcal{C}}(u \wedge t_i)\right) \end{aligned} \tag{9}$$

[7]For any predicate P, $[P] = 1$ if P is true, otherwise $[P] = 0$.

in which we remind from Theorem 3.9 that each y_i is a binary variable corresponding to a tuple $t_i \in \text{IND}(u, v)$, with $y_i = 1$ denoting that tuple t_i is arranged before u, and $y_i = 0$ that t is arranged after v.

We are now in the position to characterize the complexity of the INDARRANGE^P problem for the ER semantics.

THEOREM 5.3. *Let R^p be a probabilistic relation and $\mathcal{C}$ any set of correlation constraints. For the expected rank semantics, the* INDARRANGE^P *problem can be solved in $\mathcal{O}(poly(M_{\max}))$ time, where $M_{\max}$ is the maximum size of* IND(u, v) *over all pairs of tuples u and v in R^p such that $u \succ v$.*

PROOF. Let $\Delta_i = \Pr_{\mathcal{C}}(u \wedge t_i) - \Pr_{\mathcal{C}}(v \wedge t_i)$, so that the problem of minimizing $\sum_{t_i \in \text{IND}(u,v)} y_i \left(\Pr_{\mathcal{C}}(v \wedge t_i) - \Pr_{\mathcal{C}}(u \wedge t_i)\right)$ in Equation 9, thus the difference of probabilistic scores, is equivalent to maximize $\sum_{t_i \in \text{IND}(u,v)} y_i \Delta_i$. This can be done in polynomial time using a max-flow/min-cut algorithm, after reformulating the problem as a *project selection problem*. It is given a set of projects P, where each project i has an either positive or negative revenue r_i. A *dependency* (j, i) states that project j has project i as a prerequisite (cannot do project j without also doing project i). The project selection problem is to select a subset $P^1 \subseteq P$ of projects so as to maximize the total revenue $\sum_{i \in P^1} r_i$ by respecting all the dependencies. Since a solution to the project selection problem corresponds to a minimum-cut in a suitably defined graph, the problem can be solved in polynomial time (see, e.g., [Kleinberg and Tardos 2006] for details). Converting the problem of maximizing $\sum_{t_i \in \text{IND}(u,v)} y_i \Delta_i$ to the project selection problem is immediate: Each tuple $t_i \in \text{IND}(u, v)$ corresponds to a project $i \in P$ with revenue $r_i = \Delta_i$, and there is a dependency (j, i) iff $t_i \succ t_j$. The latter guarantees that if t_j is ordered before u ($y_j = 1$, i.e., project j is selected), then so it is the case for t_i. □

5.1.1. ER: P-domination rules.

Independent model.

For the simplest case of independent tuples, i.e., $\mathcal{C} = \emptyset$, the probabilistic score of a tuple can be written as:

$$\mathsf{Ps}^{\geqslant}_{ER,\emptyset}(u) = N - p(u) \times \sum_{t \geqslant u} p(t) - (1 - p(u)) \times \sum_{t \neq u} p(t) \tag{10}$$

from which the following bounds are immediately obtained:

$$\mathsf{Ps}^{+}_{ER,\emptyset}(u) = N - p(u) \times \sum_{t \succ u} p(t) - (1 - p(u)) \times \sum_{t \neq u} p(t)$$

$$\mathsf{Ps}^{-}_{ER,\emptyset}(u) = N - p(u) \times \sum_{\substack{u \not\succ t \\ t \neq u}} p(t) - (1 - p(u)) \times \sum_{t \neq u} p(t)$$

Concerning P-domination rules for the $u \succ v$ case, we have the following:

RULES $(ER, \emptyset)$. *Given the combination of ER ranking semantics and independent tuples, assume $u \succ v$. Then, it is $u \succ_p v$ if Rule1$_{ER,\emptyset}$ holds, or if and only if Rule2$_{ER,\emptyset}$ holds:*

$$p(u) \geq p(v) \tag{Rule1$_{ER,\emptyset}$}$$

$$p(u) \geq p(v) \vee \mathsf{Ps}^{-}_{ER,\emptyset}(u) + p(u) \times \sum_{t \in \text{IND}(u,v)} p(t) \geq \mathsf{Ps}^{+}_{ER,\emptyset}(v) \tag{Rule2$_{ER,\emptyset}$}$$

As proved in Appendix A, for the $(ER, \emptyset)$ combination the IndArrange^P problem does not need to be solved at all, i.e., its complexity is $\mathcal{O}(1)$. This reflects into Rule2$_{ER,\emptyset}$, in which the summation extends over *all* tuples in Ind(u, v), rather than on a specific subset Ind_1(u, v). Notice that, since the 1st disjunct in Rule2$_{ER,\emptyset}$ coincides with Rule1$_{ER,\emptyset}$, when Rule1$_{ER,\emptyset}$ fails only the 2nd disjunct needs to be checked.

X-relation model.

For the case of x-relations, the probabilistic score of a tuple is derived as:

$$\mathsf{Ps}^{\geqslant}_{ER,\mathcal{C}_X}(u) = N - p(u) \times \sum_{\substack{t \notin G(u) \\ t \geqslant u}} p(t) - (1 - p(u)) \times \sum_{t \notin G(u)} p(t) - \sum_{\substack{t \in G(u) \\ t \neq u}} p(t) \tag{11}$$

where $G(u)$ denotes the group of tuple u. The bound are consequently derived to be:

$$\mathsf{Ps}^{+}_{ER,\mathcal{C}_X}(u) = N - p(u) \times \sum_{\substack{t \notin G(u) \\ t \succ u}} p(t) - (1 - p(u)) \times \sum_{t \notin G(u)} p(t) - \sum_{\substack{t \in G(u) \\ t \neq u}} p(t)$$

$$\mathsf{Ps}^{-}_{ER,\mathcal{C}_X}(u) = N - p(u) \times \sum_{\substack{t \notin G(u) \\ u \not\succ t}} p(t) - (1 - p(u)) \times \sum_{t \notin G(u)} p(t) - \sum_{\substack{t \in G(u) \\ t \neq u}} p(t)$$

From Equation 9, the difference of probabilistic scores is now written as:

$$\begin{aligned}
\mathsf{Ps}^{\geqslant_{\mathbf{Y}}}_{ER,\mathcal{C}_X}(u) - \mathsf{Ps}^{\geqslant_{\mathbf{Y}}}_{ER,\mathcal{C}_X}(v) = {} & p(v) \times \left(\sum_{\substack{t \succ v \\ t \notin G(v)}} p(t) + \sum_{\substack{t_i \in \textsc{Ind}(u,v) \\ t_i \notin G(v)}} y_i p(t_i) \right) \\
& - p(u) \times \left(\sum_{\substack{t \in \textsc{Btr}(u,v) \\ t \notin G(u)}} p(t) + \sum_{\substack{t \in \textsc{IndBtr}(u,v) \\ t \notin G(u)}} p(t) + \sum_{\substack{t_i \in \textsc{Ind}(u,v) \\ t_i \notin G(u)}} y_i p(t_i) \right) \\
& + (1 - p(v)) \times \sum_{t \notin G(v)} p(t) + \sum_{\substack{t \in G(v) \\ t \neq v}} p(t) - (1 - p(u)) \times \sum_{t \notin G(u)} p(t) - \sum_{\substack{t \in G(u) \\ t \neq u}} p(t)
\end{aligned} \tag{12}$$

We first analyze the case where u and v belong to different groups. For deriving a (u, v)-adversarial order when $u \succ v$, we first partition all tuples in Ind(u, v) into three sets: Those in $G(u)$, those in $G(v)$, and those in other groups. The contribution of such tuples to the difference in Equation 12 can then be written as:

$$p(v) \times \sum_{\substack{t_i \in \textsc{Ind}(u,v) \\ t_i \in G(u)}} y_i p(t_i) - p(u) \times \sum_{\substack{t_i \in \textsc{Ind}(u,v) \\ t_i \in G(v)}} y_i p(t_i) + (p(v) - p(u)) \times \sum_{\substack{t_i \in \textsc{Ind}(u,v) \\ t_i \notin G(u) \\ t_i \notin G(v)}} y_i p(t_i) \tag{13}$$

from which we derive the following result:

Lemma 5.4. *Given a probabilistic relation R^p, let ER be the ranking semantics and assume the x-relation model for tuple correlation. Then, the* IndArrange^P *problem can be solved in $\mathcal{O}(2^g)$ time, where g is the maximum cardinality of a group in R^p.*

Proof. First consider the case $p(v) > p(u)$ and assume for the moment that no dominance relationship constraint is present. In this case the value of (13) will be

minimized by setting $y_i = 1$ for all and only those tuples $t_i \in \text{IND}(u,v) \cap G(v)$ (i.e., by placing only such tuples before u). When dominance relationship constraints are present, observe that, for any fixed arrangement of tuples in the group of v, the value of (13) is minimized by setting, for any other tuple $t_j \in \text{IND}(u,v)$, $y_j = 1$ iff there exists $t_i \in \text{IND}(u,v) \cap G(v)$ such that $y_i = 1$ and $t_j \succ t_i$. It follows that solving INDARRANGE^P reduces to determining how to arrange the tuples in $\text{IND}(u,v) \cap G(v)$.

In a similar way, we can deal with the case $p(u) \geq p(v)$ by just considering how to arrange the tuples in $\text{IND}(u,v) \cap G(u)$, which proves that in both cases the number of arrangements to test is $\mathcal{O}(2^g)$. □

From the above lemma it follows that if the cardinality of a group is limited by a constant, the INDARRANGE^P problem can be solved in $\mathcal{O}(1)$ time. In other cases the approach based on testing $\mathcal{O}(2^g)$ arrangements should be contrasted to the one implied by Theorem 5.3, which guarantees polynomial complexity for any correlation model.[8]

Concerning P-domination rules for the $u \succ v$ case, we have the following:

RULES $(ER, \mathcal{C}_X)$. *Given the combination of ER ranking semantics and x-relation correlation model, assume $G(u) \neq G(v)$ and $u \succ v$. Then, it is $u \succ_p v$ if Rule1$_{ER,\mathcal{C}_X}$ holds, or if and only if Rule2$_{ER,\mathcal{C}_X}$ holds:*

$$p(u) \geq p(v) \wedge p(u) \geq \sum_{\substack{t \in G(u) \\ t \neq u}} p(t) \qquad \text{(Rule1}_{ER,\mathcal{C}_X}\text{)}$$

$$\mathsf{Ps}^-_{ER,\mathcal{C}_X}(u) + p(u) \sum_{\substack{t \in \text{IND\_0}(u,v) \\ t \notin G(u)}} p(t) \geq \mathsf{Ps}^+_{ER,\mathcal{C}_X}(v) - p(v) \times \sum_{\substack{t \in \text{IND\_1}(u,v) \\ t \notin G(v)}} p(t) \qquad \text{(Rule2}_{ER,\mathcal{C}_X}\text{)}$$

Comparing Rule1$_{ER,\mathcal{C}_X}$ to Rule1$_{ER,\emptyset}$, it is evident that the x-relation case is harder than the independent one, since an additional condition has to be verified for the rule to succeed.

If u and v belong to the same group, $G(u) = G(v)$, from Equation 12 it is clear that, among the tuples in $\text{IND}(u,v)$, only those in other groups can influence the result. Therefore the term to minimize is just:

$$(p(v) - p(u)) \times \sum_{\substack{t_i \in \text{IND}(u,v) \\ t_i \notin G(u)}} y_i p(t_i) \qquad (14)$$

It follows that if $p(u) \geq p(v)$, the minimum is obtained by setting $y_i = 1, \forall i$, whereas $y_i = 0, \forall i$, is optimal when $p(u) < p(v)$.

5.2. Ranking Semantics Based on Top-1 Probability

We now consider three well-known ranking semantics that, for the purpose of computing the skyline, can be dealt with in a unified way.

The *U-Topk* semantics [Soliman et al. 2007] returns the most likely top-k result set, by considering the probabilities of the possible worlds of R^p, whereas *U-kRanks* [Soliman et al. 2007] returns, for each $i = 0, \ldots, k-1$, the tuple with the highest probability of being at rank i. Finally, the *Global-Topk* semantics [Zhang and Chomicki 2008] computes for each tuple u the probability that u is among the top-k tuples in the possible worlds of R^p, and returns the k tuples with the highest probabilities.

[8]Clearly, when $g = \Omega(N)$, computing the possible arrangements of tuples in a single group would lead to exponential-time complexity.

Although these semantics do not define a ranking in the strict sense, i.e., the top-k tuples need not to be a proper subset of the top-$(k+1)$ ones (since the actual ranking depends on the value of k), for the purpose of computing the skyline, which is only concerned with the top-ranked tuple on each linear order, one can safely consider the case $k=1$, i.e., ranking tuples based on their *top-1 probability*. Since in this case the three semantics behave exactly the same [Yan and Ng 2011], we collectively refer to them as *Top-1 semantics*, $T1$ for short.

In the general case the top-1 probability of a tuple can be expressed as:

$$\mathsf{Ps}^{\succ}_{T1,\mathcal{C}}(u) = \Pr_{\mathcal{C}}(\{u \text{ is the top-1 tuple under } \succ\}) = \Pr_{\mathcal{C}}\left(u \wedge \bigwedge_{t \succ u} \neg t\right) \tag{15}$$

because, for u to be the top-1 tuple in a possible world W, no tuples preceding u in $\succ$ should also appear in W.

It is easy to prove that this PRS is set-monotone, because increasing the number of tuples preceding u can never increase the probability that u is the top-1 tuple.

For arbitrary probabilistic models, there is no analogue of Lemma 5.1 for the $T1$ semantics. Intuitively, this depends on the impossibility of factorizing the probability in Equation 15 without knowledge of the specific $\mathcal{C}$ constraints. Therefore, by referring to the structure of a (u,v)-adversarial order in Figure 3, in the general case one should minimize:

$$\begin{aligned}
\mathsf{Ps}^{\succ_{\mathbf{Y}}}_{T1,\mathcal{C}}(u) - \mathsf{Ps}^{\succ_{\mathbf{Y}}}_{T1,\mathcal{C}}(v) &= \Pr_{\mathcal{C}}\left(u \wedge \bigwedge_{t \succ_{\mathbf{Y}} u} \neg t\right) - \Pr_{\mathcal{C}}\left(v \wedge \bigwedge_{t \succ_{\mathbf{Y}} v} \neg t\right) \\
&= \Pr_{\mathcal{C}}\left(u \wedge \bigwedge_{t \in \textsc{Btr}(u,v)} \neg t \wedge \bigwedge_{t \in \textsc{IndBtr}(u,v)} \neg t \wedge \bigwedge_{\substack{t_i \in \textsc{Ind}(u,v) \\ y_i = 1}} \neg t_i\right) \\
&\quad - \Pr_{\mathcal{C}}\left(v \wedge \bigwedge_{t \in \textsc{Btr}(u,v)} \neg t \wedge \bigwedge_{t \in \textsc{IndBtr}(u,v)} \neg t \wedge \bigwedge_{\substack{t_i \in \textsc{Ind}(u,v) \\ y_i = 1}} \neg t_i \wedge \neg u \wedge \bigwedge_{t \in \textsc{WrsBtr}(u,v)} \neg t\right)
\end{aligned} \tag{16}$$

From the above, we derive the following negative result:

THEOREM 5.5. *For the Top-1 ranking semantics: (1) The* INDARRANGE^P *problem is NP-hard, and (2) determining if a tuple u P-dominates another tuple v is co-NP-complete if any probabilistic score can be computed in polynomial time.*

PROOF. See Appendix A. □

5.2.1. Top-1: P-domination rules.

Independent model.

For the case of independent tuples, the top-1 probability of a tuple is expressed as:

$$\mathsf{Ps}^{\succ}_{T1,\emptyset}(u) = p(u) \prod_{t \succ u} (1 - p(t)) \tag{17}$$

Bounds are then computed as:

$$\mathsf{Ps}^{+}_{T1,\emptyset}(u) = p(u)\prod_{t\succ u}(1-p(t))$$
$$\mathsf{Ps}^{-}_{T1,\emptyset}(u) = p(u)\prod_{\substack{u\not\succ t\\ t\neq u}}(1-p(t)) = \mathsf{Ps}^{+}_{T1,\emptyset}(u)\times\prod_{\substack{t\sim u\\ t\neq u}}(1-p(t))$$

whereas Equation 16 becomes:

$$\mathsf{Ps}^{\succcurlyeq\mathbf{Y}}_{T1,\emptyset}(u) - \mathsf{Ps}^{\succcurlyeq\mathbf{Y}}_{T1,\emptyset}(v) = p(u)\prod_{\substack{t\in\textsc{Btr}(u,v)\\ \cup\textsc{IndBtr}(u,v)}}(1-p(t))\prod_{t_i\in\textsc{Ind}(u,v)}(1-y_i p(t_i))$$
$$-\,p(v)(1-p(u))\prod_{t\in\textsc{WrsBtr}(u,v)}(1-p(t))\prod_{\substack{t\in\textsc{Btr}(u,v)\\ \cup\textsc{IndBtr}(u,v)}}(1-p(t))\prod_{t_i\in\textsc{Ind}(u,v)}(1-y_i p(t_i)) \quad (18)$$

From the above equation, after some algebraic manipulations detailed in Appendix A, we derive the following:

RULES $(T1,\emptyset)$. *Given the combination of Top-1 ranking semantics and independent tuples, assume $u\succ v$. Then, it is $u\succ_p v$ if Rule1$_{T1,\emptyset}$ holds, or if and only if Rule2$_{T1,\emptyset}$ holds:*

$$p(u)\geq p(v)(1-p(u)) \qquad (\text{Rule1}_{T1,\emptyset})$$
$$\mathsf{Ps}^{+}_{T1,\emptyset}(v) = 0 \vee p(u)\geq p(v)(1-p(u))\prod_{t\in\textsc{WrsBtr}(u,v)}(1-p(t)) \qquad (\text{Rule2}_{T1,\emptyset})$$

Notice that, as with the ER semantics, the INDARRANGE^P problem needs not to be solved at all.

It can be observed that Rule1$_{T1,\emptyset}$ is less restrictive than Rule1$_{ER,\emptyset}$, because of the additional $(1-p(u))$ factor in the right-hand side. Therefore, with independent tuples the chance of success of Rule1 is higher for the Top-1 PRS than for ER.

X-relation model.

For x-relations, the probabilistic score of a tuple can be written as:

$$\mathsf{Ps}^{\succcurlyeq}_{T1,\mathcal{C}_X}(u) = p(u)\prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)}}\left(1-\sum_{\substack{t\in G\\ t\succcurlyeq u}}p(t)\right) \quad (19)$$

where $\mathcal{C}_X$ is the set of all groups in R^p. The bounds are:

$$\mathsf{Ps}^{+}_{T1,\mathcal{C}_X}(u) = p(u)\prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)}}\left(1-\sum_{\substack{t\in G\\ t\succ u}}p(t)\right)$$
$$\mathsf{Ps}^{-}_{T1,\mathcal{C}_X}(u) = p(u)\prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)}}\left(1-\sum_{\substack{t\in G\\ u\not\succ t}}p(t)\right)$$

and Equation 16 is written as:

$$\mathsf{Ps}^{\succcurlyeq_Y}_{T1,\mathcal{C}_X}(u) - \mathsf{Ps}^{\succcurlyeq_Y}_{T1,\mathcal{C}_X}(v) =$$

$$p(u)\left(1 - \sum_{\substack{t\in G(v)\\ t\in \textsc{Btr}(u,v)\\ \cup \textsc{IndBtr}(u,v)}} p(t) - \sum_{\substack{t_i\in G(v)\\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)\right) \times \prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)\\ G\neq G(v)}} \left(1 - \sum_{\substack{t\in G\\ t\in \textsc{Btr}(u,v)\\ \cup \textsc{IndBtr}(u,v)}} p(t) - \sum_{\substack{t_i\in G\\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)\right)$$

$$- p(v)\left(1 - \sum_{\substack{t\in G(u)\\ t\in \textsc{Btr}(u,v)\\ \cup \textsc{IndBtr}(u,v)}} p(t) - \sum_{\substack{t_i\in G(u)\\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i) - \sum_{\substack{t\in G(u)\\ t\in \textsc{WrsBtr}(u,v)}} p(t) - p(u)\right)$$

$$\times \prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)\\ G\neq G(v)}} \left(1 - \sum_{\substack{t\in G\\ t\in \textsc{Btr}(u,v)\\ \cup \textsc{IndBtr}(u,v)}} p(t) - \sum_{\substack{t_i\in G\\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i) - \sum_{\substack{t\in G\\ t\in \textsc{WrsBtr}(u,v)}} p(t)\right) \quad (20)$$

When u and v belong to different groups, we derive a result similar to the one obtained for the ER semantics:

LEMMA 5.6. *Given a probabilistic relation R^p, let Top-1 be the ranking semantics and assume that the x-relation correlation model is in use. Then, the* IndArrange^P *problem can be solved in $\mathcal{O}(2^g)$ time, where g is the maximum cardinality of a group in R^p.*

PROOF. See Appendix A. □

Overall, the following rules apply to the case $G(u) \neq G(v)$:

RULES $(T1, \mathcal{C}_X)$. *Given the combination of Top-1 ranking semantics and x-relation correlation model, assume $G(u) \neq G(v)$ and $u \succ v$. Then, it is $u \succ_p v$ if Rule1$_{T1,\mathcal{C}_X}$ holds, or if and only if Rule2$_{T1,\mathcal{C}_X}$ holds:*

$$p(u)\left(1 - \sum_{\substack{t\in G(v)\\ t\neq v}} p(t)\right) \geq p(v)\,(1 - p(u)) \qquad (\text{Rule1}_{T1,\mathcal{C}_X})$$

$$p(u) \prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)}} \left(1 - \sum_{\substack{t\in G\\ u\not\succ t}} p(t) + \sum_{\substack{t\in G\\ t\in \textsc{Ind\_0}(u,v)}} p(t)\right)$$

$$\geq p(v) \prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(v)}} \left(1 - \sum_{\substack{t\in G\\ t\succ v}} p(t) - \sum_{\substack{t\in G\\ t\in \textsc{Ind\_1}(u,v)}} p(t)\right) \qquad (\text{Rule2}_{T1,\mathcal{C}_X})$$

As with the expected rank PRS, also for Top-1 the x-relation case is harder than the independent one, since Rule1$_{T1,\mathcal{C}_X}$ is more restrictive than Rule1$_{T1,\emptyset}$. On the other hand, differently from Rule2$_{ER,\mathcal{C}_X}$, now we cannot exploit precomputed bounds when checking Rule2$_{T1,\mathcal{C}_X}$: This is because the latter contains a product of sums, thus the contribution of tuples in IND_1(u,v) cannot be factorized.

We conclude the analysis of the Top-1 PRS by considering the case of tuples belonging to the same group. By using arguments that follow those applied for the ER semantics, it is derived that the optimal solution of INDARRANGE is to have $y_i = 0, \forall i$. It follows that Rule1 reduces to Rule1$_{T1,\emptyset}$, whereas Rule2 is obtained by setting IND_1$(u,v) = \emptyset$ in Rule2$_{T1,\mathcal{C}_X}$.

5.3. Parameterized Probabilistic Ranking Functions (PRF)

We consider now the parameterized probabilistic ranking functions (PRFs) introduced in [Li et al. 2011]. In particular, we study here the $PRF^e(\alpha)$ function which, according to [Li et al. 2011] "is the most suitable for ranking in probabilistic databases". For such PRS the probabilistic score of a tuple u is expressed as:

$$\mathsf{Ps}^{\succcurlyeq}_{PRF^e(\alpha),\mathcal{C}}(u) = \sum_{i\geq 0} \alpha^i \Pr(\mathbf{rank}^{\succ}(u) = i) \tag{21}$$

where $\alpha < 1$ is a positive real constant. Notice that $\alpha = 1$ would lead to rank tuples based solely on their probabilities [Li et al. 2011], thus P-domination would reduce to check the inequality $p(u) > p(v)$ (independently of attribute values).

We first prove that $PRF^e(\alpha)$ is set-monotone. Clearly, $PRF^e(\alpha)$ is set-dependent, since the specific ordering of tuples preceding u has no influence on the probabilistic score of u. $PRF^e(\alpha)$ is also rank-monotone, since when $\mathrm{Up}^{\succ}(u) \subset \mathrm{Up}^{\succ'}(u)$, it is $\mathrm{rank}^{\succcurlyeq}_W(u) \leq \mathrm{rank}^{\succcurlyeq'}_W(u)$ in any possible world W containing u, thus $\mathsf{Ps}^{\succcurlyeq}_{PRF^e(\alpha),\mathcal{C}}(u) \geq \mathsf{Ps}^{\succcurlyeq'}_{PRF^e(\alpha),\mathcal{C}}(u)$, due to $\alpha < 1$.

The following Lemma 5.7 is the key to exploit, with appropriate modifications, all the results we derived for the Top-1 PRS for the analysis of $PRF^e(\alpha)$.

LEMMA 5.7. *Assuming* $0^0 = 1$, $PRF^e(0)$ *coincides with the Top-1 ranking semantics for any correlation model* $\mathcal{C}$.

PROOF. It is $\mathsf{Ps}^{\succcurlyeq}_{PRF^e(0),\mathcal{C}}(u) = \sum_{i\geq 0} 0^i \Pr(\mathbf{rank}^{\succ}(u) = i) = \Pr(\mathbf{rank}^{\succ}(u) = 0) = \mathsf{Ps}^{\succcurlyeq}_{T1,\mathcal{C}}(u)$. □

The above Lemma immediately leads us to conclude that even for the $PRF^e(\alpha)$ ranking semantics the INDARRANGE^P problem, and consequently P-domination, is NP-hard.

Notice that Lemma 5.7 uses the fact that the first term of the sum in Equation 21 is given weight 1, rather than α as in [Li et al. 2011]. Clearly, this slight difference has no consequences on the ranking of tuples, yet it allows the probabilistic score to be well-defined even when $\alpha = 0$.

5.3.1. $PRF^e(\alpha)$: P-domination rules.

Independent model.

To write $PRF^e(\alpha)$ under the tuple independence assumption, we exploit the concept of a *generating function* of a tuple u [Li et al. 2011], i.e., a polynomial function in the x variable, $\sum_{j\geq 0} c_j x^j$, for which the coefficient c_j equals the probability that u is at rank j. For independent tuples, the generating function of u is a product (over all tuples t_i

preceding u in the linear order $\succcurlyeq$) of factors $(1 - p(t_i) + p(t_i)x)$, because the rank of u is increased by 1 only if t_i is present. Then, following the definition of $PRF^e(\alpha)$, the probabilistic score of u is obtained as the generating function of u computed for $x = \alpha$:

$$\mathsf{Ps}^{\succcurlyeq}_{PRF^e(\alpha),\emptyset}(u) = p(u) \prod_{t \succcurlyeq u} (1 - p(t)(1 - \alpha)) \tag{22}$$

where it is evident that setting $\alpha = 0$ yields Equation 17, i.e., the probabilistic score under the Top-1 semantics.

The bounds are derived to be:

$$\begin{aligned}
\mathsf{Ps}^{+}_{PRF^e(\alpha),\emptyset}(u) &= p(u) \prod_{t \succ u} (1 - p(t)(1 - \alpha)) \\
\mathsf{Ps}^{-}_{PRF^e(\alpha),\emptyset}(u) &= p(u) \prod_{t \in \mathrm{Up}^+(u)} (1 - p(t)(1 - \alpha)) \\
&= \mathsf{Ps}^{+}_{PRF^e(\alpha),\emptyset}(u) \times \prod_{\substack{t \sim u \\ t \neq u}} (1 - p(t)(1 - \alpha))
\end{aligned}$$

and the INDARRANGE problem has to minimize:

$$\begin{aligned}
&\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{PRF^e(\alpha),\emptyset}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{PRF^e(\alpha),\emptyset}(v) = \\
&\quad p(u) \prod_{\substack{t \in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} (1 - p(t)(1 - \alpha)) \prod_{t_i \in \textsc{Ind}(u,v)} (1 - y_i p(t_i)(1 - \alpha)) \\
&\quad - p(v)(1 - p(u)(1 - \alpha)) \prod_{t \in \textsc{WrsBtr}(u,v)} (1 - p(t)(1 - \alpha)) \\
&\quad \times \prod_{\substack{t \in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} (1 - p(t)(1 - \alpha)) \prod_{t_i \in \textsc{Ind}(u,v)} (1 - y_i p(t_i)(1 - \alpha))
\end{aligned} \tag{23}$$

Overall the P-domination check amounts to evaluating the following

RULES $(PRF^e(\alpha), \emptyset)$. *Given the combination of $PRF^e(\alpha)$ ranking semantics and independent tuples, assume $u \succ v$. Then, it is $u \succ_p v$ if Rule1$_{PRF^e(\alpha),\emptyset}$ holds, or if and only if Rule2$_{PRF^e(\alpha),\emptyset}$ holds:*

$$p(u) \geq p(v)(1 - p(u)(1 - \alpha)) \tag{Rule1$_{PRF^e(\alpha),\emptyset}$}$$

$$\mathsf{Ps}^{+}_{PRF^e(\alpha),\emptyset}(v) = 0 \vee p(u) \geq p(v)(1 - p(u)(1 - \alpha)) \prod_{t \in \textsc{WrsBtr}(u,v)} (1 - p(t)(1 - \alpha)) \tag{Rule2$_{PRF^e(\alpha),\emptyset}$}$$

Comparing $PRF^e(\alpha)$ with the expected rank and Top-1 semantics, we note that Rule1$_{PRF^e(\alpha),\emptyset}$ is more restrictive than Rule1$_{T1,\emptyset}$, but less restrictive than Rule1$_{ER,\emptyset}$.

X-relation model.

With x-relations, the generating function of a tuple u is obtained as a product of contributions for each group, where the contribution of a group G other than $G(u)$ is $(1 - \sum_{\substack{t \in G \\ t \succcurlyeq u}} p(t) + \sum_{\substack{t \in G \\ t \succcurlyeq u}} p(t)x)$, because each group of mutually exclusive tuples can

contribute to increase the rank of u at most by 1. Then, $PRF^e(\alpha)$ is written as:

$$\mathsf{Ps}^{\succcurlyeq}_{PRF^e(\alpha),\mathcal{C}_X}(u) = p(u) \prod_{\substack{G\in\mathcal{C}_X \\ G\neq G(u)}} \left(1 - \sum_{\substack{t\in G \\ t\succ u}} p(t)(1-\alpha)\right) \tag{24}$$

Bounds are immediately computed as:

$$\mathsf{Ps}^{+}_{PRF^e(\alpha),\mathcal{C}_X}(u) = p(u) \prod_{\substack{G\in\mathcal{C}_X \\ G\neq G(u)}} \left(1 - \sum_{\substack{t\in G \\ t\succ u}} p(t)(1-\alpha)\right)$$

$$\mathsf{Ps}^{-}_{PRF^e(\alpha),\mathcal{C}_X}(u) = p(u) \prod_{\substack{G\in\mathcal{C}_X \\ G\neq G(u)}} \left(1 - \sum_{\substack{t\in G \\ u\not\succ t}} p(t)(1-\alpha)\right)$$

and the difference of probabilistic scores becomes:

$$\begin{aligned}
&\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{PRF^e(\alpha),\mathcal{C}_X}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{PRF^e(\alpha),\mathcal{C}_X}(v) = \\
&\qquad p(u)\left(1 - \sum_{\substack{t\in G(v) \\ t\in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} p(t)(1-\alpha) - \sum_{\substack{t_i\in G(v) \\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)(1-\alpha)\right) \\
&\qquad \times \prod_{\substack{G\in\mathcal{C}_X \\ G\neq G(u) \\ G\neq G(v)}} \left(1 - \sum_{\substack{t\in G \\ t\in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} p(t)(1-\alpha) - \sum_{\substack{t_i\in G \\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)(1-\alpha)\right) \\
&-p(v)\left(1 - \sum_{\substack{t\in G(u) \\ t\in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} p(t)(1-\alpha) - \sum_{\substack{t_i\in G(u) \\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)(1-\alpha) - \sum_{\substack{t\in G(u) \\ t\in \textsc{WrsBtr}(u,v)}} p(t)(1-\alpha) - p(u)(1-\alpha)\right) \\
&\times \prod_{\substack{G\in\mathcal{C}_X \\ G\neq G(u) \\ G\neq G(v)}} \left(1 - \sum_{\substack{t\in G \\ t\in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} p(t)(1-\alpha) - \sum_{\substack{t_i\in G \\ t_i\in \textsc{Ind}(u,v)}} y_i p(t_i)(1-\alpha) - \sum_{\substack{t\in G \\ t\in \textsc{WrsBtr}(u,v)}} p(t)(1-\alpha)\right)
\end{aligned} \tag{25}$$

Overall, the following rules apply to the case $G(u) \neq G(v)$:

RULES $(PRF^e(\alpha), \mathcal{C}_X)$. *Given the combination of $PRF^e(\alpha)$ ranking semantics and x-relation correlation model, assume $G(u) \neq G(v)$ and $u \succ v$. Then, it is $u \succ_p v$ if*

Rule1$_{PRF^e(\alpha),\mathcal{C}_X}$ *holds, or if and only if Rule2*$_{PRF^e(\alpha),\mathcal{C}_X}$ *holds:*

$$p(u)\left(1-\sum_{\substack{t\in G(v)\\ t\neq v}} p(t)(1-\alpha)\right) \geq p(v)\left(1-p(u)(1-\alpha)\right) \qquad (\text{Rule1}_{PRF^e(\alpha),\mathcal{C}_X})$$

$$p(u)\prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(u)}}\left(1-\sum_{\substack{t\in G\\ u\not\succ t}} p(t)(1-\alpha)+\sum_{\substack{t\in G\\ t\in \text{IND\_0}(u,v)}} p(t)(1-\alpha)\right)$$

$$\geq p(v)\prod_{\substack{G\in\mathcal{C}_X\\ G\neq G(v)}}\left(1-\sum_{\substack{t\in G\\ t\succ v}} p(t)(1-\alpha)-\sum_{\substack{t\in G\\ t\in \text{IND\_1}(u,v)}} p(t)(1-\alpha)\right) \qquad (\text{Rule2}_{PRF^e(\alpha),\mathcal{C}_X})$$

5.4. Summary of Results and Discussion

We first summarize in Table II the results obtained for the complexity of the IndArrange^P problem (the table also include results to be introduced later in this section). According to Definition 3.10, we remind that this is the complexity of determining how to arrange the tuples in the set $\text{IND}(u,v)$ (i.e., those indifferent to both tuples u and v) so as to obtain a (u,v)-adversarial order, measured as a function of $|\text{IND}(u,v)|$ only, thus disregarding the costs needed to compute probabilities and probabilistic scores.

Table II. Time complexity of IndArrange^P for different combinations of ranking semantics and correlation model. $M_{\max}$ is as in Theorem 5.3 and g is the maximum cardinality of a group. Results for the $PRF^e(\alpha)$ semantics derive from Lemma 5.7.

PRS\ Correlation model	independent		x-relation		general	
Expected rank	$\mathcal{O}(1)$	§5.1	$\mathcal{O}(2^g)$	Lemma 5.4	$\mathcal{O}(poly(M_{\max}))$	Thm. 5.3
Top-1, $PRF^e(\alpha)$	$\mathcal{O}(1)$	§5.2	$\mathcal{O}(2^g)$	Lemma 5.6	NP-hard	Thm. 5.5
Expected score	$\mathcal{O}(1)$		$\mathcal{O}(1)$		$\mathcal{O}(1)$	Lemma 5.8
Set-monotone	NP-hard	§5.4				

It has to be remarked that the bound $\mathcal{O}(2^g)$ that arises in case of x-relations has a different flavor for the ER and Top-1 (and $PRF^e(\alpha)$ as well) semantics: While for ER one can always resort to the polynomial solution available for arbitrary correlation models (Theorem 5.3), this is not possible for Top-1 and $PRF^e(\alpha)$. In particular, if the size of groups grows linearly with the number of tuples in R^p, solving IndArrange^P and therefore checking P-domination becomes NP-hard for Top-1 and $PRF^e(\alpha)$.

The last row in the table refers to a generic set-monotone PRS, and indeed shows that even the independent case, for which all the analyzed semantics exhibit constant-time complexity, can represent a challenging scenario for checking P-domination. To see why this is the case, consider the following PRS:

$$\mathsf{Ps}^{\geq}_{SS,\emptyset}(u)=\begin{cases}1+\sum_t p(t)-\sum_{t\succ u} p(t) & \text{if } p(u)>\sum_{t\succ u} p(t)\\ p(u) & \text{if } p(u)=\sum_{t\succ u} p(t)\\ 0 & \text{otherwise}\end{cases} \qquad (26)$$

that is easily demonstrated to be set-monotone. Now, assume $u \succ v$ and $p(v) = 2p(u) + \sum_{t \in \text{WRSBTR}(u,v)} p(t)$. The difference to be minimized becomes:

$$\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{SS,\emptyset}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{SS,\emptyset}(v) = \begin{cases} p(u) + \sum_{t \in \text{WRSBTR}(u,v)} p(t) & \text{if } p(u) > \sum_{t \succcurlyeq_{\mathbf{Y}} u} p(t) \\ -p(u) - \sum_{t \in \text{WRSBTR}(u,v)} p(t) & \text{if } p(u) = \sum_{t \succcurlyeq_{\mathbf{Y}} u} p(t) \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

Therefore, the INDARRANGE^P problem now amounts to determining whether there exists a subset IND_1(u,v) of tuples in IND(u,v) that can be arranged before u and such that $\sum_{t \in \text{IND\_1}(u,v)} p(t) = p(u) - \sum_{t \succ u} p(t)$. Since this is an instance of the NP-hard SUBSETSUM problem [Garey and Johnson 1979] the result in Table II follows. Notice that the problem remains NP-hard even if no dominance relationship exists among the tuples in IND(u,v). Although we do not expect this kind of "esoteric" PRS's to be used in practice, this result is of interest in that it shows that checking P-domination can become an expensive operation even when no correlation is present on the data.

5.4.1. Other Ranking Semantics. As shown in Section 5, each ranking semantics has its own P-domination rules, that Algorithm 2 can exploit to speedup skyline evaluation. A legitimate question is whether, given another (set-monotone) PRS Ψ, one is guaranteed that corresponding rules can be effectively derived for that PRS. In general, the answer is affirmative provided the probabilistic score $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(t)$ of a tuple t can be defined using a numerical expression. Indeed, if this is the case it is always possible to compute score bounds (due to Lemma 3.13), and consequently to derive Rule3$_{\Psi,\mathcal{C}}$. Similar arguments can be used to show that even Rule2$_{\Psi,\mathcal{C}}$ can be derived. On the other hand, the possibility of deriving Rule1$_{\Psi,\mathcal{C}}$ relies on the existence of an easy-to-compute lower bound of the difference $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v)$, a fact that in the general case is hard (if not impossible) to prove.

When $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(t)$ *cannot* be represented using a numerical expression, things become more complex to analyze. For instance, the recently introduced *quantile rank* (QR) semantics [Jestes et al. 2011] ranks tuples by considering the ϕ-quantile of their rank distribution (rather than the expected value of this distribution, as ER does).[9] QR needs to compute the rank distribution of all the tuples in R^p to yield the result of a top-k query, which can be done using dynamic programming techniques in $\mathcal{O}(N^2)$ time in the independent case and $\mathcal{O}(N|\mathcal{G}|^2)$ time in the x-relation case, where $|\mathcal{G}|$ is the number of groups in R^p. Although it can be proved that QR is set-monotone, there seems to be no "easy" way to solve the INDARRANGE problem for this semantics. Rather, for each arrangement of tuples in IND(u,v) one should compute the rank distributions of u and v resulting from that arrangement and then compare the corresponding ϕ-quantile values (note that if t follows u, then t has no influence on the rank distribution of u). Although this yields a valid method for checking P-domination, no equivalent of Rule2 seems to be available for the QR semantics.

On the positive side, there are other PRS's to which the polynomial complexity result derived for the ER semantics can be immediately extended. These are all the $(\Psi,\mathcal{C})$ combinations such that $\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{\Psi,\mathcal{C}}(u)$ is a linear function of the $\mathbf{Y}$ variables. When this is the case, the score difference to be minimized by INDARRANGE is also linear in $\mathbf{Y}$, and the arguments used in the proof of Theorem 5.3 apply almost unchanged.

Finally, one might also ask under which conditions $\mathcal{O}(1)$ complexity is obtainable for the INDARRANGE^P problem. This is the case if the difference of probabilistic scores,

[9]The *median rank* semantics, also introduced in [Jestes et al. 2011], is a particular case of QR obtained for $\phi = 0.5$.

$\mathsf{Ps}^{\succcurlyeq_Y}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{\succcurlyeq_Y}_{\Psi,\mathcal{C}}(v)$, is a monotone function of $\text{IND\_1}(u,v)$, that implies that such difference is minimized by setting either $\text{IND\_1}(u,v) = \text{IND}(u,v)$ or $\text{IND\_1}(u,v) = \emptyset$ (all tuples in $\text{IND}(u,v)$ either precede u or follow v, respectively).

5.4.2. Expected Score Semantics. According to Definition 3.1, a PRS is insensitive to tuple scores, since it only considers the linear order $\succcurlyeq$ that such scores induce on the tuples in R. It turns out that our results can be generalized to the case in which the ranking semantics, according to the terminology of [Jestes et al. 2011], is not *value-invariant*. We use the acronym V-PRS to denote semantics of this kind. Here we analyze the only V-PRS we are aware of, that is, *expected score* [Cormode et al. 2009], Appendix B discusses the general case.

The expected score (ES) semantics ranks tuples according to the value of $\mathsf{Ps}^{\succcurlyeq}_{ES,\bigstar}(u) = s(u) \times p(u)$. Note that, as in Example 3.7, we set $\mathcal{C} = \bigstar$ to denote that this semantics is not influenced by the correlation model. Besides this, the ES semantics also completely ignores the actual ordering of tuples $\succcurlyeq$.

The following simple result shows that the complexity of INDARRANGE^P is $\mathcal{O}(1)$:

LEMMA 5.8. *Let R^p be a probabilistic relation, and u and v two tuples in R^p. For the expected score semantics it is $u \succ_p v$ iff:*

$$u \succ v \wedge p(u) \geq p(v) \tag{28}$$

PROOF. (If) Since $u \succ v$ implies $s(u) \geq s(v)$ for any monotone scoring function $s()$ (whereas the converse, $s(v) \geq s(u)$, is not true), the result immediately follows.
(Only If) If $u \not\succ v$ then it is always possible to find, for given values of $p(u)$ and $p(v)$, a monotone scoring function $s()$, with $s(v) > s(u)$, such that $s(v) \times p(v) > s(u) \times p(u)$.
If $p(u) < p(v)$ and $u \succ v$, a scoring function $s()$ such that $s(u) = s(v)$ shows that u cannot P-dominate v. □

An immediate consequence of the above Lemma is that the ES-based skyline will always include all the tuples in the (deterministic) skyline of R, since the condition $u \succ v$ is necessary for P-domination to hold.

6. EXPERIMENTAL RESULTS

In this section we provide empirical evidence of the effects of the different combinations of ranking semantics and probabilistic models on the execution costs of INDARRANGE, P-domination tests, and skyline computation. Further, we also qualitatively analyze how the skyline changes when a different PRS and/or correlation model is adopted. Our results are obtained from a Java implementation on a AMD Phenom 2.29 GHz PC equipped with 1.75 GB of main memory, using both synthetic and real datasets.

For what follows, it has to be remarked that the INDARRANGE problem, whose complexity we have analyzed in Section 5, only contributes to part of the cost to be paid for checking P-domination (and therefore for computing the skyline). Indeed, as shown in Section 4.3, Rule1 or some combinations of correlation model and PRS allow us to check whether $u \succ_p v$ without solving INDARRANGE at all. Moreover, one should also consider the cost for actually computing probabilistic scores, which actually depends on the adopted PRS and correlation model.

6.1. Experiments on Synthetic Data

In order to study how performance varies with data characteristics, we used synthetic datasets, that allow the different parameters used for data generation to be freely modified. With respect to data distribution, we appropriately modified the code of the synthetic data generator introduced in [Börzsönyi et al. 2001]. With this generator,

data points are obtained by adding a fixed amount of Gaussian noise to points (i.e., tuples) in the $[0,1]^d$ space that are perfectly positively/negatively correlated:[10] For positive correlation, all points lie on the main diagonal of the space, whereas for negatively correlated points the sum of attribute values is 1 for each point (i.e., when a point is good on one attribute, it is bad on all the others). In order to finely control data correlation, we introduce a *spread* parameter, σ, that allows the amount of Gaussian noise to be freely modified. When $\sigma = -1$, we add zero noise to negatively correlated points, obtaining tuples that are all indifferent to each other. By increasing σ from -1 to 0 we progressively increase the variance of noise added to anti-correlated points. For $\sigma = 1$, on the other hand, points are positively correlated, and we have a linearly ordered dataset. When decreasing σ from 1 to 0 we progressively increase the variance of noise added to positively correlated points. Finally, when σ assumes the value 0, we get an uniform distribution, representing the middle way between negatively and positively correlated data distributions.

Regarding the probabilistic model, we first generate the probability of each tuple as a uniformly-distributed random value in $(0,1]$. For the case of x-relations, groups are built as in [Yi et al. 2008]: A fraction $f_G = 10\%$ of tuples is involved in groups containing $N_G = 2$ tuples, while all other tuples belong to "singleton" groups. Tuples inserted in groups are randomly picked from the dataset and this is repeated in case the probability of the group exceeds 1.

Table III shows the range of parameters used in the experiments; unless otherwise stated, default values are used.

Table III. Settings for the generation of synthetic datasets.

symbol	description	range	default
d	dimensionality	$2 \div 5$	4
N	cardinality	$10 \div 100K$	$50K$
σ	spread	$-1 \div 1$	0
$\mathcal{C}$	correlation model	$\{\emptyset, \mathcal{C}_X\}$	$\emptyset$

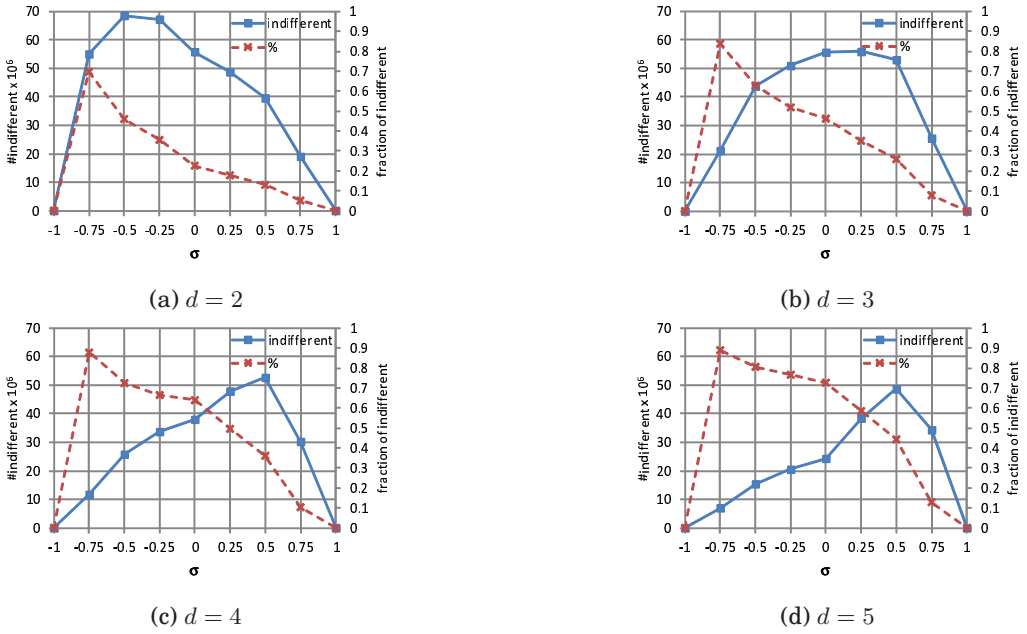
6.1.1. Size of the INDARRANGE *Problem.* Our first experiment aims to understand how the characteristics of the dataset can influence the time required to solve the INDARRANGE problem. Since this is mainly influenced by the value of $|\text{IND}(u,v)|$, we analyze how this varies with the parameters used for data generation. Notice that with datasets having positively correlated attributes, the number of times INDARRANGE has to be solved is higher with respect to the negatively correlated case (since $u \succ v$ is more likely to occur), yet the size of $|\text{IND}(u,v)|$ is expected to be smaller. Therefore, it is interesting to study not only the average fraction of indifferent tuples, $|\text{IND}(u,v)|/N$, but also the total number of such tuples (obtained as the sum of $|\text{IND}(u,v)|$ over all pairs of tuples u and v such that $u \succ v$). Figure 4 shows how such values vary when modifying the spread σ and the dimensionality d for a dataset of $N = 1000$ tuples.

As expected, when increasing σ, i.e., moving towards positively correlated datasets, the fraction of indifferent tuples decreases, while such value increases with d. Notice that for $\sigma = -1$ both graphs have value 0 because all tuples are indifferent, thus it is never the case that $u \succ v$.

By looking at the total number of indifferent tuples, we see that the hardest case varies for different d values: when $d = 2$, the most difficult scenario is obtained for $\sigma = -0.5$, whereas for $d = 5$ the graph peaks at $\sigma = 0.5$. Positively correlated datasets

[10]Note that here the term "correlation" refers to attribute values, i.e., it has nothing to do with the adopted probabilistic model. Indeed, the generator was introduced by Börzsönyi et al. [2001] to study skyline algorithms in the deterministic case.

are therefore increasingly harder for higher values of d. From now on, we will use the value $\sigma = 0$ that represents a fairly difficult scenario at all dimensionalities.



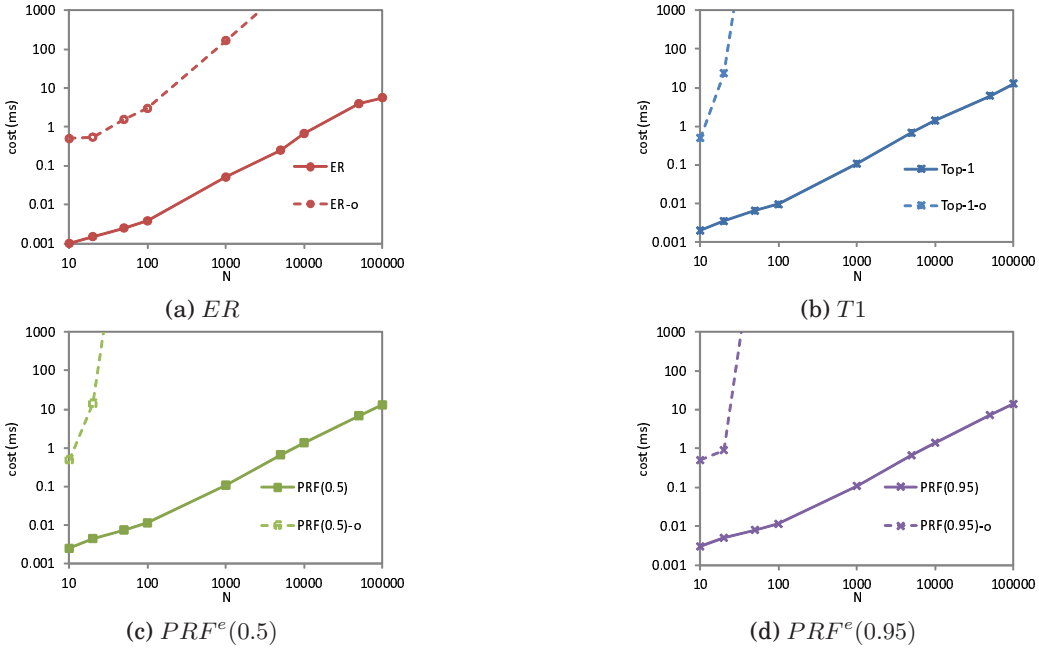
(a) $d = 2$ (b) $d = 3$

(c) $d = 4$ (d) $d = 5$

Fig. 4. Total number and average fraction of indifferent tuples vs. spread σ. $N = 1000$.

6.1.2. Cost of P-domination Tests. The objective of the next experiment is to show that checking P-domination in the tightly integrated scenario is much faster that in the loosely integrated one: In order to attain good performances, it is therefore essential for the skyline algorithm to have knowledge of the underlying probabilistic model. To this end, Figure 5 shows in a log-log scale the average time needed to perform a P-domination test under the independent probabilistic model for 4 different PRS's, namely ER, $T1$, $PRF^e(0.5)$, and $PRF^e(0.95)$, when varying the number of tuples N. In each graph, the solid line represents the tight integration scenario, while the dashed line corresponds to the loosely integrated one (dubbed "-o" in the legend). For the tight integration scenario a single test involves first checking Rule1 and, if this fails, evaluating Rule2. As expected, in all cases the tight integration scenario outperforms the loose integration one by several orders of magnitude. Similar results were obtained also for x-relations and are reported in the Electronic Appendix. Indeed, only for the ER semantics using an external module for computing probabilities allows P-domination to be checked in less than 1 second for a dataset of $N = 1000$ tuples, whereas for all other PRS's the test becomes prohibitively costly already when $N = 100$. The behavior of ER is a direct consequence of Theorem 5.3, which guarantees polynomial-time complexity of INDARRANGE^P for any probabilistic model.

Having ruled out the loose integration scenario, in Figure 6 we compare P-domination costs for the independent ("-i" suffix) and x-relation ("-x" suffix) correlation models. All the graphs exhibit a linear trend with respect to the dataset cardinality. Since all the considered PRS's require $\mathcal{O}(N)$ time to compute the probabilistic score of a tuple, this result was expected for the independent case, for which the complexity of INDARRANGE^P is $\mathcal{O}(1)$ for all the PRS's. The same also holds for the x-relation



(a) ER (b) $T1$

(c) $PRF^e(0.5)$ (d) $PRF^e(0.95)$

Fig. 5. Average cost for assessing P-domination between two tuples vs. dataset cardinality. Independent data and different ranking semantics.

case, since in this experiment the size of each group is held constant, which still guarantees $\mathcal{O}(1)$ complexity of IndArrange^P and, consequently, $\mathcal{O}(N)$ time for checking P-domination.

In order to study how costs vary with the size of groups, for the next experiment we randomly chose a number of tuples in $\text{Ind}(u, v)$ and moved all of them into the group of u (or v, see the proof of Lemma 5.4). Doing this way, we were able to vary the size g of $\text{Ind}(u, v) \cap G(u)$ ($\text{Ind}(u, v) \cap G(v)$, respectively), without altering the size of $\text{Ind}(u, v)$. In Figure 7 we plot the cost of checking P-domination in two different scenarios. Graph Top-1-x shows the cost of applying Rule2 for the $T1$ semantics, for which the complexity of IndArrange^P is $\mathcal{O}(2^g)$ (see Theorem 5.6). Results indeed confirm the exponential dependency on g (note that the graph uses a log scale). Graph ER-o is the P-domination cost of the ER semantics in the loose integration scenario. Results are in line with Theorem 5.3, ensuring that the complexity of IndArrange^P is polynomial in the size of $\text{Ind}(u, v)$, thus independent of g as long as this does not influence $|\text{Ind}(u, v)|$.

In Figure 8 we contrast the different Prs's for the two considered correlation models. From Figure 8(a) we see that in the independent case all the semantics but ER have very similar costs. This is only due to the lower number of arithmetic operations that ER needs to do for computing probabilistic scores. A similar trend is observed in the case of x-relations (Figure 8(b)). In this case, however, a major difference in costs is observed. Here the low cost paid by ER is also because this semantics is the only one that can reuse precomputed bounds when checking Rule2. On the other hand, the difference between $T1$ and the two $PRF^e(\alpha)$ semantics is mainly due to the fact that Rule1 is much more likely to succeed for $T1$ than for $PRF^e(0.5)$, and for $PRF^e(0.5)$ than for $PRF^e(0.95)$ (see also Figure 10 in Section 6.1.3 and Figure 16 in the Electronic

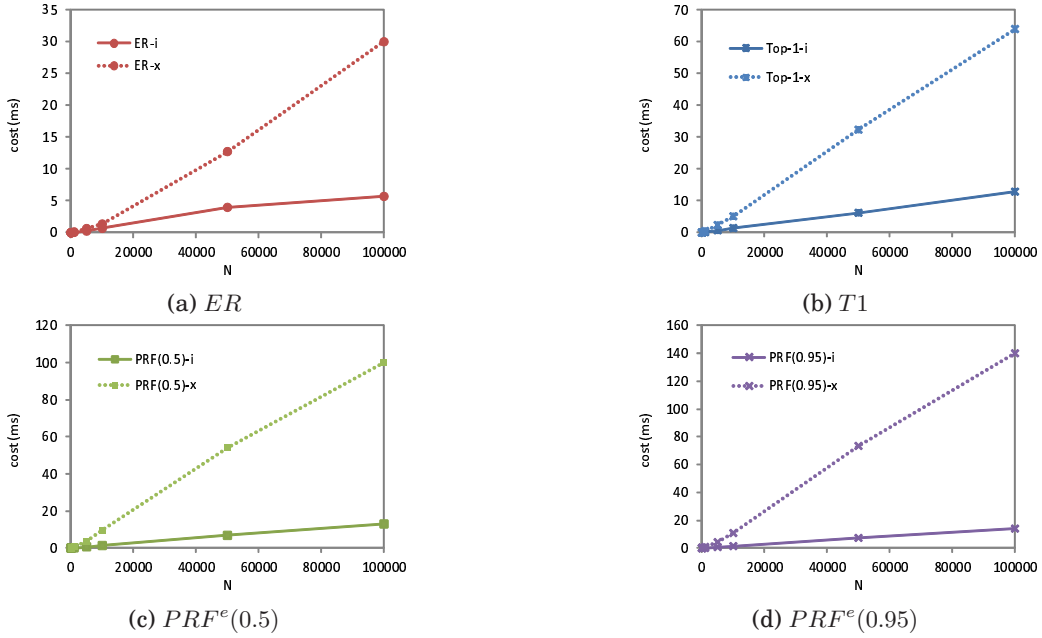


Fig. 6. Average cost for assessing P-domination between two tuples vs. dataset cardinality. Tight integration scenario, independent data/x-relations under different ranking semantics.

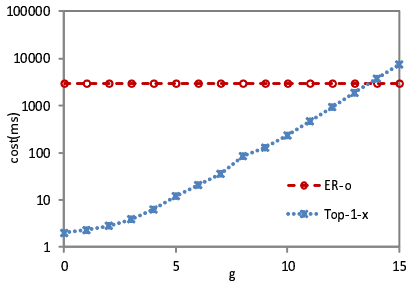


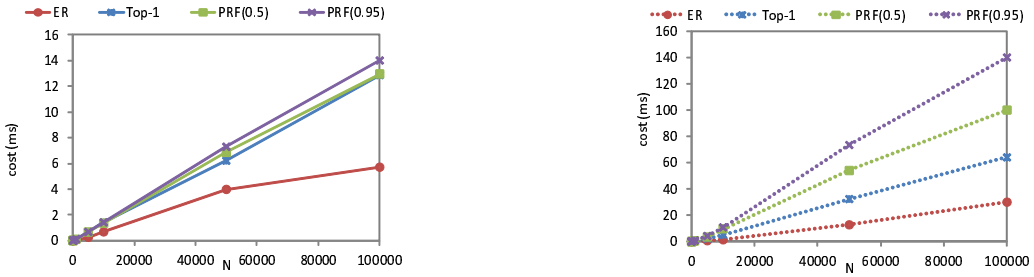
Fig. 7. Average cost of Rule2, varying the size of $g = |\text{IND}(u, v) \cap G(u)|$. $N = 5,000$. See text for details on the displayed graphs.

Appendix), which directly impacts on the number of times the more expensive Rule2 has to be checked.

We now consider the effect of data dimensionality on P-domination costs. Figure 9 shows the average time spent for a P-domination test (again, in the case $u \succ v$) under the independent (a) and x-relation (b) models. In all cases, costs increase with d, since each of the (deterministic) domination tests needed to evaluate Rule2 requires $\mathcal{O}(d)$ time in the worst case. However, it can be observed from the graphs that the cost at, say, $2d$ is less than twice the cost at d. This can be explained by noticing that increasing the number of dimensions reduces the number of cases in which domination occurs, and therefore domination tests can be interrupted earlier.

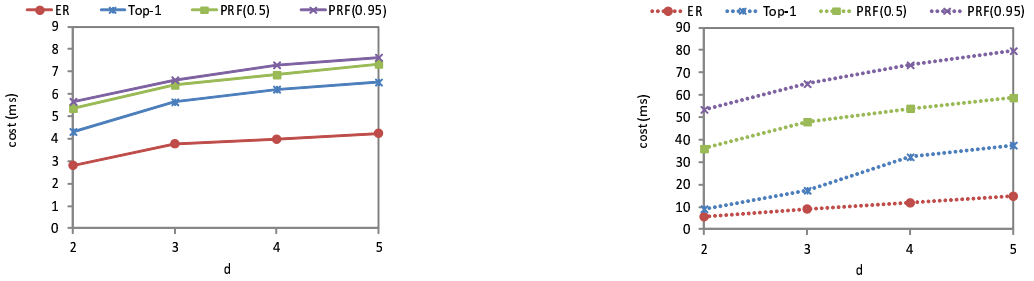
Comparing PRS's, ER is always the cheapest alternative.

6.1.3. Effectiveness of Rules. Since, as stated in Section 4.2, checking P-domination between two tuples u and v (with $u \succ v$) can be done by first exploiting the cheap Rule1



(a) independent data (b) x-relations

Fig. 8. Average cost for assessing P-domination between two tuples vs. dataset cardinality. Tight integration scenario.



(a) independent data (b) x-relations

Fig. 9. Average cost for assessing P-domination between two tuples vs. number of attributes. Tight integration scenario.

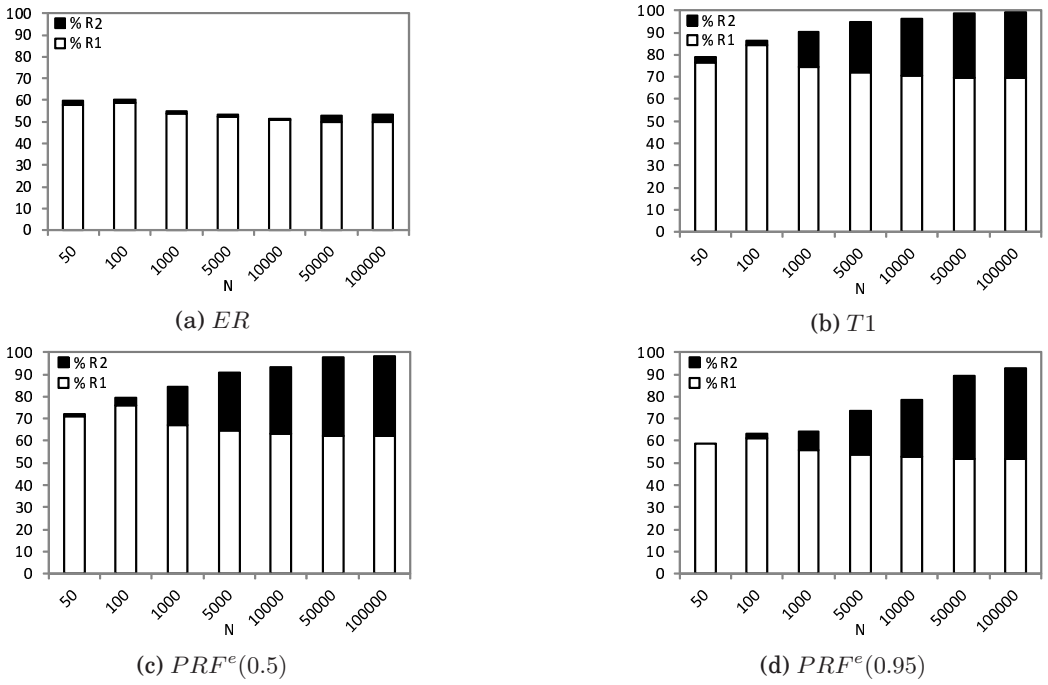
and then the costly Rule2, it is important to assess the impact of Rule1 on the cost of a single P-domination test. Figure 10 shows the chance of success of a P-domination test in the independent case. Graphs highlight some interesting facts:

— Rule1 is very effective, allowing Rule2 to be avoided always more than 50% of times for ER, and more than 68% for $T1$. With $PRF^e(\alpha)$ the effectiveness of Rule1 is lower than in the $T1$ case and decreases with increasing values of α (remind that $T1$ coincides with $PRF^e(0)$).
— When $u \succ_p v$, this is discovered by just using Rule1 in more than 94% of the cases for the ER semantics and more than 69% of the cases for the $T1$ semantics.
— The effectiveness of Rule1 slightly decreases with N. Since tuple probabilities are uniformly distributed, in the independent case the success probability of Rule1 in the limit $N \to \infty$ can be obtained as $\int_0^1 (1-x)dx = 1/2$ and $\int_0^1 (1 - \frac{x}{1+x})dx = \ln 2$, for the ER and $T1$ PRS's, respectively.

Results for x-relations, shown in the Electronic Appendix, confirm the above trend.

6.1.4. Cost of Skyline Computation. Figure 11 shows the costs for computing the skylines resulting from the ER and the $T1$ PRS's under the independent and x-relation correlation models. It is again evident that the tight integration scenario allows costs to be consistently reduced. Note that this has two different causes:

(1) The cheap Rule1 (only available in the tight integration case) can reduce the costs of the first phase of the skyline algorithm, since bounds will not be computed for already-discovered P-dominated tuples.



(a) ER (b) $T1$

(c) $PRF^e(0.5)$ (d) $PRF^e(0.95)$

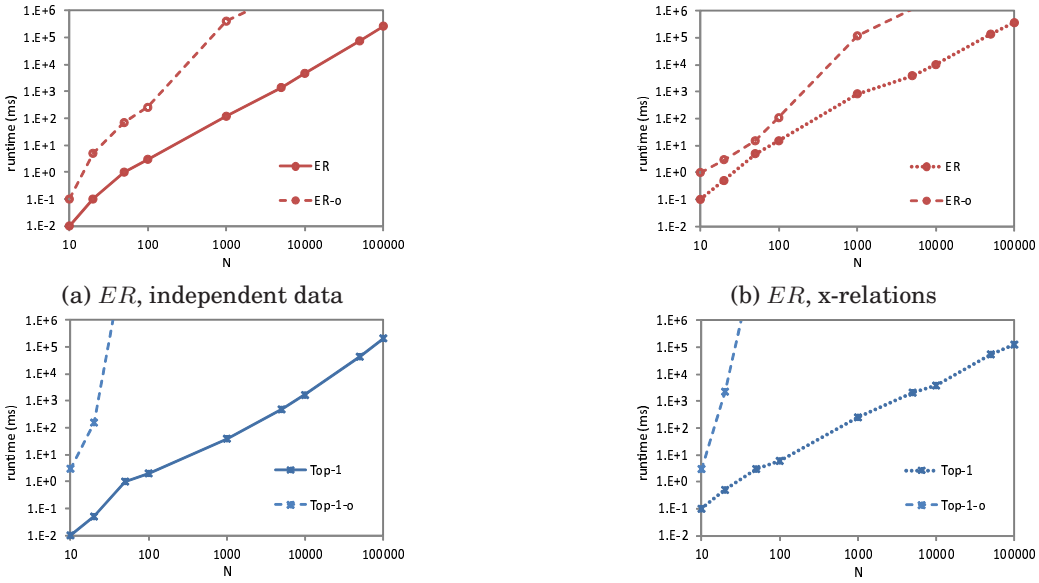
Fig. 10. Effectiveness of Rule1 (R1) and Rule2 (R2) on P-domination tests vs. dataset cardinality. Independent data.

(2) During the second phase of the skyline algorithm, the P-domination test for the tight integration scenario is cheaper than the loose integration case.

As another interesting consideration, graphs show that skyline computation costs for the ER and the $T1$ semantics are comparable. This demonstrates that estimating the cost of a single P-domination test is not sufficient for understanding if the skyline can be efficiently computed. Indeed, although the $T1$ PRS has a higher P-domination cost than ER (see Figure 6), its Rule1 has a higher chance of success (see Figures 10 and 16), thus a higher number of tuples are discarded during the first phase of Algorithm 2. This shows that the cost of skyline computation is indeed dependent on a variety of factors, which include the efficiency of P-domination tests and the effectiveness of Rule1, that can interact in a complex way. For instance, the Electronic Appendix includes skyline computation costs for the IIP dataset (this dataset will also be used in Section 6.2): For such dataset, ER is the cheapest alternative when x-relations are considered, while $T1$ is the fastest PRS for the independent case.

6.1.5. Comparison with p-skyline. Although the p-skyline approach by Pei et al. [2007] is not directly comparable to the one based on P-domination, we compared their costs and results to gain some insight about the differences of the two approaches. For computing the p-skyline we adopted an exhaustive algorithm, since the pruning techniques proposed in [Pei et al. 2007] (which are specifically derived for retrieving the best groups in a x-relation) are not applicable to our scenario, that works at the tuple level.

In order to compare the two approaches on a fair basis, we first computed $\text{SKY}(R^p)$ for the ER and the $T1$ PRS's under the x-relation model. Then, we computed the skyline probability of all tuples in R^p and retrieved only the best $|\text{SKY}(R^p)|$ tuples to form the p-skyline. Figure 12 (a) shows the fraction of tuples that are in common between



(a) ER, independent data (b) ER, x-relations

(c) $T1$, independent data (d) $T1$, x-relations

Fig. 11. Runtime of skyline computation vs. dataset cardinality. Independent data/x-relations and different ranking semantics.

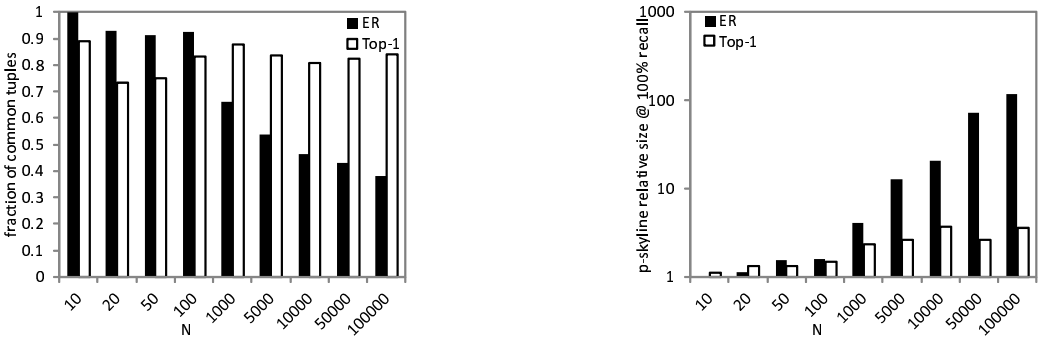
$\text{SKY}(R^p)$ and the p-skyline. Since the formulation of $\Pr_{\text{SKY}}$ closely resembles that of $T1$ (indeed, they coincide in the 1-dimensional case and $\text{Ps}^+_{T1,\mathcal{C}}(t) = \Pr_{\text{SKY}}(t)$ for any $\mathcal{C}$), it is not surprising that their results are similar (about 80% of tuples in common). Things are different with ER: the fraction of common tuples decreases with N, reaching about 38% for $N = 100,000$ tuples. It is well known that probability values have a major impact on the ranking obtained by ER, i.e., $\text{SKY}(R^p)$ is more likely to contain highly probable tuples that may have not so good attribute values [Li et al. 2009].

In Figure 12 (b) we show the size of p-skyline when all tuples in $\text{SKY}(R^p)$ have been retrieved, normalized with respect to $|\text{SKY}(R^p)|$. Note that this coincides with the worst (normalized) rank that a tuple in $\text{SKY}(R^p)$ has when tuples are ordered using skyline probabilities. With 100k tuples, when using $T1$ we have to retrieve about $3 \times |\text{SKY}(R^p)|$ tuples, while this increases to 122 times the size of $\text{SKY}(R^p)$ for ER. Therefore, one of the tuples in the ER skyline has rank $97,071 = 122 \times |\text{SKY}(R^p)|$ (i.e., in the bottom 3% of the dataset) when considering skyline probabilities.

Finally, Figure 13 contrasts execution times for computing the result in the independent and the x-relation cases. For large datasets, computing $\text{SKY}(R^p)$ is one order of magnitude faster than the exhaustive evaluation of all skyline probabilities. However, since the pruning techniques proposed by Pei et al. [2007] are able to reduce costs with respect to the exhaustive algorithm by one order of magnitude, we conclude that costs of the two approaches are comparable.

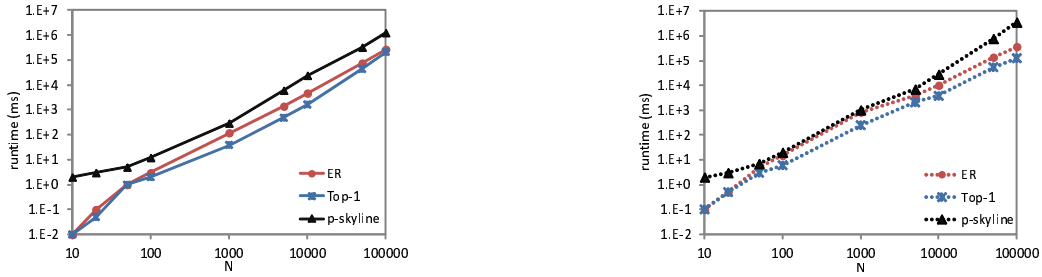
6.2. Experiments on the IIP Dataset

In our last experiment, we used the IIP Iceberg Sightings Dataset, that we already described in Example 1.1. We considered as skyline attributes the size of the iceberg and its distance to a given location, mimicking the fact that we deem dangerous those large icebergs that are drifting close to a target we consider somewhat important.



(a) (b)

Fig. 12. Comparison between results: p-skyline vs. ER and $T1$. (a) fraction of common tuples, (b) relative size of p-skyline when all tuples in $\text{SKY}(R^p)$ are retrieved.



(a) independent data (b) x-relations

Fig. 13. Average cost vs. dataset cardinality: p-Skyline, ER, and $T1$ semantics.

The size attribute for each iceberg was obtained from the symbolic value included in the dataset by converting the nine existing sizes into a decimal value. Tuple probability was generated in the same way from the eight existing confidence levels. Since the above mentioned values were obtained from symbolic values, we obtained a large number of duplicate values: To break ties, following Li et al. [2009], we added Gaussian noise to tuples' size and probability. Finally, we considered as target a randomly chosen location close to the centroid of all the iceberg sightings.

Then, we generated three different correlation models: Besides the obvious independent model, an x-relation was generated by first clustering together icebergs sighted in a same day with a L_∞ distance smaller than 0.1 degrees. Then, to ensure the transitivity of mutual exclusion we aggregated together clusters with common tuples: In this case two tuples that are far away from each other are mutually exclusive if a chain of close-enough tuples connecting them exists. At the end of this process, the average group size is 1.66, with the largest group including 503 tuples. We also re-scaled the probability of each tuple in a group so as to ensure that the overall group probability does not exceed 1. Finally, we also considered a more complex correlation model, denoted $\mathcal{C}_m$, in which mutual exclusion is not a transitive relationship, i.e., only icebergs whose L_∞ distance does not exceed 0.1 degrees cannot appear together in a same possible world. To obtain this model, the above-described cluster aggregation step was not executed.

We compared the performance of 9 different combinations of PRS's and probabilistic models, as listed in Table IV.

Table IV. Different $(\Psi, \mathcal{C})$ combinations used in the IIP experiment.

Ψ	$\mathcal{C}$
ER	$\emptyset, \mathcal{C}_X, \mathcal{C}_m$
$T1$	$\emptyset, \mathcal{C}_X$
$PRF^e(0.5)$	$\emptyset, \mathcal{C}_X$
$PRF^e(0.95)$	$\emptyset, \mathcal{C}_X$

Figure 14 compares the so-obtained skylines, also including the deterministic skyline (a graphical representation of some of these skylines can be found in the Electronic Appendix). The height of a column marked $(\Psi, \mathcal{C})$ in correspondence of a category $(\Psi', \mathcal{C}')$ on the horizontal axis equals the fraction of $(\Psi, \mathcal{C})$ skyline tuples that are also found in the $(\Psi', \mathcal{C}')$ skyline. For instance, tuples in the $(PRF^e(0.95), \emptyset)$ skyline represent about the 80% of the tuples in the $(T1, \emptyset)$ skyline, while $(ER, \emptyset)$ and $(T1, \emptyset)$ skylines have no common tuple.

As a first observation, using a different probabilistic model under the same ranking semantics yields a different skyline. As an example, only 20 out of 35 $(ER, \emptyset)$ skyline tuples are also included in the skyline of $(ER, \mathcal{C}_X)$ (that has 34 tuples).

Although each skyline presents some differences with respect to others, this is particularly evident for ER skylines. This fact confirms the results in [Li et al. 2011], where the ER PRS was observed to yield a quite different ranking with respect to other semantics.

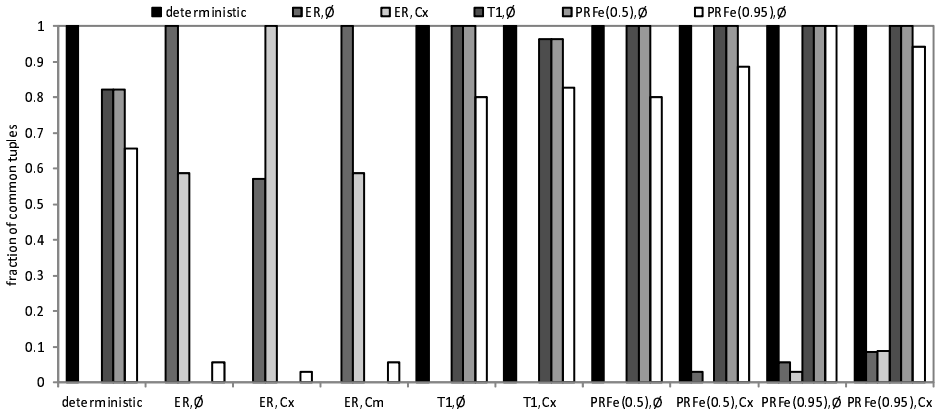


Fig. 14. Comparison of skylines for different $(\Psi, \mathcal{C})$ combinations.

7. CONCLUSIONS

In this paper we have addressed the problem of studying how to compute the skyline of a probabilistic relation for arbitrary correlation models and ranking semantics. We have shown how P-domination (i.e., domination between probabilistic tuples), which lies at the heart of our approach, can be formulated as an optimization problem (INDARRANGE), whose complexity we have characterized for a variety of combinations of ranking semantics and correlation models, including the most general case in which the probability of events can only be obtained through an oracle. Our findings, summarized in Table II, show that the complexity of checking P-domination is largely influenced by both problem coordinates (probabilistic model and ranking semantics). For each analyzed case we have also derived specific P-domination rules, which are

exploited by one of the two skyline algorithms we have detailed. Our experimental evaluation of the INDARRANGE problem, P-domination rules, and skyline algorithms confirms the theoretical analysis.

ACKNOWLEDGMENTS

We would like to thank the anonymous referees for their insightful suggestions.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

REFERENCES

AFRATI, F. N., KOUTRIS, P., SUCIU, D., AND ULLMAN, J. D. 2012. Parallel skyline queries. In *Proceedings of the 15th International Conference on Database Theory (ICDT'12)*. ACM Press, Berlin, Germany, 274–284.

AGRAWAL, P., BENJELLOUN, O., DAS SARMA, A., HAYWORTH, C., NABAR, S. U., SUGIHARA, T., AND WIDOM, J. 2006. Trio: A system for data, uncertainty, and lineage. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*. ACM Press, Seoul, Korea, 1151–1154.

ATALLAH, M. J., QI, Y., AND YUAN, H. 2011. Asymptotically efficient algorithms for skyline probabilities of uncertain data. *ACM Transactions on Database Systems 36,* 2, Article 12.

BALAZINSKA, M., DESHPANDE, A., FRANKLIN, M. J., GIBBONS, P. B., GRAY, J., HANSEN, M. H., LIEBHOLD, M., NATH, S., SZALAY, A. S., AND TAO, V. 2007. Data management in the worldwide sensor web. *IEEE Pervasive Computing 6,* 2, 30–40.

BARTOLINI, I., CIACCIA, P., AND PATELLA, M. 2008. Efficient sort-based skyline evaluation. *ACM Transactions on Database Systems 33,* 4, Article 31.

BARTOLINI, I., CIACCIA, P., AND PATELLA, M. 2013. The skyline of a probabilistic relation. *IEEE Transactions on Knowledge and Data Engineering 25,* 7, 1656–1669.

BENJELLOUN, O., DAS SARMA, A., HALEVY, A. Y., THEOBALD, M., AND WIDOM, J. 2008. Databases with uncertainty and lineage. *The VLDB Journal 17,* 2 (Mar.), 243–264.

BÖRZSÖNYI, S., KOSSMANN, D., AND STOCKER, K. 2001. The skyline operator. In *Proceedings of the 17th International Conference on Data Engineering (ICDE 2001)*. IEEE Computer Society, Heidelberg, Germany, 421–430.

CHOMICKI, J., CIACCIA, P., AND MENEGHETTI, N. 2013. Skyline queries, front and back. *SIGMOD Record 42,* 3, 6–18.

CHOMICKI, J., GODFREY, P., GRYZ, J., AND LIANG, D. 2003. Skyline with presorting. In *Proceedings of the 19th International Conference on Data Engineering (ICDE 2003)*. IEEE Computer Society, Bangalore, India, 717–719.

CHONG, C.-Y. AND KUMAR, S. P. 2003. Sensor networks: Evolution, opportunities, and challenges. *Proceedings of the IEEE 91,* 8, 1247–1256.

CORMODE, G., LI, F., AND YI, K. 2009. Semantics of ranking queries for probabilistic data and expected ranks. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2009)*. IEEE Computer Society, Shanghai, China, 305–316.

COWELL, R. G., DAWID, P., LAURITZEN, S. L., AND SPIEGELHALTER, D. J. 1999. *Probabilistic Networks and Expert Systems*. Springer, Berlin, Germany.

DALVI, N. N., RE, C., AND SUCIU, D. 2011. Queries and materialized views on probabilistic databases. *Journal of Computer and System Sciences 77,* 3, 473–490.

DALVI, N. N. AND SUCIU, D. 2004. Efficient query evaluation on probabilistic databases. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB 2004)*. Morgan Kaufmann, Toronto, ON, 864–875.

DAS SARMA, A., BENJELLOUN, O., HALEVY, A. Y., AND WIDOM, J. 2006. Working models for uncertain data. In *Proceedings of the 22nd International Conference on Data Engineering, (ICDE 2006)*. IEEE Computer Society, Atlanta, GA, Article 7.

DONG, X. L., HALEVY, A. Y., AND YU, C. 2007. Data integration with uncertainty. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB 2007)*. ACM Press, Vienna, Austria, 687–698.

EMRICH, T., KRIEGEL, H.-P., MAMOULIS, N., RENZ, M., AND ZÜFLE, A. 2012. Querying uncertain spatio-temporal data. In *Proceedings of the 28th International Conference on Data Engineering (ICDE 2012)*. IEEE Computer Society, Washington, DC, 354–365.

GAREY, M. R. AND JOHNSON, D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, CA.

GODFREY, P., SHIPLEY, R., AND GRYZ, J. 2005. Maximal vector computation in large data sets. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*. ACM Press, Trondheim, Norway, 229–240.

JESTES, J., CORMODE, G., LI, F., AND YI, K. 2011. Semantics of ranking queries for probabilistic data. *IEEE Transactions on Knowledge and Data Engineering 23,* 12, 1903–1917.

KLEINBERG, J. AND TARDOS, È. 2006. *Algorithm Design*. Addison-Wesley.

LI, J., SAHA, B., AND DESHPANDE, A. 2009. A unified approach to ranking in probabilistic databases. In *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB 2009)*. ACM Press, Lyon, France, 502–513.

LI, J., SAHA, B., AND DESHPANDE, A. 2011. A unified approach to ranking in probabilistic databases. *The VLDB Journal 20,* 2, 249–275.

LIN, X., ZHANG, Y., ZHANG, W., AND CHEEMA, M. A. 2011. Stochastic skyline operator. In *Proceedings of the 25th International Conference on Data Engineering (ICDE 2011)*. IEEE Computer Society, Hannover, Germany, 721–732.

MORSE, M. D., PATEL, J. M., AND JAGADISH, H. V. 2007. Efficient skyline computation over low-cardinality domains. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB 2007)*. ACM Press, Vienna, Austria, 267–278.

PAPADIAS, D., TAO, Y., FU, G., AND SEEGER, B. 2005. Progressive skyline computation in database systems. *ACM Transactions on Database Systems 30,* 1, 41–82.

PEARL, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA.

PEI, J., JIANG, B., LI, X., AND YUAN, Y. 2007. Probabilistic skylines on uncertain data. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB 2007)*. ACM Press, Vienna, Austria, 15–26.

SOLIMAN, M. A., ILYAS, I. F., AND CHANG, K. C.-C. 2007. Top-k query processing in uncertain databases. In *Proceedings of the 23th International Conference on Data Engineering (ICDE 2007)*. IEEE Computer Society, Istanbul, Turkey, 896–905.

SOLIMAN, M. A., ILYAS, I. F., AND CHANG, K. C.-C. 2008. Probabilistic top-k and ranking-aggregate queries. *ACM Transactions on Database Systems 33,* 3, Article 13.

TAO, Y. AND PAPADIAS, D. 2006. Maintaining sliding window skylines on data streams. *IEEE Transactions on Knowledge and Data Engineering 18,* 2, 377–391.

TRIMPONIAS, G., BARTOLINI, I., PAPADIAS, D., AND YANG, Y. 2013. Skyline processing on distributed vertical decompositions. *IEEE Transactions on Knowledge and Data Engineering 25,* 4, 850–862.

YAN, D. AND NG, W. 2011. Robust ranking of uncertain data. In *Proceedings of the 16th International Conference on Database Systems for Advanced Applications (DASFAA 2011), Part I*. Springer, Hong Kong, China, 254–268.

YI, K., LI, F., KOLLIOS, G., AND SRIVASTAVA, D. 2008. Efficient processing of top-k queries in uncertain databases with x-relations. *IEEE Transactions on Knowledge and Data Engineering 20,* 12, 1669–1682.

YUAN, Y., LIN, X., LIU, Q., WANG, W., YU, J. X., AND ZHANG, Q. 2005. Efficient computation of the skyline cube. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005)*. ACM Press, Trondheim, Norway, 241–252.

ZHANG, S., MAMOULIS, N., AND CHEUNG, D. W. 2009. Scalable skyline computation using object-based space partitioning. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. ACM Press, Providence, RI, 483–494.

ZHANG, W., LIN, X., ZHANG, Y., CHEEMA, M. A., AND ZHANG, Q. 2012. Stochastic skylines. *ACM Transactions on Database Systems 37,* 2, Article 14.

ZHANG, X. AND CHOMICKI, J. 2008. On the semantics and evaluation of top-k queries in probabilistic databases. In *Proceedings of the 2nd International Workshop on Ranking in Databases (DBRank 2008)*. IEEE Computer Society, Cancun, Mexico, 556–563.

ZHANG, X. AND CHOMICKI, J. 2009. Semantics and evaluation of top-k queries in probabilistic databases. *Distributed and Parallel Databases 26,* 1, 67–126.

APPENDICES

A. PROOFS OF FORMAL RESULTS

PROOF OF THEOREM 5.5.
For proving both facts it is convenient to first consider a generalization of the PARTITION problem, which is a basic NP-complete problem [Garey and Johnson 1979]. An instance of PARTITION consists of a set $A = \{a_1, \ldots, a_n\}$ of positive integers, and the problem is to determine if there is a subset $A' \subseteq A$ such that $\sum_{a_i \in A'} a_i = \sum_{a_i \in A \setminus A'} a_i$. The generalization we consider is that of *unbalanced* partitions, which leads to what we call the α-PARTITION problem, where α is an *a-priori given* rational number in $(0, 1)$:

α-PARTITION problem
INSTANCE: A set $S = \{w_1, \ldots, w_n\}$ of positive integers.
QUESTION: Is there a subset $S' \subseteq S$ such that $\sum_{s_i \in S'} w_i = \alpha \sum_{s_i \in S} w_i$?

Notice that there is a subtle difference between α-PARTITION and the well-known SUBSETSUM problem, which asks for a subset whose elements sum to B [Garey and Johnson 1979]. This is because in SUBSETSUM the value of B is part of the problem instance, whereas in α-PARTITION this is not the case. Also observe that $\alpha = 1/2$ is the classical PARTITION problem.

THEOREM A.1. *The α-PARTITION problem is NP-complete for any $\alpha \in (0, 1)$.*

PROOF. Without loss of generality, let $\alpha < 1/2$ (the case $\alpha > 1/2$ requires similar arguments). The problem is clearly in NP. For the hardness part, we reduce PARTITION to α-PARTITION in polynomial time as follows. For each element $a_i \in A$, we have a corresponding element $w_i \in S$, with $w_i = a_i$. The set S contains two additional elements, denoted b_1 and b_2. By denoting with $W(T)$ the sum of the elements in a set T, b_1 and b_2 are:

$$b_1 = \frac{r\alpha}{2} W(A) - \frac{W(A)}{2}$$
$$b_2 = \frac{r(1-\alpha)}{2} W(A) - \frac{W(A)}{2}$$

where r is an integer satisfying $r\alpha > 2$, and such that $r\alpha$ is integer. Note that such r always exists since α is rational.

Since $W(S) = W(A) + b_1 + b_2 = \frac{r}{2} W(A)$, a solution to α-PARTITION will consists of two subsets with sums:

$$\frac{r\alpha}{2} W(A) \qquad \text{and} \qquad \frac{r(1-\alpha)}{2} W(A)$$

Since $b_1 + b_2 = \frac{r}{2} W(A) - W(A) = \frac{r}{2} W(A)(1 - 2/r) > \frac{r}{2} W(A)(1 - \alpha)$ (because $r\alpha > 2$), b_1 and b_2 cannot be put into a same subset of S by any solution of α-PARTITION. It follows that a solution of α-PARTITION will partition the elements of A into two subsets with equal sum $W(A)/2$, adding to each of them one of the b_1 and b_2 elements. The same arguments apply to show that a solution to PARTITION can be used to derive a solution to α-PARTITION. □

We now turn back to the INDARRANGE^P problem, for which we exhibit a polynomial-time reduction from α-PARTITION. Let $S = \{w_1, \ldots, w_n\}$ be an instance of α-PARTITION, and assume $\alpha > 1/2$, such that $B = \alpha W(S)$ is integer (the exact value

of α will be specified later). The corresponding instance of IndArrange^P is generated as follows:

— The relation R consists of $2n+3$ tuples: $u, v, t, v_i (i = 1, \ldots, n)$, and $t_i (i = 1, \ldots, n)$.
— Besides $u \succ v$, the other dominance relationships are: $u \succ t$, $t \succ v$, and $t_i \succ v_i (i = 1, \ldots, n)$
— The correlation model is that of x-relations, and there are 3 groups (tuples in a same group being mutually exclusive): $G(u) = \{u\}$, $G(v) = \{v, v_i, i = 1, \ldots, n\}$, and $G(t) = \{t, t_i, i = 1, \ldots, n\}$.

With above positions, IndArrange^P has to minimize the difference:

$$\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{T1,\mathcal{C}_X}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{T1,\mathcal{C}_X}(v) =$$
$$p(u)\left(1 - \sum_{v_i \in G(v)} y_i p(v_i)\right)\left(1 - \sum_{t_i \in G(t)} y_i p(t_i)\right) - p(v)(1-p(u))\left(1 - p(t) - \sum_{t_i \in G(t)} y_i p(t_i)\right)$$

We now set $p(u) = p(v)(1-p(u))$, under the constraint $p(v) \leq p(t)$, and $p(v_i) = p(t_i) = w_i/(2B)$ $(i = 1, \ldots, n)$. Then, the above reduces to minimizing:

$$p(u)\left[p(t) - \sum_{i:y_i=1} \frac{w_i}{2B}\left(1 - \sum_{i:y_i=1} \frac{w_i}{2B}\right)\right] \tag{29}$$

which, for given values of $p(u)$ and $p(t)$, is equivalent to maximize the function:

$$f(\mathbf{Y}) = \sum_{i:y_i=1} \frac{w_i}{2B}\left(1 - \sum_{i:y_i=1} \frac{w_i}{2B}\right)$$

whose value is $\leq 1/4$, with equality obtained only if $\sum_{i:y_i=1} w_i/(2B) = 1/2$.

Now, assume there is a polynomial time algorithm $\mathcal{ALG}$ that solves the IndArrange^P problem, i.e., it maximizes $f(\mathbf{Y})$. Let $S^{\mathbf{Y}}$ be the subset of those elements w_i in S such that the solution of IndArrange^P sets $y_i = 1$ for the corresponding tuples v_i and t_i. If $f(\mathbf{Y}) = 1/4$, then $W(S^{\mathbf{Y}}) = B = \alpha W(S)$, and α-Partition has a positive solution. Conversely, any value less than $1/4$ shows that α-Partition has no solution.

To complete the proof, we need to show that the transformation satisfies constraints on tuple probabilities. In particular, since within each group the sum of probabilities cannot exceed 1, we can set $1/(2\alpha) = 1 - p(t) - c$, where $0 < c < 1/2 - p(t)$ is a positive[11] constant whose precise value will be made clear below, so that the function to maximize becomes:

$$f(\mathbf{Y}) = \sum_{i:y_i=1} (1 - p(t) - c)\frac{w_i}{W(S)}\left(1 - \sum_{i:y_i=1} (1 - p(t) - c)\frac{w_i}{W(S)}\right)$$

We now prove that testing P-domination is co-NP-complete. The problem is in co-NP, since to show that $u \not\succ_p v$ it is sufficient to exhibit an order $\succcurlyeq$ that extends $\succ$ and verify that it is $\mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(v) > \mathsf{Ps}^{\succcurlyeq}_{\Psi,\mathcal{C}}(u)$. Since verifying that $\succcurlyeq$ is a linear extension of $\succ$ requires polynomial time, and by hypothesis this is also the case for computing probabilistic scores, P-domination belongs to co-NP.

[11] $c < 0$ does not guarantee that the sum of the probabilities of the tuples in $G(t)$ is ≤ 1.

For the hardness part we show that the answer to an instance of α-PARTITION is 'yes' if and only if the answer to the corresponding P-domination instance is 'no'. For what previously shown, this is the case iff $f(\mathbf{Y}) = 1/4$. From Equation 29 we immediately derive $p(t) < 1/4$. Since $f(\mathbf{Y}) > p(t)$ has to hold only when $f(\mathbf{Y}) = 1/4$, which in turn corresponds to $W(S^{\mathbf{Y}}) = \alpha W(S)$, it is sufficient to determine a lower bound to the value of $p(t)$ so that $W(S^{\mathbf{Y}}) \neq \alpha W(S)$ implies $f(\mathbf{Y}) \leq p(t)$.

Since the w_i's are integer, it is easy to show that if $W(S^{\mathbf{Y}}) \neq \alpha W(S)$, then $f(\mathbf{Y}) \leq 1/4 - 1/(2\alpha W(S))^2$. Thus, any value of $p(t)$ in the open interval $(1/4 - 1/(2\alpha W(S))^2, 1/4)$ will work equally well. At this point, having chosen $p(t)$, we consider the equation $c = 1 - p(t) - 1/(2\alpha) > 3/4 - 1/(2\alpha)$. Since it has to be $c > 0$, this requires $\alpha \geq 2/3$, which completes the proof.

As an example of how constants are set, let $W(S) = 275$ and $\alpha W(S) = 190$, so that $\alpha = 190/275 = 38/55$ and $1/(2\alpha) = 1 - p(t) - c = 55/76 \approx 0.723684$. The value of $p(t)$ has to be $> 1/4 - 1/(380)^2 \approx 0.249993075$, from which it results $c \approx 0.02632271$.

PROOF OF LEMMA 5.6.
As in the proof of Lemma 5.4, we partition all tuples in $\text{IND}(u, v)$ into three sets: Those in $G(u)$, those in $G(v)$, and those in other groups. We also observe that, since for testing P-domination only the *sign* of an INDARRANGE^P solution is relevant, it is possible to consider the *ratio* of probabilistic scores, which can be compactly written as:

$$\frac{\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{\Psi,\mathcal{C}}(u)}{\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{\Psi,\mathcal{C}}(v)} = \frac{p(u)}{p(v)} \frac{Z^{\succcurlyeq_{\mathbf{Y}}}_{G(v)}(u)}{Z^{\succcurlyeq_{\mathbf{Y}}}_{G(u)}(v)} \prod_{G \notin \{G(u), G(v)\}} \frac{Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(u)}{Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(v)} \tag{30}$$

where, say, $Z^{\succcurlyeq_{\mathbf{Y}}}_{G(v)}(u)$ is the contribution due to the tuples in $G(v)$ to the score of u, and similarly for other factors. Clearly, it is $u \succ_p v$ iff above ratio is ≥ 1 for all orders $\succcurlyeq_{\mathbf{Y}}$.

From Equation 20 it is apparent that, for any group $G \notin \{G(u), G(v)\}$, it is $Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(u) \geq Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(v)$, thus tuples in such groups should possibly follow v (i.e., yielding $y_i = 0$) so as to avoid further decreasing the ratio $Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(u)/Z^{\succcurlyeq_{\mathbf{Y}}}_{G}(v)$.

For tuples in $G(u)$ or in $G(v)$ arguments similar to those in the proof of Lemma 5.4 for the ER case apply here, thus one should verify at most $\mathcal{O}(2^g)$ arrangements, which proves the result. □

B. SEMANTICS BASED ON TUPLE SCORES

Can the results we have derived be generalized to arbitrary semantics that are not value-invariant? Such semantics, which we denote as V-PRS's, are those that also consider the tuple scores for ranking probabilistic tuples. Notice that, besides the expected score semantics, that we have already analyzed in Section 5.4.2, no other V-PRS semantics has been proposed. Therefore, what follows has a purely speculative interest.

We start by considering the set-monotonicity property, that for a PRS ensures that enlarging the set of tuples preceding u cannot increase the probabilistic score of u. The equivalent of set-monotonicity for a V-PRS Ψ is to demand that, in addition to being insensitive to the specific ordering of tuples preceding u (set-dependency), Ψ is also *score-monotone* (rather than just rank-monotone), that is, decreasing the value of $s(u)$ cannot increase the probabilistic score of u. Since now we consider scoring functions, rather than just linear orders, the latter is conveniently denoted as $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u)$ (i.e., with s replacing $\succcurlyeq$).

For checking P-domination one should now determine the minimum value of $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(v)$, and a (u, v)-*adversarial scoring function* is consequently defined as any scoring function for which this minimum is attained. Note that, unlike the case of PRS's, for which the search space has finite size (the number of linear extensions is

finite), for V-PRS's we have an infinite number of alternatives (the number of monotone scoring function is infinite).

When $u \succ v$, consider the INDARRANGE problem, that for a PRS requires to deciding how tuples in the set $\text{IND}(u, v)$ should be arranged so as to maximally favor v with respect to u. For a V-PRS we have a similar problem, but in the infinite space of the monotone scoring functions. To this end, we distinguish two relevant cases:

a. The probabilistic score of a tuple u depends on $s(u)$, $p(u)$ and on the probabilities of all and only those tuples t in $\text{Up}^{\succ}(u) = \{t \,|\, t \succ u\}$.
b. In addition to the above, $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u)$ also depends on the *scores* of the tuples in $\text{Up}^{\succ}(u)$.

For the scenario in (a), P-domination can be checked by setting $s(u) = s(v)$ and then solving as usual the INDARRANGE problem. This follows because $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u)$ is not influenced at all by the scores of the tuples in $\text{Up}^{\succ_{\mathbf{Y}}}(u)$. Therefore, $s(u) = s(v)$ has no side-effects on the probabilistic score of v.

As a simple example, consider the following hypothetical V-PRS:

$$\mathsf{Ps}^{S}_{V\_ER,\emptyset}(u) = s(u) \times \mathsf{Ps}^{\succcurlyeq_s}_{ER,\emptyset}(u)$$

where $\succcurlyeq_S$ is the linear order induced by $s()$. After setting $s(u) = s(v)$ it will be $u \not\succ_p v$ iff there exists an order $\succcurlyeq$ such that $\mathsf{Ps}^{\succcurlyeq}_{ER,\emptyset}(u) < \mathsf{Ps}^{\succcurlyeq}_{ER,\emptyset}(v)$, which amounts to solving the INDARRANGE problem for the $(ER, \emptyset)$ combination.

Scenario (b) is (much) more complex to analyze, and no immediate result seems to be available in the absence of additional hypotheses. Indeed, if the dependency of $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u)$ on the scores of the tuples in $\text{Up}^{\succ}(u)$ is an arbitrary one, each specific arrangement of tuples in $\text{IND}(u, v)$ should be considered, and for each of them the minimum of $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u) - \mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(v)$ evaluated. Note that each of these optimization subproblems has a number of variables (the numerical tuple scores) equal to the number of tuples preceding v in that arrangement, and the values of these variables are only constrained by $s()$ to be a monotone function (if $t_i \succ t_j$, then it has to be $s(t_i) \geq s(t_j)$). We argue that in case (b) a reasonable thing to demand is that $\mathsf{Ps}^{S}_{\Psi,\mathcal{C}}(u)$ is also monotone in the scores of all the tuples preceding u, but even in this case the problem remains a challenging one to solve.

Received June 2013; revised November 2013; accepted February 2014

Online Appendix to: Domination in the Probabilistic World: Computing Skylines for Arbitrary Correlations and Ranking Semantics

ILARIA BARTOLINI, PAOLO CIACCIA, and MARCO PATELLA, Università di Bologna

A. DERIVATION OF P-DOMINATION RULES

In this appendix we detail how the P-domination rules presented in Section 5 can be derived. We omit those derivations that are trivial to obtain, and those for the $PRF^e(\alpha)$ ranking semantics, that are very similar to the ones given for the $T1$ semantics.

A.1. Expected Rank – Independent Model

To prove that $p(u) \geq p(v)$ is a sufficient condition to derive $u \succ_p v$ when $u \succ v$, we start from Equation 9, that for independent tuples can be written as:

$$\begin{aligned}
&\mathsf{Ps}^{\succcurlyeq \mathbf{Y}}_{ER,\emptyset}(u) - \mathsf{Ps}^{\succcurlyeq \mathbf{Y}}_{ER,\emptyset}(v) \\
&\quad = p(v) \times \left(\sum_{t \succ v} p(t) + \sum_{t_i \in \textsc{Ind}(u,v)} y_i p(t_i) \right) \\
&\qquad - p(u) \times \left(\sum_{t \in \textsc{Btr}(u,v)} p(t) + \sum_{t \in \textsc{IndBtr}(u,v)} p(t) + \sum_{t_i \in \textsc{Ind}(u,v)} y_i p(t_i) \right) \\
&\qquad + (1 - p(v)) \times \sum_{t \neq v} p(t) - (1 - p(u)) \times \sum_{t \neq u} p(t) \\
&\quad = p(u) \times \left(1 + \sum_{t} p(t) - p(u) - \sum_{t \in \textsc{Btr}(u,v)} p(t) - \sum_{t \in \textsc{IndBtr}(u,v)} p(t) - \sum_{t_i \in \textsc{Ind}(u,v)} y_i p(t_i) \right) \\
&\qquad - p(v) \times \left(1 + \sum_{t} p(t) - p(v) - \sum_{t \succ v} p(t) - \sum_{t_i \in \textsc{Ind}(u,v)} y_i p(t_i) \right)
\end{aligned} \tag{31}$$

Since $u \succ v$, it is $\{t \mid t \succ v\} = \textsc{Btr}(u,v) \cup \textsc{IndBtr}(u,v) \cup \textsc{WrsBtr}(u,v) \cup \{u\}$, therefore $\sum_{t \succ v} p(t) > p(u) + \sum_{t \in \textsc{Btr}(u,v)} p(t) + \sum_{t \in \textsc{IndBtr}(u,v)} p(t)$. The assert is then proved since, by hypothesis, it is $p(u) \geq p(v)$, thus $\mathsf{Ps}^{\succcurlyeq \mathbf{Y}}_{ER,\emptyset}(u) \geq \mathsf{Ps}^{\succcurlyeq \mathbf{Y}}_{ER,\emptyset}(v)$ whatever the value of $\mathbf{Y}$ is.

On the other hand, when $p(v) > p(u)$, having $y_i = 1$ would increase the value of (31), thus it should be $y_i = 0, \forall i$, i.e., all tuples indifferent to both u and v should follow v.

DOI: http://dx.doi.org/10.1145/0000000.0000000

Consequently, the difference of probabilistic scores reduces to:

$$\begin{aligned}
&\mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{ER,\emptyset}(u) - \mathsf{Ps}^{\succcurlyeq_{\mathbf{Y}}}_{ER,\emptyset}(v) \\
&\quad = p(v) \times \sum_{t \succ v} p(t) - p(u) \times \left(\sum_{t \in \textsc{Btr}(u,v)} p(t) + \sum_{t \in \textsc{IndBtr}(u,v)} p(t) \right) \\
&\qquad + (1 - p(v)) \times \sum_{t \neq v} p(t) - (1 - p(u)) \times \sum_{t \neq u} p(t) \\
&\quad = N - \mathsf{Ps}^{+}_{ER,\emptyset}(v) - \left(N - \left(\mathsf{Ps}^{-}_{ER,\emptyset}(u) + p(u) \times \sum_{t \in \textsc{Ind}(u,v)} p(t) \right) \right)
\end{aligned}$$

from which the 2nd disjunct in Rule$2_{ER,\emptyset}$ follows.

A.2. Expected Rank – X-relation Model

Rule$1_{ER,\mathcal{C}_X}$ is derived by assuming that $\textsc{WrsBtr}(u,v) = \emptyset$, all tuples in $G(v)$ other than v dominate u, and all tuples in $G(u)$ other than u do not dominate v. Tuples preceding u are: (i) those that dominate u in groups other than $G(u)$ and $G(v)$ and (ii) all tuples in $G(v)$ other than v. Tuples preceding v are (i) and u.

Rule$2_{ER,\mathcal{C}_X}$ is derived as for the independent case, by considering that we now have to solve IndArrange.

A.3. Top-1 – Independent Model

When $u \succ v$, if $\mathsf{Ps}^{+}_{T1,\emptyset}(v) = 0$ we have that $u \succ_p v$, since it will always be $\mathsf{Ps}^{\succcurlyeq}_{T1,\emptyset}(u) \geq \mathsf{Ps}^{\succcurlyeq}_{T1,\emptyset}(v)$ for any $\succcurlyeq$ that extends $\succ$. Therefore, assume that $\mathsf{Ps}^{+}_{T1,\emptyset}(v) > 0$, i.e., no tuple t exists that dominates v and for which $p(t) = 1$ (thus, it is also $\mathsf{Ps}^{+}_{T1,\emptyset}(u) > 0$). Moreover, any tuple in $\textsc{Ind}(u,v)$ with a probability equal to 1 should be arranged behind v (otherwise both probabilistic scores would be 0), thus no tuple in $\textsc{Ind\_1}(u,v)$ will have probability 1. The factor

$$\prod_{\substack{t \in \textsc{Btr}(u,v) \\ \cup \textsc{IndBtr}(u,v)}} (1 - p(t)) \prod_{t_i \in \textsc{Ind}(u,v)} (1 - y_i p(t_i))$$

in Equation 18 has therefore a value greater than 0, regardless of the optimal choice of $\mathbf{Y}$. Thus, omitting such factor from Equation 18 does not influence the P-domination test, which reduces to checking whether:

$$p(u) - p(v)(1 - p(u)) \prod_{t \in \textsc{WrsBtr}(u,v)} (1 - p(t)) \geq 0$$

without the need of solving the IndArrange problem.

Rule$1_{T1,\emptyset}$ is obtained by postulating that $\textsc{WrsBtr}(u,v) = \emptyset$ in the above test.

A.4. Top-1 – X-relation Model

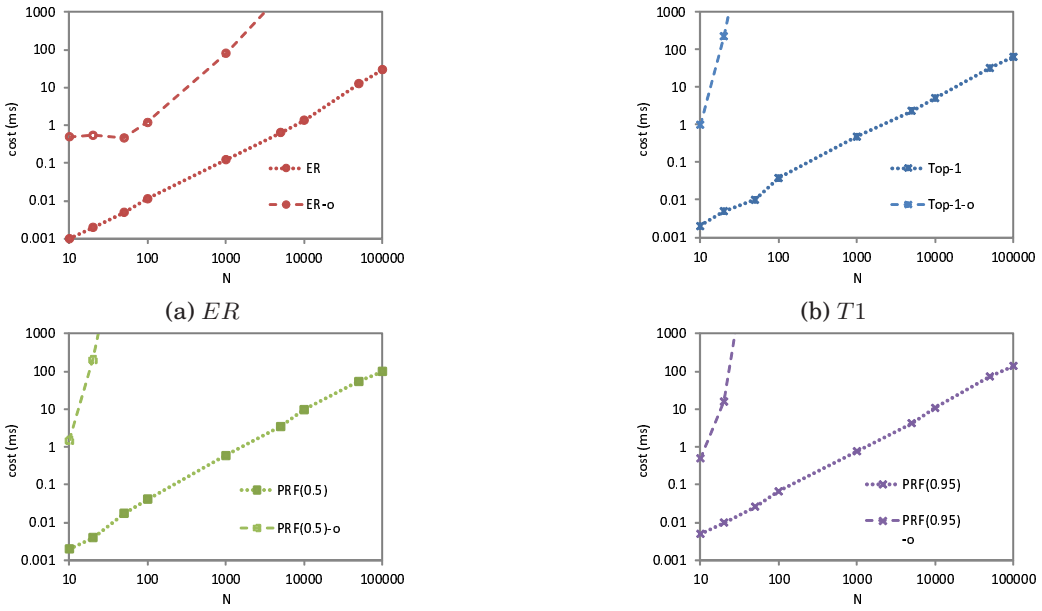
Rule$2_{T1,\mathcal{C}_X}$ directly derives from Equation 19, by considering that tuples preceding u in a (u,v)-adversarial order are those in $\textsc{Btr}(u,v) \cup \textsc{IndBtr}(u,v) \cup \textsc{Ind\_1}(u,v) = \{t \,|\, u \not\succ t\} \setminus \textsc{Ind\_0}(u,v)$, while those preceding v are $\textsc{Btr}(u,v) \cup \textsc{IndBtr}(u,v)\textsc{WrsBtr}(u,v) \cup \{u\} \cup \textsc{Ind\_1}(u,v) = \{t \,|\, t \succ v\} \cup \textsc{Ind\_1}(u,v)$.

Rule$1_{T1,\mathcal{C}_X}$ is obtained by postulating that, in the above test, it is $\textsc{WrsBtr}(u,v) = \emptyset$,, all tuples in $G(v)$ other than v dominate u, and all tuples in $G(u)$ other than u do not

dominate v. It follows that, for groups other than $G(u)$ and $G(v)$, tuples preceding u and v coincide and their contribution can be therefore neglected, this obtaining Rule1$_{T1,\mathcal{C}_X}$.

B. ADDITIONAL EXPERIMENTAL RESULTS

Figure 15 shows in a log-log scale the average time needed to perform a P-domination test under the x-relation model, when varying the number of tuples N. In each graph, the dotted line represents the tight integration scenario, while the dashed line corresponds to the loosely integrated one (dubbed "-o" in the legend). As also noted for Figure 5, the tight integration scenario outperforms the loose integration one by several orders of magnitude.



(a) ER (b) $T1$

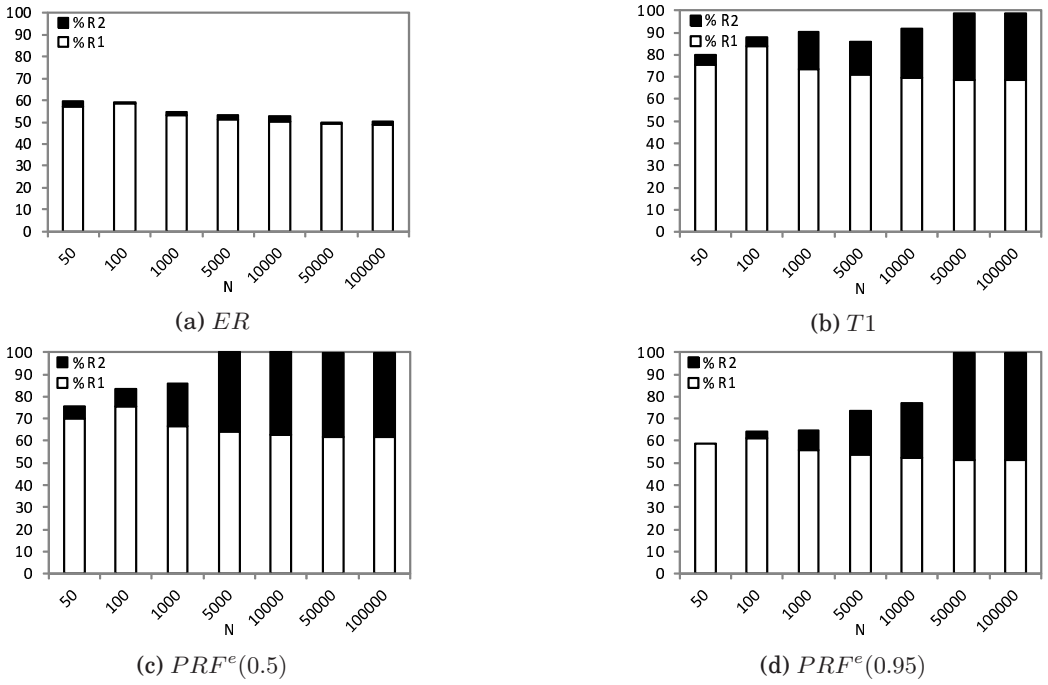
(c) $PRF^e(0.5)$ (d) $PRF^e(0.95)$

Fig. 15. Average cost for assessing P-domination between two tuples vs. cardinality. X-relations and different ranking semantics.

Figure 16 shows the chance of success of a P-domination test in the x-relation model. The graphs exhibit a trend similar to those obtained for independent data (Figure 10), although the chance of success of Rule1 is always lower than in the independent case, as pointed out in Section 5.

The skylines obtained for IIP real dataset (see Section 6.2) are shown in Figure 17. In displaying the tuples in the (distance, size) space, we have highlighted tuples in common with other skylines. Thus, for example, in Figure 17(c), representing the $(T1, \emptyset)$ and $(PRF^e(0.5), \emptyset)$ combinations, we show that, besides all the 23 tuples also appearing in the deterministic skyline and displayed as black crosses, other five tuples, depicted as red triangles, appear in the skyline. Comparing Figures 17(b) with Figure 17(a) we see that no tuples in the deterministic skyline are included in the skyline for the ER PRS.

Finally, Table V shows skyline computation costs for the different $(\Psi, \mathcal{C})$ combinations. For each combination, the table also reports the chance of success of Rule1, the number of tuples which are not P-dominated using Rule1, i.e., the tuples for which



(a) ER (b) $T1$

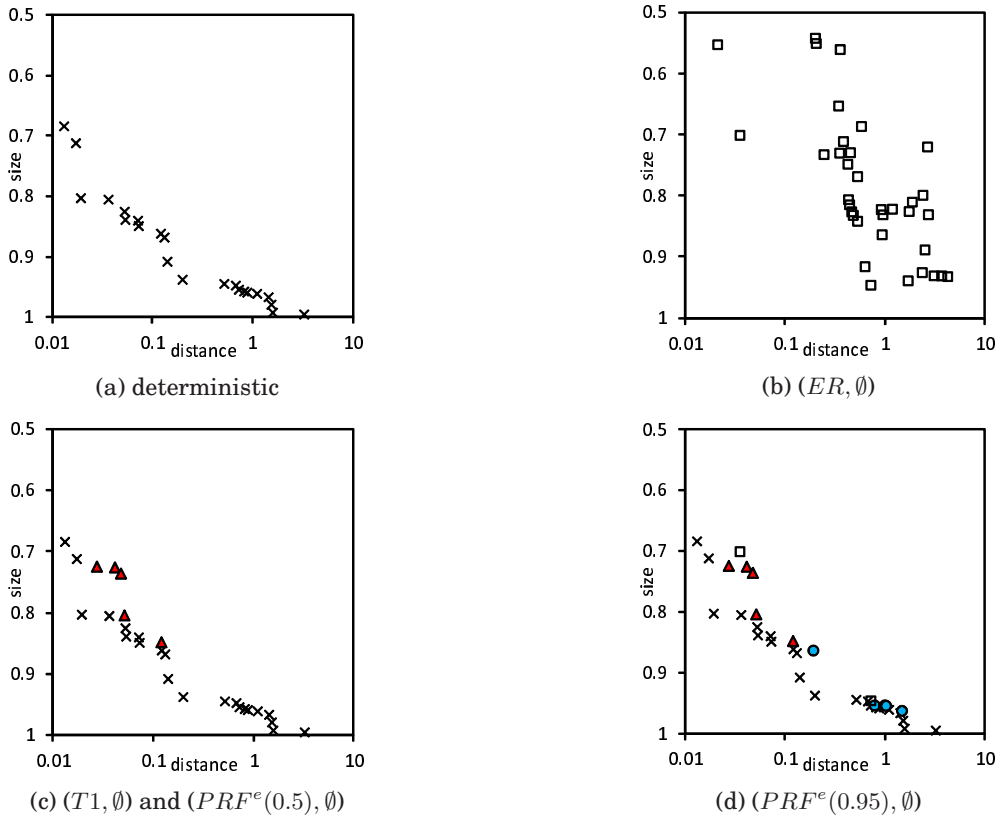
(c) $PRF^e(0.5)$ (d) $PRF^e(0.95)$

Fig. 16. Effectiveness of Rule1 (R1) and Rule2 (R2) on P-domination tests vs. dataset cardinality, x-relations.

bounds have to be computed and the second phase of Algorithm 2 performed, and the final size of the skyline.

Table V. Skyline cost for $(\Psi, \mathcal{C})$ combinations used in the experiments: residual tuples are those which are not P-dominated using Rule1.

$(\Psi, \mathcal{C})$	Skyline cost (msecs)	Effectiveness of Rule1	Number of residual tuples	$\lvert\text{SKY}(R^p)\rvert$
$(ER, \emptyset)$	82017	26.8%	92	35
$(ER, \mathcal{C}_X)$	85469	22.0%	102	34
$(ER, \mathcal{C}_m)$	187365	28.6%	92	35
$(T1, \emptyset)$	76325	47.3%	28	28
$(T1, C_X)$	88603	34.9%	40	32
$(PRF^e(0.5), \emptyset)$	77971	47.2%	28	28
$(PRF^e(0.5), \mathcal{C}_X)$	89442	33.3%	42	35
$(PRF^e(0.95), \emptyset)$	77239	36.8%	41	35
$(PRF^e(0.95), \mathcal{C}_X)$	115570	25.9%	58	37



(a) deterministic

(b) $(ER, \emptyset)$

(c) $(T1, \emptyset)$ and $(PRF^e(0.5), \emptyset)$

(d) $(PRF^e(0.95), \emptyset)$

Fig. 17. Skyline tuples for different $(\Psi, \mathcal{C})$ combinations.