Semantics and Evaluation of Top-k Queries in Probabilistic Databases*

Xi Zhang and Jan Chomicki

Department of Computer Science and Engineering
University at Buffalo, SUNY, U.S.A.
{xizhang,chomicki}@cse.buffalo.edu

Abstract. We study here fundamental issues involved in top-k query evaluation in probabilistic databases. We consider *simple* probabilistic databases in which probabilities are associated with individual tuples, and *general* probabilistic databases in which, additionally, exclusivity relationships between tuples can be represented. In contrast to other recent research in this area, we do not limit ourselves to injective scoring functions. We formulate three intuitive postulates that the semantics of top-k queries in probabilistic databases should satisfy, and introduce a new semantics, Global-Top$k$, that satisfies those postulates to a large degree. We also show how to evaluate queries under the Global-Top$k$ semantics. For simple databases we design dynamic-programming based algorithms, and for general databases we show polynomial-time reductions to the simple cases. For example, we demonstrate that for a fixed k the time complexity of top-k query evaluation is as low as linear, under the assumption that probabilistic databases are simple and scoring functions are injective.

1 Introduction

The study of incompleteness and uncertainty in databases has long been an interest of the database community [2–8]. Recently, this interest has been rekindled by an increasing demand for managing rich data, often incomplete and uncertain, emerging from scientific data management, sensor data management, data cleaning, information extraction etc. [9] focuses on query evaluation in traditional probabilistic databases; ULDB [10] supports uncertain data and data lineage in Trio [11]; MayBMS [12] uses the vertical World-Set representation of uncertain data [13]. The standard semantics adopted in most works is the *possible worlds* semantics [2, 6, 7, 10, 9, 13].

On the other hand, since the seminal papers of Fagin [14, 15], the top-k problem has been extensively studied in multimedia databases [16], middleware systems [17], data cleaning [18], core technology in relational databases [19, 20] etc. In the top-k problem, each tuple is given a *score*, and users are interested in k tuples with the highest scores.

More recently, the top-k problem has been studied in probabilistic databases [21, 22]. Those papers, however, are solving two essentially different top-k problems. Soliman et al. [21] assumes the existence of a scoring function to rank tuples. Probabilities

* Research partially supported by NSF grant IIS-0119186. An earlier version of some of the results in this paper was presented in [1].

provide information on how likely tuples will appear in the database. In contrast, in [22], the ranking criterion for top-k is the probability associated with each query answer. In many applications, it is necessary to deal with tuple probabilities and scores at the same time. Thus, in this paper, we use the model of [21]. Even in this model, different semantics for top-k queries are possible, so a part of the challenge is to define a reasonable semantics.

As a motivating example, let us consider the following graduate admission example.

Example 1. A graduate admission committee need to select two winners of a fellowship. They narrow the candidates down to the following short list:

Name	Overall Score	Prob. of Coming
Aidan	0.65	0.3
Bob	0.55	0.9
Chris	0.45	0.4

where the *overall score* is the normalized score of each candidate based on their qualifications, and the *probability of acceptance* is derived from historical statistics on candidates with similar qualifications and background.

The committee want to make offers to the best two candidates who will take the offer. This decision problem can be formulated as a top-k query over the above probabilistic relation, where $k = 2$.

In Example 1, each tuple is associated with an *event*, which is that the candidate will accept the offer. The probability of the event is shown next to each tuple. In this example, all the events of tuples are independent, and tuples are therefore said to be *independent*. Such a relation is said to be *simple*. In contrast, Example 2 illustrates a more general case.

Example 2. In a sensor network deployed in a habitat, each sensor reading comes with a confidence value *Prob*, which is the probability that the reading is valid. The following table shows the temperature sensor readings at a given sampling time. These data are from two sensors, Sensor 1 and Sensor 2, which correspond to two *parts* of the relation, marked C_1 and C_2 respectively. Each sensor has only one *true* reading at a given time, therefore tuples from the same part of the relation correspond to exclusive events.

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

Our question is:

“What’s the temperature of the warmest spot?”

The question can be formulated as a top-k query, where $k = 1$, over a probabilistic relation containing the above data. The scoring function is the temperature. However, we must take into consideration that the tuples in each part $C_i, i = 1, 2$, are exclusive.

Our contributions in this paper are the following:

- We formulate three intuitive semantic postulates and use them to analyze and compare different top-k semantics in probabilistic databases (Section 3.1);
- We propose a new semantics for top-k queries in probabilistic databases, called Global-Top$k$, which satisfies the above postulates to a large degree (Section 3.2);
- We exhibit efficient algorithms for evaluating top-k queries under the Global-Top$k$ semantics in *simple* probabilistic databases (Section 4.1) and general probabilistic databases, under injective scoring functions (Section 4.3).
- We generalize Global-Top$k$ semantics to general scoring functions, where ties are allowed, by introducing the notion of *allocation policy*. We propose dynamic programming based algorithms for query evaluation under the *Equal* allocation policy (Section 5).

2 Background

2.1 Probabilistic Relations

To simplify the discussion in this paper, we assume that a probabilistic database contains a single *probabilistic relation*. We refer to a traditional database relation as a *deterministic relation*. A deterministic relation R is a set of tuples. A *partition* $\mathcal{C}$ of R is a collection of non-empty subsets of R such that every tuple belongs to one and only one of the subsets. That is, $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ such that $C_1 \cup C_2 \cup \ldots \cup C_m = R$ and $C_i \cap C_j = \emptyset, 1 \leq i \neq j \leq m$. Each subset $C_i, i = 1, 2, \ldots, m$ is a *part* of the partition $\mathcal{C}$. A *probabilistic relation* R^p has three components, a *support (deterministic) relation* R, a probability function p and a partition $\mathcal{C}$ of the support relation R. The probability function p maps every tuple in R to a probability value in $(0, 1]$. The partition $\mathcal{C}$ divides R into subsets such that the tuples within each subset are exclusive and therefore their probabilities sum up to at most 1. In the graphical presentation of R, we use horizontal lines to separate tuples from different parts.

Definition 1 (Probabilistic Relation). *A probabilistic relation R^p is a triplet $\langle R, p, \mathcal{C} \rangle$, where R is a support deterministic relation, p is a probability function $p : R \mapsto (0, 1]$ and $\mathcal{C}$ is a partition of R such that $\forall C_i \in \mathcal{C}, \sum_{t \in C_i} p(t) \leq 1$.*

In addition, we make the assumption that tuples from different parts of of $\mathcal{C}$ are independent, and tuples within the same part are exclusive. Definition 1 is equivalent to the model used in Soliman et al. [21] with exclusive tuple generation rules. Ré et al. [22] proposes a more general model, however only a restricted model equivalent to Definition 1 is used in top-k query evaluation.

Example 2 shows an example of a probabilistic relation whose partition has two parts. Generally, each part corresponds to a real world entity, in this case, a sensor. Since there is only one true state of an entity, tuples from the same part are exclusive. Moreover, the probabilities of all possible states of an entity sum up to at most 1. In Example 2, the sum of probabilities of tuples from Sensor 1 is 1, while that from Sensor 2 is 0.7. This can happen for various reasons. In the above example, we might encounter a physical difficulty in collecting the sensor data, and end up with partial data.

Definition 2 (Simple Probabilistic Relation). *A probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ is simple iff the partition $\mathcal{C}$ contains only singleton sets.*

The probabilistic relation in Example 1 is *simple* (individual parts not illustrated). Note that in this case, $|R| = |\mathcal{C}|$.

We adopt the well-known *possible worlds* semantics for probabilistic relations [2, 6, 7, 10, 9, 13].

Definition 3 (Possible World). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a deterministic relation W is a* possible world *of R^p iff*

1. *W is a subset of the support relation, i.e. $W \subseteq R$;*
2. *For every part C_i in the partition $\mathcal{C}$, at most one tuple from C_i is in W, i.e. $\forall C_i \in \mathcal{C}, |C_i \cap W| \leq 1$;*
3. *The probability of W (defined by Equation 1) is positive, i.e. $Pr(W) > 0$.*

$$Pr(W) = \prod_{t \in W} p(t) \prod_{C_i \in \mathcal{C}'} (1 - \sum_{t \in C_i} p(t)) \tag{1}$$

where $\mathcal{C}' = \{C_i \in \mathcal{C} | W \cap C_i = \emptyset\}$.

Denote by $pwd(R^p)$ *the set of all possible worlds* of R^p.

2.2 Total order v.s. Weak order

A binary relation $\succ$ is

– irreflexive: $\forall x.\ x \not\succ x$,
– asymmetric:$\forall x, y.\ x \succ y \Rightarrow y \not\succ x$,
– transitive: $\forall x, y, z.\ (x \succ y \wedge y \succ z) \Rightarrow x \succ z$,
– negatively transitive: $\forall x, y, z.\ (x \not\succ y \wedge y \not\succ z) \Rightarrow x \not\succ z$,
– connected: $\forall x, y.\ x \succ y \vee y \succ x \vee x = y$.

A *strict partial order* is an irreflexive, transitive (and thus symmetric) binary relation. A *weak order* is a negatively transitive strict partial order. A *total order* is a connected strict partial order.

2.3 Scoring function

A *scoring function over a deterministic relation* R is a function from R to real numbers, i.e. $s : R \mapsto \mathbb{R}$. The function s induces a *preference relation* $\succ_s$ and an *indifference relation* $\sim_s$ on R. For any two distinct tuples t_i and t_j from R,

$$t_i \succ_s t_j \text{ iff } s(t_i) > s(t_j);$$
$$t_i \sim_s t_j \text{ iff } s(t_i) = s(t_j).$$

A *scoring function over a probabilistic relation* $R^p = \langle R, p, \mathcal{C} \rangle$ is a scoring function s over its support relation R. In general, a scoring function establishes a *weak order* over R, where tuples from R can tie in score. However, when the scoring function s is *injective*, $\succ_s$ is a *total order*. In such a case, no two tuples tie in score.

2.4 Top-k Queries

Definition 4 (Top-k Answer Set over Deterministic Relation). *Given a deterministic relation R, a non-negative integer k and a scoring function s over R, a top-k answer in R under s is a set T of tuples such that*

1. $T \subseteq R$;
s 2. If $|R| < k$, $T = R$, otherwise $|T| = k$;
3. $\forall t \in T\ \forall t' \in R - T.\ t \succ_s t'$ or $t \sim_s t'$.

According to Definition 4, given k and s, there can be more than one top-k answer set in a deterministic relation R. The evaluation of a top-k query over R returns one of them nondeterministically, say S. However, if the scoring function s is injective, S is unique, denoted by $top_{k,s}(R)$.

3 Semantics of Top-k Queries

In the following two sections, we restrict our discussion to *injective* scoring functions. We will discuss the generalization to general scoring functions in Section 5.

3.1 Semantic Postulates for Top-*k* Answers

Probability opens the gate for various possible semantics for top-k queries. As the semantics of a probabilistic relation involves a set of worlds, it is to be expected that there may be more than one top-k answer, even under an injective scoring function. The answer to a top-k query over a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ should clearly be a set of tuples from its support relation R. We formulate below three desirable *postulates*, which serve as a benchmark to compare different semantics.

In the following discussion, denote by $Ans_{k,s}(R^p)$ the collection of all top-k answer sets of R^p under the function s.

Postulates

- **Static Postulates**
  1. *Exact k*: When R^p is sufficiently large ($|\mathcal{C}| \geq k$), the cardinality of every top-k set S is exactly k;
  $$|\mathcal{C}| \geq k \Rightarrow [\forall S \in Ans_{k,s}(R^p).\ |S| = k].$$
  2. *Faithfulness*: For every top-k set S and any two tuples $t_1, t_2 \in R$, if both the score and the probability of t_1 are higher than those of t_2 and $t_2 \in S$, then $t_1 \in S$;
  $$\forall S \in Ans_{k,s}(R^p)\ \forall t_1, t_2 \in R.\ s(t_1) > s(t_2) \wedge p(t_1) > p(t_2) \wedge t_2 \in S \Rightarrow t_1 \in S.$$
- **Dynamic Postulate**
  $\cup\, Ans_{k,s}(R^p)$ denotes the union of all top-k answer sets of $R^p = \langle R, p, \mathcal{C} \rangle$ under the function s. For any $t \in R$,
  $$t \text{ is a } \textit{winner} \text{ iff } t \in \cup\, Ans_{k,s}(R^p)$$
  $$t \text{ is a } \textit{loser} \text{ iff } t \in R - \cup\, Ans_{k,s}(R^p)$$

3. *Stability*:
   - Raising the score/probability of a winner will not turn it into a loser;
     (a) If a scoring function s' is such that $s'(t) > s(t)$ and for every $t' \in R - \{t\}$, $s'(t) = s(t)$, then
     $$t \in \cup Ans_{k,s}(R^p) \Rightarrow t \in \cup Ans_{k,s'}(R^p).$$
     (b) If a probability function p' is such that $p'(t) > p(t)$ and for every $t' \in R - \{t\}$, $p'(t) = p(t)$, then
     $$t \in \cup Ans_{k,s}(R^p) \Rightarrow t \in \cup Ans_{k,s}((R^p)'),$$
     where $(R^p)' = \langle R, p', \mathcal{C} \rangle$.
   - Lowering the score/probability of a loser will not turn it into a winner.
     (a) If a scoring function s' is such that $s'(t) < s(t)$ and for every $t' \in R - \{t\}$, $s'(t) = s(t)$, then
     $$t \in R - \cup Ans_{k,s}(R^p) \Rightarrow t \in R - \cup Ans_{k,s'}(R^p).$$
     (b) If a probability function p' is such that $p'(t) < p(t)$ and for every $t' \in R - \{t\}$, $p'(t) = p(t)$, then
     $$t \in R - \cup Ans_{k,s}(R^p) \Rightarrow t \in R - \cup Ans_{k,s}((R^p)'),$$
     where $(R^p)' = \langle R, p', \mathcal{C} \rangle$.

All of those postulates reflect basic intuitions about top-k answers.

Exact k expresses user expectations about the size of the result. Typically, a user issues a top-k query in order to restrict the size of the result and get a subset of cardinality k (cf. Example 1). Therefore, k is a crucial parameter specified by the user that should be complied with.

Faithfulness reflects the significance of score and probability in a static environment. It plays an important role in designing efficient query evalution algorithms. The satisfaction of *Faithfulness* allows the application of a set of pruning techniques based on *monotonicity*.

Stability reflects the significance of score and probability in a dynamic environment. In a dynamic world, it is common that user might update score/probability on-the-fly. *Stability* requires that the consequences of such changes should not be counterintuitive.

3.2 Global-Top*k* Semantics

We propose here a new top-k answer semantics in probabilistic relations, namely **Global-Top**k, which satisfies the postulates formulated in Section 3.1 to a large degree:

- **Global-Top**k: return k highest-ranked tuples according to their probability of being in the top-k answers in possible worlds.

Considering a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ under an injective scoring function s, any $W \in pwd(R^p)$ has a unique top-k answer set $top_{k,s}(W)$. Each tuple from the support relation R can be in the top-k answer (in the sense of Definition 4) in zero, one or more possible worlds of R^p. Therefore, the sum of the probabilities of those possible worlds provides a global ranking criterion.

Definition 5 (Global-Top$k$ Probability). *Assume a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a non-negative integer k and an injective scoring function s over R^p. For any tuple t in R, the Global-Top$k$ probability of t, denoted by $P_{k,s}^{R^p}(t)$, is the sum of the probabilities of all possible worlds of R^p whose top-k answer contains t.*

$$P_{k,s}^{R^p}(t) = \sum_{\substack{W \in pwd(R^p) \\ t \in top_{k,s}(W)}} Pr(W). \tag{2}$$

For simplicity, we skip the superscript in $P_{k,s}^{R^p}(t)$, i.e. $P_{k,s}(t)$, when the context is unambiguous.

Definition 6 (Global-Top$k$ Answer Set over Probabilistic Relation). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a non-negative integer k and an injective scoring function s over R^p, a top-k answer in R^p under s is a set T of tuples such that*

1. $T \subseteq R$;
2. If $|R| < k$, $T = R$, otherwise $|T| = k$;
3. $\forall t \in T, \forall t' \in R - T, P_{k,s}(t) \geq P_{k,s}(t')$.

Notice the similarity between Definition 6 and Definition 4. In fact, the probabilistic version only changes the last condition, which restates the preferred relationship between two tuples by taking probability into account. This semantics preserves the nondeterministic nature of Definition 4. For example, if two tuples are of the same Global-Top$k$ probability, and there are $k-1$ tuples with higher Global-Top$k$ probability, Definition 4 allows one of the two tuples to be added to the top-k answer nondeterministically. Example 3 gives an example of the Global-Top$k$ semantics.

Example 3. Consider the top-2 query in Example 1. Clearly, the scoring function here is the *Overall Score* function. The following table shows all the possible worlds and their probabilities. For each world, the names of the people in the top-2 answer set of that world are underlined.

Possible World	Prob
$W_1 = \emptyset$	0.042
$W_2 = \{\underline{Aidan}\}$	0.018
$W_3 = \{\underline{Bob}\}$	0.378
$W_4 = \{\underline{Chris}\}$	0.028
$W_5 = \{\underline{Aidan}, \underline{Bob}\}$	0.162
$W_6 = \{\underline{Aidan}, \underline{Chris}\}$	0.012
$W_7 = \{\underline{Bob}, \underline{Chris}\}$	0.252
$W_8 = \{\underline{Aidan}, \underline{Bob}, Chris\}$	0.108

Chris is in the top-2 answer of W_4, W_6, W_7, so the top-2 probability of Chris is $0.028 + 0.012 + 0.252 = 0.292$. Similarly, the top-2 probability of Aidan and Bob are 0.9 and 0.3 respectively. $0.9 > 0.3 > 0.292$, therefore Global-Top$k$ will return $\{Aidan, Bob\}$.

Note that top-k answer sets may be of cardinality less than k for some possible worlds. We refer to such possible worlds as *small* worlds. In Example 3, $W_{1...4}$ are all small worlds.

3.3 Other Semantics

Soliman et al. [21] proposes two semantics for top-k queries in probabilistic relations.

- *U-Top$k$*: return the most probable top-k answer set that belongs to possible world(s);
- *U-kRanks*: for $i = 1, 2, \ldots, k$, return the most probable i^{th}-ranked tuples across all possible worlds.

Hua et al. [23] independently proposes PT-k, a semantics based on Global-Top$k$ probability as well. PT-k takes an additional parameter: probability threshold $p_\tau \in (0, 1]$.

- *PT-k*: return every tuple whose probability of being in the top-k answers in possible worlds is at least p_τ.

Example 4. Continuing Example 3, under U-Top$k$ semantics, the probability of top-2 answer set $\{Bob\}$ is 0.378, and that of $\{Aidan, Bob\}$ is $0.162 + 0.108 = 0.27$. Therefore, $\{Bob\}$ is more probable than $\{Aidan, Bob\}$ under U-Top$k$. In fact, $\{Bob\}$ is the most probable top-2 answer set in this case, and will be returned by U-Top$k$.

Under U-kRanks semantics, Aidan is in 1^{st} place in the top-2 answer of W_2, W_5, W_6, W_8, therefore the probability of Aidan being in 1^{st} place in the top-2 answers in possible worlds is $0.018 + 0.162 + 0.012 + 0.108 = 0.3$. However, Aidan is not in 2^{nd} place in the top-2 answer of any possible world, therefore the probability of Aidan being in 2^{nd} place is 0. In fact, we can construct the following table.

	Aidan	Bob	Chris
Rank 1	0.3	$\underline{0.63}$	0.028
Rank 2	0	$\underline{0.27}$	0.264

U-kRanks selects the tuple with the highest probability at each rank (underlined) and takes the union of them. In this example, Bob wins at both Rank 1 and Rank 2. Thus, the top-2 answer returned by U-kRanks is $\{Bob\}$.

PT-k returns every tuple with Global-Top$k$ probability above the user specified threshold p_τ, therefore the answer depends on p_τ. Say $p_\tau = 0.6$, then PT-k return $\{Aidan\}$, as it is the only tuple with Global-Top$k$ probability at least 0.6.

The postulates introduced in Section 3.1 lay the ground for comparing different semantics. In Table 1, a single "✓" (resp. "×") indicates that postulate is (resp. is not) satisfied under that semantics. "✓/×" indicates that, the postulate is satisfied by that semantics in *simple* probabilistic relations, but not in the general case.

Semantics	Exact k	Faithfulness	Stability
Global-Top$k$	✓	✓/×	✓
PT-k	×	✓/×	✓
U-Top$k$	×	✓/×	✓
U-kRanks	×	×	×

Table 1. Postulate Satisfaction for Different Semantics

For *Exact k*, Global-Top$k$ is the only semantics that satisfies this postulate. Example 4 illustrates the case where U-Top$k$, U-kRanks and PT-k violate this postulate. It is not satisfied by U-Top$k$ because a *small* possible world with high probability could dominate other worlds. In that case, the dominating possible world might not have enough tuples. It is also violated by U-kRanks because a single tuple can win at multiple ranks in U-kRanks. In PT-k, if the threshold parameter p_τ is set too high, then less than k tuples will be returned (as in Example 4). As p_τ decreases, PT-k return more tuples. In the extreme case when p_τ approaches 0, any tuple with a positive Global-Top$k$ probability will be returned.

For *Faithfulness*, Global-Top$k$ violates it when exclusion rules lead to a highly restricted distribution of possible worlds, and are combined with an unfavorable scoring function. PT-k violates *Faithfulness* for the same reason. U-Top$k$ violates *Faithfulness* since it requires all tuples in a top-k answer set to be compatible, this postulate can be violated when a high-score/probability tuple could be dragged down arbitrarily by its compatible tuples if they are not very likely to appear. U-kRanks violates both *Faithfulness* and *Stability*. Under U-kRanks, instead of a set, a top-k answer is an ordered vector, where ranks are significant. A change in a tuple's probability/score might have unpredictable consequence on ranks, therefore those two postulates are not guaranteed to hold.

Faithfulness is a postulate which can lead to significant pruning in practice. Even though it is not fully satisfied by any of the four semantics, some degree of satisfaction is still desirable, as it will help us find pruning rules. For example, our optimization in Section 4.2 explores the *Faithfulness* of Global-Top$k$ in simple probabilistic databases. Another example is that one of the pruning techniques in [23] explores the *Faithfulness* of exclusive tuples in general probabilistic databases as well.

Proofs of the results in Table 1 are in Appendix.

4 Query Evaluation under Global-Top$k$

4.1 Simple Probabilistic Relations

We first consider a *simple* probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ under an injective scoring function s.

Proposition 1. *Given a simple probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ and an injective scoring function s over R^p, if $R = \{t_1, t_2, \ldots, t_n\}$ and $t_1 \succ_s t_2 \succ_s \ldots \succ_s t_n$, the following recursion on Global-Top$k$ queries holds:*

$$q(k,i)=\begin{cases}0 & k=0\\ p(t_i) & 1\leq i\leq k\\ (q(k,i-1)\dfrac{\bar{p}(t_{i-1})}{p(t_{i-1})}+q(k-1,i-1))p(t_i) & \textit{otherwise}\end{cases} \tag{3}$$

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

Proof. See Appendix.

Notice that Equation 3 involves probabilities only, while the scores are used to determine the order of computation.

Example 5. Consider a simple probabilistic relation $R^p=\langle R,p,\mathcal{C}\rangle$, where $R=\{t_1, t_2, t_3, t_4\}$, $p(t_i)=p_i, 1\leq i\leq 4, \mathcal{C}=\{\{t_1\},\{t_2\},\{t_3\},\{t_4\}\}$ and an injective scoring function s such that $t_1\succ_s t_2\succ_s t_3\succ_s t_4$. The following table shows the Global-Top$k$ probability of t_i, where $0\leq k\leq 2$.

k	t_1	t_2	t_3	t_4
0	0	0	0	0
1	p_1	$\bar{p}_1p_2$	$\bar{p}_1\bar{p}_2p_3$	$\bar{p}_1\bar{p}_2\bar{p}_3p_4$
2	$\mathbf{p_1}$	$\mathbf{p_2}$	$(\mathbf{\bar{p}_2+\bar{p}_1p_2})\mathbf{p_3}$	$((\mathbf{\bar{p}_2+\bar{p}_1p_2})\mathbf{\bar{p}_3}$ $+\mathbf{\bar{p}_1\bar{p}_2p_3})\mathbf{p_4}$

Row 2 (bold) is each t_i's Global-Top2 probability. Now, if we are interested in top-2 answer in R^p, we only need to pick the two tuples with the highest value in Row 2.

Theorem 1 (Correctness of Algorithm 1). *Given a simple probabilistic relation* $R^p=\langle R,p,\mathcal{C}\rangle$*, a non-negative integer* k *and an injective scoring function* s*, Algorithm 1 correctly computes a Global-Top*k *answer set of* R^p *under the scoring function* s.

Proof. Algorithm 1 maintains a priority queue to select the k tuples with the highest Global-Top$k$ value. Notice that the nondeterminism is reflected in Line 6 as the algorithm for maintaining the priority queue in the presence of tying elements. As long as Line 2 in Algorithm 1 correctly computes the Global-Top$k$ probability of each tuple in R, Algorithm 1 returns a valid Global-Top$k$ answer set. By Proposition 1, Algorithm 2 correctly computes the Global-Top$k$ probability of tuples in R.

Algorithm 1 is a one-pass computation on the probabilistic relation, which can be easily implemented even if secondary storage is used. The overhead is the initial sorting cost (not shown in Algorithm 1), which would be amortized by the workload of consecutive top-k queries.

Algorithm 2 takes $O(kn)$ to compute the DP table. In addition, Algorithm 1 uses a priority queue to maintain the k highest values, which takes $O(n\log k)$. Altogether, Algorithm 1 takes $O(kn)$.

4.2 Threshold Algorithm Optimization

Fagin [15] proposes *Threshold Algorithm (TA)* for processing top-k queries in a middleware scenario. In a middleware system, an *object* has m attributes. For each attribute,

Algorithm 1 (Ind_Topk) Evaluate Global-Top$k$ Queries in a Simple Probabilistic Relation under an Injective Scoring Function

Require: $R^p = \langle R, p, \mathcal{C} \rangle, k$
Ensure: tuples in R are sorted in the decreasing order based on the scoring function s

Initialize a fixed cardinality $(k+1)$ priority queue Ans of $\langle t, prob \rangle$ pairs, which compares pairs on $prob$, i.e. the Global-Top$k$ probability of t;
Calculate Global-Top$k$ probabilities using Algorithm 2, i.e.

$$q(0 \ldots k, 1 \ldots |R|) = \text{Ind\_Topk\_Sub}(R^p, k);$$

for $i = 1$ to $|R|$ **do**
  Add $\langle t_i, q(k, i) \rangle$ to Ans;
  if $|Ans| > k$ **then**
    remove the pair with the smallest $prob$ value from Ans;
  end if
end for
return $\{t_i | \langle t_i, q(k, i) \rangle \in Ans\}$;

Algorithm 2 (Ind_Topk_Sub) Compute Global-Top$k$ Probabilities in a Simple Probabilistic Relation under an Injective Scoring Function

Require: $R^p = \langle R, p, \mathcal{C} \rangle, k$
Ensure: tuples in R are sorted in the decreasing order based on s

$q(0, 1) = 0$;
for $k' = 1$ to k **do**
  $q(k', 1) = p(t_1)$;
end for
for $i = 2$ to $|R|$ **do**
  for $k' = 0$ to k **do**
    if $k' = 0$ **then**
      $q(k', i) = 0$;
    else
      $q(k', i) = p(t_i)(q(k', i-1)\frac{\bar{p}(t_{i-1})}{p(t_{i-1})} + q(k'-1, i-1))$;
    end if
  end for
end for
return $q(0 \ldots k, 1 \ldots |R|)$;

there is a sorted list ranking objects in the decreasing order of its score on that attribute. An *aggregation function* f combines the individual attribute scores x_i, i=$1, 2, \ldots, m$ to obtain the overall object score $f(x_1, x_2, \ldots, x_m)$. An aggregation function is *monotonic* iff $f(x_1, x_2, \ldots, x_m) \leq f(x'_1, x'_2, \ldots, x'_m)$ whenever $x_i \leq x'_i$ for every i. Fagin [15] shows that TA is cost-optimal in finding the top-k objects in such a system.

TA is guaranteed to work as long as the aggregation function is monotonic. For a simple probabilistic relation, if we regard *score* and *probability* as two special attributes, Global-Top$k$ probability $P_{k,s}$ is an aggregation function of *score* and *probability*. The *Faithfulness* postulate in Section 3.1 implies the monotonicity of Global-Top$k$ probability. Consequently, assuming that we have an index on probability as well, we can guide the dynamic programming (DP) in Algorithm 2 by TA. Now, instead of computing all kn entries for DP, where $n = |R|$, the algorithm can be stopped as early as possible. A subtlety is that Global-Top$k$ probability $P_{k,s}$ is *only* well-defined for $t \in R$, unlike in [15], where an aggregation function is well-defined over the domain of all possible attribute values. Therefore, compared to the original TA, we need to achieve the same behavior without referring to virtual tuples which are not in R.

U-Top$k$ satisfies *Faithfulness* in simple probabilistic relations. An adaption of the TA algorithm in this case is available in [21]. TA is not applicable to U-kRanks. Even though we can define an aggregation function per $rank$, $rank = 1, 2, \ldots, k$, for tuples under U-kRanks, the violation of *Faithfulness* in Table 1 suggests a violation of monotonicity of those k aggregation functions. PT-k computes Global-Top$k$ probability as well, and is therefore a natural candidate for TA in simple probabilistic relations.

Denote T and P for the list of tuples in the decreasing order of score and probability respectively. Following the convention in [15], $\underline{t}$ and $\underline{p}$ are the last value seen in T and P respectively.

Algorithm 1' (TA_Ind_Topk)

(1) Go down T list, and fill in entries in the DP table. Specifically, for $\underline{t} = t_j$, compute the entries in the j^{th} column up to the k^{th} row. Add t_j to the top-k answer set Ans, if any of the following conditions holds:
   (a) Ans has less than k tuples, i.e. $|Ans| < k$;
   (b) The Global-Top$k$ probability of t_j, i.e. $q(k, j)$, is greater than the lower bound of Ans, i.e. LB_{Ans}, where $LB_{Ans} = \min_{t_i \in Ans} q(k, i)$.

   In the second case, we also need to drop the tuple with the lowest Global-Top$k$ probability in order to preserve the cardinality of Ans.

(2) After we have seen at least k tuples in T, we go down P list to find the first p whose tuple t has not been seen. Let $\underline{p} = p$, and we can use $\underline{p}$ to estimate the *threshold*, i.e. upper bound (UP) of the Global-Top$k$ probability of any unseen tuple. Assume $\underline{t} = t_i$,

$$UP = (q(k, i)\frac{\bar{p}(t_i)}{p(t_i)} + q(k-1, i))\underline{p}.$$

(3) If $UP > LB_{Ans}$, we can expect Ans will be updated in the future, so go back to (1). Otherwise, we can safely stop and report Ans.

Theorem 2 (Correctness of Algorithm 1'). *Given a simple probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, a non-negative integer k and an injective scoring function s over R^p, the above TA-based algorithm correctly find a top-k answer under Global-Top$k$ semantics.*

Proof. See Appendix.

The optimization above aims at an early stop. Bruno et al. [24] carries out an extensive experimental study on the effectiveness of applying TA in RDMBS. They consider various aspects of query processing. One of their conclusions is that if at least one of the indices available for the attributes[1] is a *covering index*, that is, it is defined over all other attributes and we can get the values of all other attributes directly without performing a primary index lookup, then the improvement by TA can be up to two orders of magnitude. The cost of building a useful set of indices once would be amortized by a large number of top-k queries that subsequently benefit form such indices. Even in the lack of covering indices, if the data is highly correlated, in our case, that means high-score tuples having high probabilities, TA would still be effective.

4.3 Arbitrary Probabilistic Relations

Induced Event Relation In the general case of probabilistic relation, each part of the partition $\mathcal{C}$ can contain more than one tuple. The crucial *independence* assumption in Algorithm 1 no longer holds. However, even though tuples in one part of the partition $\mathcal{C}$ are not independent, tuples in different parts are. In the following definition, we assume an identifier function id. For any tuple t, $id(t)$ identifies the part where t belongs.

Definition 7 (Induced Event Relation). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, an injective scoring function s over R^p and a tuple $t \in C_{id(t)} \in \mathcal{C}$, the event relation induced by t, denoted by $E^p = \langle E, p^E, \mathcal{C}^E\rangle$, is a probabilistic relation whose support relation E has only one attribute, $Event$. The relation E and the probability function p^E are defined by the following two generation rules:*

- *Rule 1:* $t_{e_t} \in E$ *and* $p^E(t_{e_t}) = p(t)$;
- *Rule 2:* $\forall C_i \in \mathcal{C} \wedge C_i \neq C_{id(t)}$.

$$(\exists t' \in C_i \wedge t' \succ_s t) \Rightarrow (t_{e_{C_i}} \in E) \text{ and } p^E(t_{e_{C_i}}) = \sum_{\substack{t' \in C_i \\ t' \succ_s t}} p(t').$$

No other tuples belong to E. The partition $\mathcal{C}^E$ is defined as the collection of singleton subsets of E.

Except for one special tuple generated by *Rule 1*, each tuple in the induced event relation (generated by *Rule 2*) represents an event e_{C_i} associated with a part $C_i \in \mathcal{C}$. Given the tuple t, the *event* e_{C_i} is defined as "some tuple from the part C_i has the score higher than the score of t". The probability of this event, denoted by $p(t_{e_{C_i}})$, is the probability that e_{C_i} occurs.

The role of the special tuple t_{e_t} and its probability $p(t)$ will become clear in Proposition 3. Let us first look at an example of an induced event relation.

[1] Probability is typically supported as a special attribute in DBMS.

Example 6. Given R^p as in Example 2, we would like to construct the induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ for tuple t=(Temp: 15) from C_2. By Rule 1, we have $t_{e_t} \in E$, $p^E(t_{e_t}) = 0.6$. By Rule 2, since $t \in C_2$, we have $t_{e_{C_1}} \in E$ and $p^E(t_{e_{C_1}}) = \sum_{\substack{t' \in C_1 \\ t' \succ_s t}} p(t') = p((\text{Temp: } 22)) = 0.6$. Therefore,

E: p^E:

Event	Prob
t_{e_t}	0.6
$t_{e_{C_1}}$	0.6

Proposition 2. *An induced event relation in Definition 7 is a simple probabilistic relation.*

Evaluating Global-Top*k* Queries With the help of *induced event relation*, we can reduce Global-Top$k$ in the general case to Global-Top$k$ in simple probabilistic relations.

Lemma 1. *Let $R^p = \langle R, p, \mathcal{C} \rangle$ be a probabilistic relation, s an injective scoring function, $t \in R$, and $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ the event relation induced by t. Define $Q^p = \langle E - \{t_{e_t}\}, p^E, \mathcal{C}^E - \{\{t_{e_t}\}\} \rangle$. Then, the Global-Top$k$ probability of t satisfies the following:*

$$P^{R^p}_{k,s}(t) = p(t) \sum_{\substack{W_e \in pwd(Q^p) \\ |W_e| < k}} Pr(W_e).$$

Proposition 3. *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ and an injective scoring function s, for any $t \in R^p$, the Global-Top$k$ probability of t equals the Global-Top$k$ probability of t_{e_t} when evaluating top-k in the induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ under the injective scoring function $s^E : E \rightarrow \mathbb{R}, s^E(t_{e_t}) = \frac{1}{2}$ and $s^E(t_{e_{C_i}}) = i$:*

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

Proof. See Appendix.

In Proposition 3, the choice of the function s^E is rather arbitrary. In fact, any injective functioin giving t_{e_t} the lowest score will do. Every tuple other than t in the induced event relation corresponds to the event that a tuple with a score higher than that of t occurs. We want to track the case that at most $k-1$ such events happen. Since any induced event relation is simple (Proposition 2), Proposition 3 illustrates how we can reduce the computation of $P^{R^p}_{k,s}(t)$ in the original probabilistic relation to a top-k computation in a simple probabilistic relation, where we can apply the DP technique described in Section 4.1. The complete algorithms are shown as Algorithm 3 and Algorithm 4.

In Algorithm 4, we first find the part $C_{id(t)}$ where t belongs. In Line 4, we initialize the support relation E of the induced event relation by the tuple generated by Rule 1 in Definition 7. For any part C_i other than $C_{id(t)}$, we compute the probability of the event e_{C_i} according to Definition 7 (Line 4), and add it to E if its probability is non-zero (Line 5-7). Since all the tuples from the same part are exclusive, this probability is the sum of the probabilities of all tuples that qualify in that part. Note that if no tuple

Algorithm 3 (IndEx_Topk) Evaluate Global-Top$k$ Queries in a General Probabilistic Relation under an Injective Scoring Function

Require: $R^p = \langle R, p, \mathcal{C} \rangle, k, s$

Initialize a fixed cardinality $k+1$ priority queue Ans of $\langle t, prob \rangle$ pairs, which compares pairs on $prob$, i.e. the Global-Top$k$ probability of t;
for $t \in R$ **do**
  Calculate $P^{R^p}_{k,s}(t)$ using Algorithm 4, i.e.

$$P^{R^p}_{k,s}(t) = \text{IndEx\_Topk\_Sub}(R^p, k, s, t);$$

  Add $\langle t, P^{R^p}_{k,s}(t) \rangle$ to Ans;
  if $|Ans| > k$ **then**
    remove the pair with the smallest $prob$ value from Ans;
  end if
end for
return $\{t | \langle t, P^{R^p}_{k,s}(t) \rangle \in Ans\}$;

Algorithm 4 (IndEx_Topk_Sub) Calculate $P^{R^p}_{k,s}(t)$ using an induced event relation

Require: $R^p = \langle R, p, \mathcal{C} \rangle, k, s, t \in R$

Find the part $C_{id(t)} \in \mathcal{C}$ such that $t \in C_{id(t)}$;
$E = \{t_{e_t}\}$, where $p^E(t_{e_t}) = p(t)$;
for $C_i \in \mathcal{C}$ and $C_i \neq C_{id(t)}$ **do**

$$p(e_{C_i}) = \sum_{\substack{t' \in C_i \\ t' \succ_s t}} p(t');$$

  if $p(e_{C_i}) > 0$ **then**
    $E = E \cup \{t_{e_{C_i}}\}$, where $p^E(t_{e_{C_i}}) = p(e_{C_i})$;
  end if
end for
Use Algorithm 2 to compute Global-Top$k$ probabilities in $E^p = \langle E, p^E, \mathcal{C}^E \rangle$, i.e.

$$q(0 \dots k, 1 \dots |E|) = \text{Ind\_Topk\_Sub}(E^p, k)$$

$P^{R^p}_{k,s}(t) = P^{E^p}_{k,s^E}(t_{e_t}) = q(k, |E|)$;
return $P^{R^p}_{k,s}(t)$;

from C_i qualifies, this probability is zero. In this case, we do not care whether any tuple from C_i will be in the possible world or not, since it does not have any influence on whether t will be in top-k or not. The corresponding event tuple is therefore excluded from E. By default, any probabilistic database assumes that any tuple not in the support relation is with probability zero. Line 4 uses Algorithm 2 to compute $P_{k,s}^{E^p}(t_{e_t})$. Note that Algorithm 2 requires all tuples be sorted on score, but this is not a problem for us. Since we already know the scoring function s^E, we simply need to organize tuples based on s^E when generating E. No extra sorting is necessary.

Theorem 3 (Correctness of Algorithm 3). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, a non-negative integer k and an injective scoring function s, Algorithm 3 correctly computes a Global-Top$k$ answer set of R^p under the scoring function s.*

Proof. The top-level structure with the priority queue in Algorithm 3 resemble those in Algorithm 1. Therefore, as long as Line 3 in Algorithm 3 correctly computes the Global-Top$k$ probability of each tuple in R, Algorithm 3 returns a valid Global-Top$k$ answer set. Line 1-8 in Algorithm 4 computes the event relation induced by tuple t. By Proposition 3, Line 9-10 in Algorithm 4 correctly computes the Global-Top$k$ probability of tuple t.

In Algorithm 4, Line 4-4 takes $O(n)$ to build E (we need to scan all tuples within each part). The call to Algorithm 2 in Line 4 takes $O(k|E|)$, where $|E|$ is no more than the number of parts in partition $\mathcal{C}$, which is in turn no more than n. So Algorithm 4 takes $O(kn)$. Algorithm 3 make n calls to Algorithm 4 to compute $P_{k,s}^{R^p}(t)$ for every tuple $t \in R$. Again, Algorithm 3 uses a priority queue to select the final answer set, which takes $O(n \log k)$. The entire algorithm takes $O(kn^2 + n \log k) = O(kn^2)$.

5 Global-Top*k* under General Scoring Functions

5.1 Semantics and Postulates

Global-Top*k* Semantics with Allocation Policy Under a general scoring function, the Global-Top$k$ semantics remains the same. However, the definition of Global-Top$k$ probability in Definition 5 needs to be generalized to handle *ties*.

Recall that under an injective scoring function s, there is a unique top-k answer set S in every possible world W. When the scoring function s is non-injective, there may be multiple top-k answer sets $S_1, \ldots, S_d$, each of which is returned nondeterministically. Therefore, for any tuple $t \in \cap S_i, i = 1, \ldots, d$, the world W contributes $Pr(W)$ to the Global-Top$k$ probability of t. One the other hand, for any tuple $t \in (\cup S_i - \cap S_i), i = 1 \ldots, d$, the world W contributes only a *fraction* of $Pr(W)$ to the Global-Top$k$ probability of t. The *allocation policy* determines the value of this fraction, i.e. the *allocation coefficient*. Denote by $\alpha(t, W)$ the allocation coefficient of a tuple t in a world W. Let $all_{k,s}(W) = \cup S_i, i = 1, \ldots, d$.

Definition 8 (Global-Top$k$ Probability under a General Scoring Function). *Assume a probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, a non-negative integer k and a scoring function s over R^p. For any tuple t in R, the Global-Top$k$ probability of t, denoted by $P_{k,s}^{R^p}(t)$, is* *the sum of the (partial) probabilities of all possible worlds of R^p whose top-k answer may contain t.*

$$P^{R^p}_{k,s}(t) = \sum_{\substack{W \in pwd(R^p) \\ t \in all_{k,s}(W)}} \alpha(t,W) Pr(W). \tag{4}$$

With no prior bias towards any tuple, it is natural to assume that each of $S_1, \ldots, S_d$ is returned nondeterministically with *equal* probability. Notice that this probability has nothing to do with tuple probabilities. Rather, it is the determined by the number of equally qualified top-k answer sets. Hence, we have the following *Equal* allocation policy.

Definition 9 (Equal Allocation Policy). *Assume a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a non-negative integer k and a scoring function s over R^p. For a possible world $W \in pwd(R^p)$ and a tuple $t \in W$, let $a = |\{t' \in W | t' \succ_s t\}|$ and $b = |\{t' \in W | t' \sim_s t\}|$*

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

Satisfaction of Postulates The semantic postulates in Section 3.1 are directly applicable to Global-Top$k$ with allocation policy. In the Appendix, we show that the *Equal* allocation policy preserves the semantic postulates of Global-Top$k$.

5.2 Query Evaluation in Simple Probabilistic Relations

Definition 10. *Let $R^p = \langle R, p, \mathcal{C} \rangle$ be a probabilistic relation, k a non-negative integer and s a general scoring function over R^p. Assume that $R = \{t_1, t_2, \ldots, t_n\}$, $t_1 \succeq_s t_2 \succeq_s \ldots \succeq_s t_n$. Let $T^{R^p}_{k,[i]}$, $k \leq i$, be the sum of the probabilities of all possible worlds of exactly k tuples from $\{t_1, \ldots, t_i\}$:*

$$T^{R^p}_{k,[i]} = \sum_{\substack{W \in pwd(R^p) \\ |W \cap \{t_1, \ldots, t_i\}| = k}} Pr(W)$$

As usual, we omit the superscript in $T^{R^p}_{k,[i]}$, i.e. $T_{k,[i]}$, when the context is unambiguous. Remark 1 shows that in a simple probabilistic relation $T_{k,[i]}$ can be computed efficiently.

Remark 1. Let $R^p = \langle R, p, \mathcal{C} \rangle$ be a simple probabilistic relation, k a non-negative integer and s a general scoring function over R^p. Assume that $R = \{t_1, t_2, \ldots, t_n\}$, $t_1 \succeq_s t_2 \succeq_s \ldots \succeq_s t_n$. For any i, $1 \leq i \leq n-1$, $T^{R^p}_{k,[i]}$ can be computed using the DP table for computing the Global-Top$k$ probabilities in R^p under an order-preserving injective scoring function s' such that $t_1 \succ_{s'} t_2 \succ_{s'} \ldots \succ_{s'} t_n$.

Proof. We show by case study.

– Case 1: If $k = 0$, $1 \leq i \leq n-1$, then

$$T^{R^p}_{k,[i]} = \prod_{1 \leq j \leq i} \overline{p}(t_j) = \frac{P^{(R^p)}_{1,s'}(t_{i+1})}{p(t_{i+1})}$$

– Case 2: For every $1 \leq k \leq i \leq n-1$, by the definition of $T^{R^p}_{k,[i]}$, we have

$$T^{R^p}_{k,[i]} = \sum_{\substack{W \in pwd(R^p) \\ |W \cap \{t_1,\ldots,t_i\}| \leq k}} Pr(W) - \sum_{\substack{W \in pwd(R^p) \\ |W \cap \{t_1,\ldots,t_i\}| \leq k-1}} Pr(W)$$

In the DP table computing the Global-Top$k$ probabilities in R^p under function s', we have

$$\begin{aligned} P^{R^p}_{k+1,s'}(t_{i+1}) &= \sum_{\substack{W \in pwd(R^p) \\ t_{i+1} \in top_{k+1,s'}(W)}} Pr(W) && (s' \text{ is injective}) \\ &= \sum_{\substack{W \in pwd(R^p) \\ |W \cap \{t_1,\ldots,t_i\}| \leq k \\ t_{i+1} \in W}} Pr(W) \\ &= p(t_{i+1}) \sum_{\substack{W \in pwd(R^p) \\ |W \cap \{t_1,\ldots,t_i\}| \leq k}} Pr(W) && (\text{tuples are independent}) \end{aligned}$$

Therefore,

$$T^{R^p}_{k,[i]} = \frac{P^{R^p}_{k+1,s'}(t_{i+1})}{p(t_{i+1})} - \frac{P^{R^p}_{k,s'}(t_{i+1})}{p(t_{i+1})}$$

Since $1 \leq k \leq i \leq n-1$, both $P^{R^p}_{k+1,s'}(t_{i+1})$ and $P^{R^p}_{k,s'}(t_{i+1})$ can be computed using the DP table used to compute the Global-Top$k$ probabilities of tuples in R^p under the injective scoring function s'.

Remark 2 shows that we can compute Global-Top$k$ probability under a general scoring function in polynomial time for an extreme case, where the probabilistic relation is simple and all tuples tie in scores. As we will see shortly, this special case plays an important role in our major result Proposition 4.

Remark 2. Let $R^p = \langle R, p, \mathcal{C} \rangle$ be a simple probabilistic relation, k a non-negative integer and s a general scoring function over R^p. Assume that $R = \{t_1, \ldots, t_m\}$ and $t_1 \sim_s t_2 \sim_s \ldots \sim_s t_m$. For any tuple t_i, $1 \leq i \leq m$, the Global-Top$k$ probability of t_i, i.e. $P^{R^p}_{k,s}(t_i)$, can be computed using Remark 1.

Proof. If $k > m$, it is trivial that $P^{R^p}_{k,s}(t_i) = p(t_i)$. Therefore, we only prove the case when $k \leq m$. According to Equation 4, for any i, $1 \leq i \leq m$,

$$
\begin{aligned}
P^{R^p}_{k,s}(t_i) &= \sum_{j=1}^{m} \sum_{\substack{W \in pwd(R^p) \\ t_i \in all_{k,s}(W), |W|=j}} \alpha(t_i, W) Pr(W) \\
&= \sum_{j=1}^{m} \sum_{\substack{W \in pwd(R^p) \\ t_i \in W, |W|=j}} \alpha(t_i, W) Pr(W) \qquad \text{(Since all tuple tie , } all_{k,s}(W) = W) \\
&= \sum_{j=1}^{k} \sum_{\substack{W \in pwd(R^p) \\ t_i \in W, |W|=j}} \alpha(t_i, W) Pr(W) + \sum_{j=k+1}^{m} \sum_{\substack{W \in pwd(R^p) \\ t_i \in W, |W|=j}} \alpha(t_i, W) Pr(W) \\
&= \sum_{j=1}^{k} \sum_{\substack{W \in pwd(R^p) \\ t_i \in W, |W|=j}} Pr(W) + \sum_{j=k+1}^{m} \frac{k}{j} \sum_{\substack{W \in pwd(R^p) \\ t_i \in W, |W|=j}} Pr(W)
\end{aligned}
$$

With out loss of generality, assume $i = m$, then the above equation becomes

$$
\begin{aligned}
P^{R^p}_{k,s}(t_m) &= \sum_{j=1}^{k} \sum_{\substack{W \in pwd(R^p) \\ t_m \in W, |W|=j}} Pr(W) + \sum_{j=k+1}^{m} \frac{k}{j} \sum_{\substack{W \in pwd(R^p) \\ t_m \in W, |W|=j}} Pr(W) \\
&= p(t_i) (\sum_{j=1}^{k} T^{R^p}_{j-1,[m-1]} + \sum_{j=k+1}^{m} \frac{k}{j} T^{R^p}_{j-1,[m-1]}) \qquad (5)
\end{aligned}
$$

By Remark 1, every $T^{R^p}_{j-1,[m-1]}$ can be computed using the DP table computing Global-Top$k$ probabilities in R^p under an order preserving injective scoring function s'. Therefore, Equation 5 can be computed using Remark 1.

Based on Remark 1 and Remark 2, we design Algorithm 5 and prove its correctness in Theorem 4 using Proposition 4.

Assume $R^p = \langle R, p, \mathcal{C} \rangle$ where $R = \{t_1, t_2, \ldots, t_n\}$ and $t_1 \succeq_s t_2 \succeq_s \ldots \succeq_s t_n$. For any $t_l \in R$, i_l is the largest index such that $t_{i_l} \succ_s t_l$, and j_l is the largest index such that $t_{j_l} \succeq_s t_l$.

Intuitively, Algorithm 5 and Proposition 4 convey the idea that, in a simple probabilistic relation, the computation of Global-Top$k$ under the *Equal* allocation policy can be simulated by the following procedure:

(S1) Independently flip a biased coin with probability $p(t_j)$ for each tuple $t_j \in R = \{t_1, t_2, \ldots, t_n\}$, which gives us a possible world $W \in pwd(R^p)$;

(S2) Return a top-k answer set S of W nondeterministically (with equal probability in the presence of multiple top-k sets). The Global-Top$k$ probability of t_l is the probability that $t_l \in S$.

The above Step (S1) can be further refined into:

(S1.1) Independently flip a biased coin with probability $p(t_j)$ for each tuple $t_j \in R_A = \{t_1, t_2 \ldots, t_{i_l}\}$, which gives us a collection of tuples W_A;

(S1.2) Independently flip a biased coin with probability $p(t_j)$ for each tuple $t_j \in R_B = \{t_{i_l+1}, \ldots, t_n\}$, which gives us a collection of tuples W_B. $W = W_A \cup W_B$ is a possible world from $pwd(R^p)$;

In order for t_l to be in S, W_A can have at most $k-1$ tuples. Let $|W_A| = k'$, then $k' < k$. Every top-k answer set S of W contains all k' tuples from W_A, plus the top-$(k-k')$ tuples from W_B. For t_l to be in S, it has to be in the top-$(k-k')$ set of W_B. Consequently, the probability of $t_l \in S$, i.e. the Global-Top$k$ probability of t_l, is the joint probability that $|W_A| = k' < k$ and t_l belongs to the top-$(k-k')$ set of W_B. The former is $T_{k',[i_l]}$ and the latter is $P^{R^p_B}_{k-k',s}(t_l)$, where R^p_B is R^p restricted to R_B. Again, due to the independence among tuples, Step (S1.1) and Step (S1.2) are independent, and their joint probability is simply the product of the two.

Further notice that since t_l has the highest score in R_B and all tuples are independent in R_B, any tuple with score lower than that of t_l does not have influence on $P^{R^p_B}_{k-k',s}(t_l)$. In other words, $P^{R^p_B}_{k-k',s}(t_l) = P^{R^p_s(t_l)}_{k-k',s}(t_l)$, where $R^p_s(t_l)$ is R^p restricted to all tuples tying with t_l in R. Notice that the computation of $P^{R^p_s(t_l)}_{k-k',s}(t_l)$ is the extreme case addressed in Remark 2.

Algorithm 5 elaborates the algorithm based on the idea above, where $m = j_l - i_l$ is the number of tuples tying with t_l (including t_l).

Furthermore, Algorithm 5 exploits the overlapping among DP tables and makes the following two optimizations:

1. Use a single DP table to collect the information needed to compute all $T_{k',[i_l]}$, $k' = 0, \ldots, k-1$, $l = 1, \ldots, n$ and $k' \leq i_l$ (Line 2).

   Notice that for $1 \leq l \leq n$, $1 \leq i_l \leq n-1$. It is easy to see that the DP table computing $T_{k-1,[n-1]}$ subsumes all other DP tables.

2. Use a single DP table to compute all $P^{R^p_s(t_l)}_{k-k',s}(t_l)$, $k' = 0, \ldots, k-1$, for a tuple t_l (Line 8-18).

   For different k', the computation of $P^{R^p_s(t_l)}_{k-k',s}(t_l)$ requires the computation of the same set of $T^{R^p_s(t_l)}_{j,[m-1]}$. In Line 8-18, $P^{R^p_s(t_l)}_{k-k',s}(t_l)$ is abbreviated as $P_l(k-k')$ to emphasize the changing parameter k'.

Each DP table computation uses a call to Algorithm 2 (Line 2 in Algorithm 5, Line 3 in Algorithm 6).

Algorithm 5 (Ind_Topk_Gen) Evaluate Global-Top$k$ Queries in a Simple Probabilistic Relation under a General Scoring Function

Require: $R^p = \langle R, p, \mathcal{C}\rangle, k$

Ensure: tuples in R are sorted in the non-increasing order based on s

Initialize a fixed cardinality $(k+1)$ priority queue Ans of $\langle t, prob\rangle$ pairs, which compares pairs on $prob$, i.e. the Global-Top$k$ probability of t;

Get the DP table for computing $T_{k',[i]}, k' = 0, \ldots k-1, i = 1, \ldots, n-1$, $k' \leq i$ using Algorithm 2, i.e.

$$q(0 \ldots k, 1 \ldots |R|) = \text{Ind\_Topk\_Sub}(R^p, k);$$

for $l = 1$ to $|R|$ **do**

  $m = j_l - i_l$;

  if $m == 1$ **then**

    Add $\langle t_l, q(k,l)\rangle$ to Ans;

  else

    Get the DP table for computing $P^{R^p_s(t_l)}_{k-k',s}(t_l)$, i.e. $P_l(k-k')$, $k' = 0, \ldots, k-1$

$$q_{tie}(0 \ldots m, 1 \ldots m) = \text{Ind\_Topk\_Gen\_Sub}(R^p_s(t_l), t_l, m);$$

    $P_l(0 \ldots \max(m,k)) = 0$;

    for $k'' = 1$ to $\min(k,m)$ **do**

      $P_l(k'') = P_l(k''-1) + q_{tie}(k'', m)$;

    end for

    for $k'' = k+1$ to m **do**

      $P_l(k'') = P_l(k''-1) + \frac{k}{k''} q_{tie}(k'', m)$;

    end for

    for $k'' = m+1$ to k **do**

      $P_l(k'') = p(t_l)$;

    end for

    $P^{R^p}_{k,s}(t_l) = 0$;

    for $k' = 0$ to $k-1$ **do**

$$T_{k',[i_l]} = \frac{q(k'+1, i_l+1) - q(k', i_l+1)}{p(t_{i_l+1})};$$

$$P^{R^p}_{k,s}(t_l) = P^{R^p}_{k,s}(t_l) + T_{k',[i_l]} \cdot P_l(k-k');$$

    end for

    Add $\langle t_l, P^{R^p}_{k,s}(t_l)\rangle$ to Ans;

  end if

  if $|Ans| > k$ **then**

    remove the pair with the smallest $prob$ value from Ans;

  end if

end for

return $\{t_i | \langle t_i, prob\rangle \in Ans\}$;

Algorithm 6 (Ind_Topk_Gen_Sub) Compute the DP table for Global-Top$k$ probabilities in a Simple Probabilistic Relation under an All-Tie Scoring Function

Require: $R_s^p(t_{target}) = \langle R, p, \mathcal{C}\rangle, t_{target}, m$

Ensure: $|R| = m$, $t_{target} \in R$

Rearrange tuples in R such that $R = \{t_1, \ldots, t_{m-1}, t_m\}$ and $t_m = t_{target}$;

Assume the injective scoring function s' is such that $t_1 \succ_{s'} \ldots \succ_{s'} t_{m-1} \succ_{s'} t_{target}$;

Get the DP table

$$q_{tie}(0 \ldots m, 1 \ldots m) = \text{Ind\_Topk\_Sub}(R_s^p(t_{target}), m);$$

return $q_{tie}(0 \ldots m, 1 \ldots m)$;

Proposition 4. *Let $R^p = \langle R, p, \mathcal{C}\rangle$ be a simple probabilistic relation where $R = \{t_1, \ldots, t_n\}$, $t_1 \succeq_s t_2 \succeq_s \ldots \succeq_s t_n$, k a non-negative integer and s a scoring function. For every $t_l \in R$, the Global-Top$k$ probability of t_l can be computed by the following equation:*

$$P_{k,s}^{R^p}(t_l) = \sum_{k'=0}^{k-1} T_{k',[i_l]} \cdot P_{k-k',s}^{R_s^p(t_l)}(t_l) \tag{6}$$

where $R_s^p(t_l)$ is R^p restricted to $\{t \in R | t \sim_s t_l\}$.

Proof. *See* Appendix.

Theorem 4 (Correctness of Algorithm 5). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, a non-negative integer k and a general scoring function s, Algorithm 5 correctly computes a Global-Top$k$ answer set of R^p under the scoring function s.*

Proof. In Algorithm 5, by Remark 1, Line 2 and Line 9 correctly computes $T_{k',[i]}$ for $0 \le k' \le k-1$, $1 \le i \le n-1$, $k' \le i$. In Line 8, each entry $q_{tie}(k'', m) = p(t_l) T_{k''-1,[m-1]}^{R_s^p(t_l)}$, $1 \le k'' \le m$. By Remark 2, Line 8 collects the information for computing $P_{k-k',s}^{R_s^p(t_l)}(t_l)$, $1 \le k-k' \le m$. Line 9-15 correctly compute those cases based on the definition. If $m < k-k' \le k$, then it is trivial that $P_{k-k',s}^{R_s^p(t_l)}(t_l) = p(t_l)$ (Line 16-18). By Proposition 4, Line 19-23 correctly computes the Global-Top$k$ probability of t_l. Also notice that in Line 6, the Global-Top$k$ probability of a tuple without tying tuples is retrieved directly. It is an optimization as the code handling the general case (i.e. $m > 1$, Line 7-24) works for this special case as well. Again, the top-level structure with the priority queue in Algorithm 5 ensures that a Global-Top$k$ answer set is correctly computed.

In Algorithm 5, Line 2 takes $O(kn)$, and for each tuple, there is one call to Algorithm 6 in Line 8, which takes $O(m_{\max}^2)$, where $m_{\max}$ is the maximal number of tying tuples. Therefore, Algorithm 5 takes $O(n \max(k, m_{\max}^2))$ altogether.

5.3 Query Evaluation in General Probabilistic Relations

Recall that under an injective scoring function, every tuple t in a general probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ induces a *simple* event relation E^p, and we reduce the computation of t's Global-Top$k$ probability in R^p to the computation of t_{e_t}'s Global-Top$k$ probability in E^p.

In the case of general scoring functions, we use the same reduction idea. However, now for each part $C_i \in \mathcal{C}, C_i \neq C_{id(t)}$, tuple t induces in E^p two *exclusive* tuples $t_{e_{C_i,\succ}}$ and $t_{e_{C_i,\sim}}$, corresponding to the *event* $e_{C_i,\succ}$ that "some tuple from the part C_i has the score *higher than* that of t" and the *event* $e_{C_i,\sim}$ that "some tuple from the part C_i has the score *equal to* that of t", respectively. In addition, in Definition 11, we allow the existence of tuples with probability 0, in order to simplify the description of query evaluation algorithms. This is an artifact whose purpose will become clear in Theorem 5.

Definition 11 (Induced Event Relation under General Scoring Functions). *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a scoring function s over R^p and a tuple $t \in C_{id(t)} \in \mathcal{C}$, the event relation induced by t, denoted by $E^p = \langle E, p^E, \mathcal{C}^E \rangle$, is a probabilistic relation whose support relation E has only one attribute, $Event$. The relation E and the probability function p^E are defined by the following four generation rules and the postprocess step:*

- *Rule 1.1:* $t_{e_{t,\sim}} \in E$ *and* $p^E(t_{e_{t,\sim}}) = p(t)$;
- *Rule 1.2:* $t_{e_{t,\succ}} \in E$ *and* $p^E(t_{e_{t,\succ}}) = 0$;
- *Rule 2.1:*

$$\forall C_i \in \mathcal{C} \wedge C_i \neq C_{id(t)}.(t_{e_{C_i,\succ}} \in E) \text{ and } p^E(t_{e_{C_i,\succ}}) = \sum_{\substack{t' \in C_i \\ t' \succ_s t}} p(t');$$

- *Rule 2.2:*

$$\forall C_i \in \mathcal{C} \wedge C_i \neq C_{id(t)}.(t_{e_{C_i,\sim}} \in E) \text{ and } p^E(t_{e_{C_i,\sim}}) = \sum_{\substack{t' \in C_i \\ t' \sim_s t}} p(t').$$

Postprocess step: only when $p^E(t_{e_{C_i,\succ}})$ and $p^E(t_{e_{C_i,\sim}})$ are both 0, delete both tuple $t_{e_{C_i,\succ}}$ and $t_{e_{C_i,\sim}}$.

Proposition 5. *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ and a scoring function s, for any $t \in R^p$, the Global-Top$k$ probability of t equals the Global-Top$k$ probability of $t_{e_{t,\sim}}$ when evaluating top-k in the induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ under the scoring function $s^E : E \rightarrow \mathbb{R}$, $s^E(t_{e_t}) = \frac{1}{2}$, $s^E(t_{e_{t,\sim}}) = \frac{1}{2}$ and $s^E(t_{e_{C_i,\succ}}) = i$:*

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

Proof. See Appendix.

Notice that the induced event relation E^p in Definition 11, unlike its counterpart under an injective scoring function, is not simple. Therefore, we cannot utilize the algorithm in Proposition 4. Rather, the induced relation E^p is a special general probabilistic relation, where each part of the partition contains *exactly* two tuples. For this special general probabilistic relation, the recursion in Theorem 5 (Equation 7,8) collects enough information to compute the Global-Top$k$ probability of $t_{e_{t,\sim}}$ in E^p (Equation 9).

Definition 12 (Secondary Induced Event Relations). *Let $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ be the event relation induced by tuple t under a general scoring function s. Without loss of generality, assume*

$$E = \{t_{e_{C_1,\succ}}, t_{e_{C_1,\sim}}, \ldots, t_{e_{C_{m-1},\succ}}, t_{e_{C_{m-1},\sim}}, t_{e_{t,\succ}}, t_{e_{t,\sim}}\}$$

we can split E into two non-overlapping subsets $E_\succ$ and $E_\sim$ such that

$$E_\succ = \{t_{e_{C_1,\succ}}, \ldots, t_{e_{C_{m-1},\succ}}, t_{e_{t,\succ}}\}$$
$$E_\sim = \{t_{e_{C_1,\sim}}, \ldots, t_{e_{C_{m-1},\sim}}, t_{e_{t,\sim}}\}$$

The two secondary induced event relation *$E^p_\succ$ and $E^p_\sim$ are E^p restricted to $E^p_\succ$ and $E^p_\sim$ respectively. They are both mutually related and simple probabilistic relations. For every $1 \leq i \leq m-1$, tuple $t_{i,\succ}$ ($t_{i,\sim}$ resp.) refers to $t_{e_{C_i,\succ}}$ ($t_{e_{C_i,\sim}}$ resp.). The tuple $t_{m,\succ}$ ($t_{m,\sim}$ resp.) refers to $t_{e_{t,\succ}}$ ($t_{e_{t,\sim}}$ resp.).*

In spirit, the recursion in Theorem 5 is close to the recursion in Proposition 1, even though they are not computing the same measure. The following table does a comparison between the measure q in Proposition 1 and the measure u in Theorem 5:

Measure	$= \sum Pr(W)$	$\lvert\{t_j \lvert t_j \in W, j \leq i, t_j \sim_s t\}\rvert$
$q(k,i)$	(1) W contains t_i (2) W has *no more than* k tuples from $\{t_1, t_2, \ldots, t_i\}$	-
$u_{\succ/\sim}(k,i,b)$	(1) W contains t_i (2) W has *exactly* k tuples from $\{t_1, t_2, \ldots, t_i\}$	b

Under the general scoring function s^E, a possible world of an induced relation E^p may partially contribute to tuple $t_{m,\sim}$'s Global-Top$k$ probability. The allocation coefficient depends on the combination of two factors: the number of tuples that are strictly better than $t_{m,\sim}$ and the number of tuples tying with $t_{m,\sim}$. Therefore, in the new measure u, first, we add one more dimension to keep track of b, i.e. the number of tying tuples of a subscript no more than i in a world. Second, we keep track of distinct (k,b) pairs. Furthermore, the recursion on measure u differentiates between two cases: a non-tying tuple (handled by $u_\succ$) and a tying tuple (handled by $u_\sim$), since those two types of tuples have different influence on the values of k and b.

Formally, let $u_\succ(k', i, b)$ ($u_\sim(k', i, b)$ resp.) be the sum of the probabilities of all the possible worlds W of E^p such that

1. $t_{i,\succ} \in W$ ($t_{i,\sim} \in W$ resp.)
2. i is the k'th smallest tuple subscript in world W
3. the world W contains b tuples from $E^p_\sim$ with subscript less than or equal to i.

Equation 7,8 resemble Equation 3, except that now, since we introduce tuples with probability 0 to ensure that each part of $\mathcal{C}^E$ has exactly two tuples, we need to address the special cases when divisor can be zero. Notice that, for any $i, 1 \leq i \leq m$, at least one of $p^E(t_{i,\succ})$ and $p^E(t_{i,\sim})$ is non-zero, otherwise, they are not in E^p by definition.

Theorem 5. *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a scoring function s, $t \in R^p$, and its induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$, where $|E| = 2m$, the following recursion on $u_{\succ}(k', i, b)$ and $u_{\sim}(k', i, b)$ holds, where $b_{\max}$ is the number of tuples with positive probability in $E^p_{\sim}$.*
When $i = 1, 0 \leq k' \leq m$ and $0 \leq b \leq b_{\max}$,

$$u_{\succ}(k', 1, b) = \begin{cases} p^E(t_{1,\succ}) & k' = 1, b = 0 \\ 0 & \textit{otherwise} \end{cases}$$

$$u_{\sim}(k', 1, b) = \begin{cases} p^E(t_{1,\sim}) & k' = 1, b = 1 \\ 0 & \textit{otherwise} \end{cases}$$

For every i, $2 \leq i < m$, $0 \leq k' \leq m$ and $0 \leq b \leq b_{\max}$,

$$u_{\succ}(k', i, b) = \begin{cases} 0 & k' = 0 \\ (u_{\succ}(k', i-1, b)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\succ})} & 1 \leq k' \leq m \\ \quad + u_{\succ}(k'-1, i-1, b) & \textit{and } p^E(t_{i-1,\succ}) > 0 \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \\ (u_{\sim}(k', i-1, b+1)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\sim})} & b < b_{\max} \\ \quad + u_{\succ}(k'-1, i-1, b) & \textit{and } 1 \leq k' \leq m \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \textit{and } p^E(t_{i-1,\succ}) = 0 \\ (u_{\succ}(k'-1, i-1, b) & \textit{otherwise} \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \end{cases} \tag{7}$$

$$u_{\sim}(k', i, b) = \begin{cases} 0 & k' = 0 \textit{ or } b = 0 \\ (u_{\sim}(k', i-1, b)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\sim})} & b > 0 \\ \quad + u_{\succ}(k'-1, i-1, b-1) & \textit{and } 1 \leq k' \leq m \\ \quad + u_{\sim}(k'-1, i-1, b-1))p^E(t_{i,\sim}) & \textit{and } p^E(t_{i-1,\sim}) > 0 \\ (u_{\succ}(k', i-1, b-1)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\succ})} & \textit{otherwise} \\ \quad + u_{\succ}(k'-1, i-1, b-1) & \\ \quad + u_{\sim}(k'-1, i-1, b-1))p^E(t_{i,\sim}) & \end{cases} \tag{8}$$

The Global-Top$k$ probability of $t_{e_t,\sim}$ in E^p under the scoring function s^E can be computed by the following equation:

$$\begin{aligned} P^{E^p}_{k,s^E}(t_{e_t,\sim}) &= P^{E^p}_{k,s^E}(t_{m,\sim}) \\ &= \sum_{b=1}^{b_{\max}} \Big(\sum_{k'=1}^{k} u_{\sim}(k', m, b) + \sum_{k'=k+1}^{k+b-1} \frac{k - (k' - b)}{b} u_{\sim}(k', m, b) \Big) \end{aligned} \tag{9}$$

Proof. See Appendix.

Recall that we design Algorithm 1 based on the recursion in Proposition 1. Similarly, a DP algorithm based on the mutual recursion in Theorem 5 is available. We are going skip the details. Instead, we show how the algorithm works using the following example.

The complexity of the recursion in Theorem 5 determines the complexity of the algorithm. It takes $O(b_{\max} n^2)$ for one tuple, and $O(m_{\max} n^3)$ for computing all n tuples. Recall that $m_{\max}$ is the maximal number of tying tuples in R. Again, the priority queue takes $O(n \log k)$. Altogether, the algorithm takes $O(m_{\max} n^3)$ time.

Example 7. When evaluating a top-2 query in $R^p = \langle R, p, \mathcal{C} \rangle$, consider a tuple $t \in R$ and its induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$

$E_\succ$	$t_{e_{C_1},\succ}$ (t_1)	$t_{e_{C_2},\succ}$ (t_3)	$t_{e_{C_3},\succ}$ (t_5)	$t_{e_t,\succ}$ (t_7)
p^E	0.6	0.5	0.2	0

$E_\sim$	$t_{e_{C_1},\sim}$ (t_2)	$t_{e_{C_2},\sim}$ (t_4)	$t_{e_{C_3},\sim}$ (t_6)	$t_{e_t,\sim}$ (t_8)
p^E	0	0.25	0.6	0.4

In order to compute the Global-Top$k$ probability of t_8 (i.e. $t_{e_t,\sim}$) in E^p, Theorem 5 leads to the following DP tables, each for a distinct combination of a b value and a secondary induced relation, where $b_{\max} = 3$.

($b = 0$, $E^p_\succ$)

$k \backslash t$	t_1	t_3	t_5	t_7
0	0	0	0	0
1	0.6	0.2	0.02	0
2	0	0.3	0.07	0
3	0	0	0.06	0
4	0	0	0	0

($b = 0$, $E^p_\sim$)

$k \backslash t$	t_2	t_4	t_6	t_8
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

($b = 1$, $E^p_\succ$)

$k \backslash t$	t_1	t_3	t_5	t_7
0	0	0	0	0
1	0	0	0	0
2	0	0	0.02	0
3	0	0	0.03	0
4	0	0	0	0

($b = 1$, $E^p_\sim$)

$k \backslash t$	t_2	t_4	t_6	t_8
0	0	0	0	0
1	0	0.1	0.06	0.008
2	0	0.15	0.21	0.036
3	0	0	0.18	0.052
4	0	0	0	0.024

($b = 2$, $E^p_\succ$)

$k \backslash t$	t_1	t_3	t_5	t_7
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

($b = 2$, $E^p_\sim$)

$k \backslash t$	t_2	t_4	t_6	t_8
0	0	0	0	0
1	0	0	0	0
2	0	0	0.06	0.032
3	0	0	0.09	0.104
4	0	0	0	0.084

$(b=3,\, E^p_{\succ})$

$k\backslash t$	t_1	t_3	t_5	t_7
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

$(b=3,\, E^p_{\sim})$

$k\backslash t$	t_2	t_4	t_6	t_8
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0.024
4	0	0	0	0.036

The computation of each entry follows the mutual recursion in Theorem 5, for example,

$$\begin{aligned}
u_{\succ}(2,5,0) &= (u_{\succ}(1,3,0) + u_{\sim}(1,4,0) + u_{\succ}(2,3,0)\frac{1-p^E(t_3)-p^E(t_4)}{p^E(t_3)})p^E(t_5) \\
&= (0.2 + 0 + 0.3\frac{1-0.5-0.25}{0.5})0.2 \\
&= 0.07 \\
u_{\sim}(2,6,1) &= (u_{\succ}(1,3,0) + u_{\sim}(1,4,0) + u_{\sim}(2,4,1)\frac{1-p^E(t_3)-p^E(t_4)}{p^E(t_3)})p^E(t_6) \\
&= (0.2 + 0 + 0.15\frac{1-0.5-0.25}{0.25})0.6 \\
&= 0.21
\end{aligned}$$

Finally, under the scoring function s^E defined in Proposition 5

$$\begin{aligned}
P^{E^p}_{k,s^E}(t_{e_t,\sim}) &= P^{E^p}_{2,s^E}(t_8) \\
&= \sum_{b=1}^{3}(\sum_{k'=1}^{2} u_{\sim}(k',8,b) + \sum_{k'=2+1}^{2+b-1}\frac{2-(k'-b)}{b}u_{\sim}(k',8,b)) \\
&= u_{\sim}(1,8,1) + u_{\sim}(2,8,1) \\
&\quad + u_{\sim}(1,8,2) + u_{\sim}(2,8,2) + \frac{1}{2}u_{\sim}(3,8,2) \\
&\quad + u_{\sim}(2,8,3) + u_{\sim}(2,8,3) + \frac{2}{3}u_{\sim}(3,8,3) + \frac{1}{3}u_{\sim}(3,8,4) \\
&= 0.156
\end{aligned}$$

6 Conclusion

We study the semantic and computational problems for top-k queries in probabilistic databases. We propose three desired postulates for a top-k semantics and discuss their satisfaction by all the semantics in the literature. Those postulates are our first step to benchmark different semantics. From the postulates, it is inconclusive that a single semantics is overwhelmingly better. We deem that the choice of the semantics should be guided by the application, which in turn, supports our efforts to explore postulates in order to create a profile of each semantics. Our Global-Top$k$ semantics satisfies those postulates to a large degree. We study the computational problem of query evaluation under Global-Top$k$ semantics for simple and general probabilistic relations when the scoring function is injective. For the former, we propose a dynamic programming algorithm and effectively optimize it with Threshold Algorithm. For the latter, we show a polynomial reduction to the simple case. Furthermore, we extend our Global-Top$k$ semantics to general scoring functions and introduce the concept of allocation policy to handle ties in score. To the best of our knowledge, this is the first attempt to address the tie problem rigorously. Previous work either does not consider ties or uses an arbitrary tie-breaking mechanism. Advanced dynamic programming algorithms are proposed for query evaluation under general scoring functions for both simple and general probabilistic relations.

For completeness, we list in Table 2 the complexity of the best known algorithm for each semantics in the literature. Since no other work address general scoring functions in a systematical way, those results are restricted to injective scoring functions.

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

Table 2. Time Complexity of Different Semantics

7 Future Work

So far, almost unanimously, only independent and exclusive relationship among tuples are considered in the literature [21, 23, 25]. It will be interesting to investigate other complex relationships between tuples. Other possible directions include top-k evaluation in other uncertain database models proposed in the literature [13] and more general preference queries in probabilistic databases.

8 Acknowledgment

We acknowledge the input of Graham Cormode who showed that Faithfulness in general probabilistic relations is problematic. Jan Chomicki acknowledges the discussions with Sergio Flesca.

9 Appendix

9.1 Proofs of Semantic Postulates

Semantics	Exact k	Faithfulness	Stability
†Global-Top$k$	✓(1)	✓/× (5)	✓(9)
PT-k	× (2)	✓/× (6)	✓(10)
U-Top$k$	× (3)	✓/× (7)	✓(11)
U-kRanks	× (4)	× (8)	× (12)

† Postulates of Global-Top$k$ semantics are proved under general scoring functions with *Equal* allocation policy.

Table 3. Postulate Satisfaction for Different Semantics in Table 1

Proof. The following proofs correspond to the numbers next to each entry in the above table.

Assume that we are given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a non-negative integer k and an injective scoring function s.

(1) Global-Top$k$ satisfies *Exact* k.
We compute the Global-Top$k$ probability for each tuple in R. If there is at least k tuples in R, we are always able to pick the k tuples with the highest Global-Top$k$ probability. In case when there are more than $k - r + 1$ tuple(s) with the rth highest Global-Top$k$ probability, where $r = 1, 2 \ldots, k$, only $k - r + 1$ of them will be picked nondeterministically.

(2) PT-k violates *Exact* k
Example 4 illustrates a counterexample in a simple probabilistic relation.

(3) U-Top$k$ violates *Exact* k.
Example 4 illustrates a counterexample in a simple probabilistic relation.

(4) U-kRanks violates *Exact* k.
Example 4 illustrates a counterexample in a simple probabilistic relation.

(5) Global-Top$k$ satisfies *Faithfulness* in simple probabilistic relations while it violates *Faithfulness* in general probabilistic relations.
Simple Probabilistic Relations
Proof. By the assumption, $t_1 \succ_s t_2$ and $p(t_1) > p(t_2)$, so we need to show that $P_{k,s}(t_1) > P_{k,s}(t_2)$.
For every $W \in pwd(R^p)$ such that $t_2 \in all_{k,s}(W)$ and $t_1 \notin all_{k,s}(W)$, obviously $t_1 \notin W$. Otherwise, since $t_1 \succ_s t_2$, t_1 would be in $all_{k,s}(W)$. Since all tuples are independent, there is always a world $W' \in pwd(R^p)$, $W' = (W \backslash \{t_2\}) \cup \{t_1\}$ and $Pr(W') = Pr(W) \frac{p(t_1)\bar{p}(t_2)}{\bar{p}(t_1)p(t_2)}$. Since $p(t_1) > p(t_2)$, $Pr(W') > Pr(W)$. Moreover, t_1 will substitute for t_2 in the top-k answer to W'. It is easy to see that $\alpha(t_1, W') = 1$ in W' and also in any world W such that both t_1 and t_2 are in $all_{k,s}(W)$, $\alpha(t_1, W) = 1$.
Therefore, for the Global-Top$k$ probability of t_1 and t_2, we have

$$
\begin{aligned}
P_{k,s}(t_2) &= \sum_{\substack{W \in pwd(R^p)\\ t_1 \in all_{k,s}(W)\\ t_2 \in all_{k,s}(W)}} \alpha(t_2, W) Pr(W) + \sum_{\substack{W \in pwd(R^p)\\ t_1 \notin all_{k,s}(W)\\ t_2 \in all_{k,s}(W)}} \alpha(t_2, W) Pr(W) \\
&< \sum_{\substack{W \in pwd(R^p)\\ t_1 \in all_{k,s}(W)\\ t_2 \in all_{k,s}(W)}} Pr(W) + \sum_{\substack{W' \in pwd(R^p)\\ t_1 \in all_{k,s}(W')\\ t_2 \notin W'}} Pr(W') \\
&= \sum_{\substack{W \in pwd(R^p)\\ t_1 \in all_{k,s}(W)\\ t_2 \in all_{k,s}(W)}} \alpha(t_1, W) Pr(W) + \sum_{\substack{W' \in pwd(R^p)\\ t_1 \in all_{k,s}(W')\\ t_2 \notin W'}} \alpha(t_1, W') Pr(W') \\
&\leq \sum_{\substack{W \in pwd(R^p)\\ t_1 \in all_{k,s}(W)\\ t_2 \in all_{k,s}(W)}} \alpha(t_1, W) Pr(W) + \sum_{\substack{W' \in pwd(R^p)\\ t_1 \in all_{k,s}(W')\\ t_2 \notin W'}} \alpha(t_1, W') Pr(W') \\
&\quad + \sum_{\substack{W'' \in pwd(R^p)\\ t_1 \in all_{k,s}(W'')\\ t_2 \in W''\\ t_2 \notin all_{k,s}(W'')}} \alpha(t_1, W'') Pr(W'') \\
&= P_{k,s}(t_1).
\end{aligned}
$$

The equality in $\leq$ holds when $s(t_2)$ is among the k highest scores and there are at most k tuples (including t_2) with higher or equal scores. Since there is at least one inequality in the above equation, we have

$$P_{k,s}(t_1) > P_{k,s}(t_2).$$

General Probabilistic Relations
The following is a counterexample.
Say $k = 1$, $R = \{t_1, \ldots, t_9\}$, $t_1 \succ_s \ldots \succ_s t_9$, $\{t_1, \ldots, t_7, t_9\}$ are exclusive. $p(t_i) = 0.1, i = 1 \ldots 7$, $p(t_8) = 0.4$, $p(t_9) = 0.3$.
By Global-Top$k$, the top-1 answer is $\{t_9\}$, while $t_8 \succ_s t_9$ and $p(t_8) > p(t_9)$, which violates *Faithfulness*.

(6) PT-k satisfies *Faithfulness* in simple probabilistic relations while it violates *Faithfulness* in general probabilistic relations.
For simple probabilistic relations, we can use the same proof in (5) to show that PT-k satisfies *Faithfulness*. The only change would be that we need to show $P_{k,s}(t_1) > p_\tau$ as well. Since $P_{k,s}(t_2) > p_\tau$ and $P_{k,s}(t_1) > P_{k,s}(t_2)$, this is obviously true. For general probabilistic relations, we can use the same counterexample in (5) and set threshold $p_\tau = 0.15$.

(7) U-Top$k$ satisfies *Faithfulness* in simple probabilistic relations while it violates *Faithfulness* in general probabilistic relations.
Simple Probabilistic Relations

Proof. By contradiction. If U-Top$k$ violates *Faithfulness* in a simple probabilistic relation, there exists $R^p = \langle R, p, \mathcal{C} \rangle$ and exists $t_i, t_j \in R, t_i \succ_s t_j, p(t_i) > p(t_j)$, and by U-Top$k$, t_j is in the top-k answer to R^p under the scoring function s while t_i is not.
S is a top-k answer to R^p under the function s by the U-Top$k$ semantics, $t_j \in S$ and $t_i \notin S$. Denote by $Q_{k,s}(S)$ the probability of S under the U-Top$k$ semantics. That is,

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

For any world W contributing to $Q_{k,s}(S)$, $t_i \notin W$. Otherwise, since $t_i \succ_s t_j$, t_i would be in $top_{k,s}(W)$, which is S. Define a world $W' = (W \backslash \{t_j\}) \cup \{t_i\}$. Since t_i is independent of any other tuple in R, $W' \in pwd(R^p)$ and $Pr(W') = Pr(W)\frac{p(t_i)\bar{p}(t_j)}{\bar{p}(t_i)p(t_j)}$. Moreover, $top_{k,s}(W') = (S \backslash \{t_j\}) \cup \{t_i\}$. Let $S' = (S \backslash \{t_j\}) \cup \{t_i\}$, then W' contributes to $Q_{k,s}(S')$.

$$\begin{aligned}
Q_{k,s}(S') &= \sum_{\substack{W \in pwd(R^p) \\ S' = top_{k,s}(W)}} Pr(W) \\
&\geq \sum_{\substack{W \in pwd(R^p) \\ S = top_{k,s}(W)}} Pr((W \backslash \{t_j\}) \cup \{t_i\}) \\
&= \sum_{\substack{W \in pwd(R^p) \\ S = top_{k,s}(W)}} Pr(W)\frac{p(t_i)\bar{p}(t_j)}{\bar{p}(t_i)p(t_j)} \\
&= \frac{p(t_i)\bar{p}(t_j)}{\bar{p}(t_i)p(t_j)} \sum_{\substack{W \in pwd(R^p) \\ S = top_{k,s}(W)}} Pr(W) \\
&= \frac{p(t_i)\bar{p}(t_j)}{\bar{p}(t_i)p(t_j)} Q_{k,s}(S) \\
&> Q_{k,s}(S),
\end{aligned}$$

which is a contradiction.
General Probabilistic Relations
The following is a counterexample.
Say $k = 2$, $R = \{t_1, t_2, t_3, t_4\}$, $t_1 \succ_s t_2 \succ_s t_3 \succ_s t_4$, t_1 and t_2 are exclusive, t_3 and t_4 are exclusive. $p(t_1) = 0.5$, $p(t_2) = 0.45$, $p(t_3) = 0.4$, $p(t_4) = 0.3$.
By U-Top$k$, the top-2 answer is $\{t_1, t_3\}$, while $t_2 \succ_s t_3$ and $p(t_2) > p(t_3)$, which violates *Faithfulness*.

(8) U-kRanks violates *Faithfulness*.
The following is a counterexample.
Say $k = 2$, R^p is simple. $R = \{t_1, t_2, t_3\}$, $t_1 \succ_s t_2 \succ_s t_3$, $p(t_1) = 0.48, p(t_2) = 0.8, p(t_3) = 0.78$.
The probabilities of each tuple at each rank are as follows:

	t_1	t_2	t_3
rank 1	0.48	0.416	0.08112
rank 2	0	0.384	0.39936
rank 3	0	0	0.29952

By U-kRanks, the top-2 answer set is $\{t_1, t_3\}$ while $t_2 \succ t_3$ and $p(t_2) > p(t_3)$, which contradicts *Faithfulness*.

(9) Global-Top$k$ satisfies *Stability*.

Proof. In the rest of this proof, let A be the set of all winners under the Global-Top$k$ semantics.

Part I: Probability.

Case 1: Winners.

For any winner $t \in A$, if we only raise the probability of t, we have a new probabilistic relation $(R^p)' = \langle R, p', \mathcal{C} \rangle$, where the new probability function p' is such that $p'(t) > p(t)$ and for any $t' \in R, t' \neq t, p'(t') = p(t')$. Note that $pwd(R^p) = pwd((R^p)')$. In addition, assume $t \in C_t$, where $C_t \in \mathcal{C}$. By Global-Top$k$,

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

and

$$\begin{aligned} P_{k,s}^{(R^p)'}(t) &= \sum_{\substack{W \in pwd(R^p) \\ t \in all_{k,s}(W)}} \alpha(t, W) Pr(W) \frac{p'(t)}{p(t)} \\ &= \frac{p'(t)}{p(t)} P_{k,s}^{R^p}(t). \end{aligned}$$

For any other tuple $t' \in R, t' \neq t$, we have the following equation:

$$
\begin{aligned}
P_{k,s}^{(R^p)'}(t') = & \sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W), t\in W}} \alpha(t',W)Pr(W)\frac{p'(t)}{p(t)} \\
& + \sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W),\ t\notin W\\ (C_t\setminus\{t\})\cap W=\emptyset}} \alpha(t',W)Pr(W)\frac{c-p'(t)}{c-p(t)} \\
& + \sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W),\ t\notin W\\ (C_t\setminus\{t\})\cap W\neq\emptyset}} \alpha(t',W)Pr(W) \\
\leq & \frac{p'(t)}{p(t)}(\sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W)\\ t\in W}} \alpha(t',W)Pr(W) \\
& + \sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W),\ t\notin W\\ (C_t\setminus\{t\})\cap W=\emptyset}} \alpha(t',W)Pr(W) \\
& + \sum_{\substack{W\in pwd(R^p)\\ t'\in all_{k,s}(W),\ t\notin W\\ (C_t\setminus\{t\})\cap W\neq\emptyset}} \alpha(t',W)Pr(W)) \\
= & \frac{p'(t)}{p(t)}P_{k,s}^{R^p}(t'),
\end{aligned}
$$

where $c = 1 - \sum_{t''\in C_t\setminus\{t\}} p(t'')$.

Now we can see that, t's Global-Top$k$ probability in $(R^p)'$ will be raised to *exactly* $\frac{p'(t)}{p(t)}$ times of that in R^p under the same weak order scoring function s, and for any tuple other than t, its Global-Top$k$ probability in $(R^p)'$ can be raised to *as much as* $\frac{p'(t)}{p(t)}$ times of that in R^p under the same scoring function s. As a result, $P_{k,s}^{(R^p)'}(t)$ is still among the highest k Global-Top$k$ probabilities in $(R^p)'$ under the function s, and therefore still a winner.

Case 2: Losers.

This case is similar to *Case 1*.

Part II: Score.

Case 1: Winners.

For any winner $t \in A$, we evaluate R^p under a new general scoring function s'. Comparing to s, s' only raises the score of t. That is, $s'(t) > s(t)$ and for any $t' \in R, t' \neq t, s'(t') = s(t')$. Then, in addition to all the worlds already *totally* (i.e. $\alpha(t,W) = 1$) or *partially* (i.e. $\alpha(t,W) < 1$) contributing to t's Global-Top$k$ probability when evaluating R^p under s, some other worlds may now totally or partially contribute to t's Global-Top$k$ probability. Because, under the function s',

t might climb high enough to be in the top-k answer set of those worlds. Moreover, if a possible world W contributes paritally under scoring function s, it is easy to see that it contributes totally under scoring function s'.

For any tuple t'' other than t in R,

(i) If $s(t'') \neq s(t)$, then its Global-Top$k$ probability under the function s' either stays the same (if the "climbing" of t does not knock that tuple out of the top-k answer in some possible world) or decreases (otherwise);

(ii) If $s(t'') = s(t)$, then for any possible world W contributing to t'''s Global-Top$k$ under scoring function s, $\alpha(t'', W) = \frac{k-a}{b}$, and now under scoring function s', $\alpha'(t'', W) = \frac{k-a-1}{b-1} < \frac{k-a}{b} = \alpha(t'', W)$. Therefore the Global-Top$k$ of t'' under scoring function s' is less than that under scoring function s.

Consequently, t is still a winner when evaluating R^p under the function s'.

Case 2: Losers.

This case is similar to *Case 1*.

(10) PT-k satisfies *Stability*.

Proof. In the rest of this proof, let A be the set of all winners under the PT-k semantics.

Part I: Probability.

Case 1: Winners.

For any winner $t \in A$, if we only raise the probability of t, we have a new probabilistic relation $(R^p)' = \langle R, p', \mathcal{C} \rangle$, where the new probability function p' is such that $p'(t) > p(t)$ and for any $t' \in R, t' \neq t, p'(t') = p(t')$. Note that $pwd(R^p) = pwd((R^p)')$. In addition, assume $t \in C_t$, where $C_t \in \mathcal{C}$. The Global-Top$k$ probability of t is such that

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

and

$$\begin{aligned} P^{(R^p)'}_{k,s}(t) &= \sum_{\substack{W \in pwd(R^p) \\ t \in top_{k,s}(W)}} Pr(W) \frac{p'(t)}{p(t)} \\ &= \frac{p'(t)}{p(t)} P^{R^p}_{k,s}(t) > P^{R^p}_{k,s}(t) \geq p_\tau. \end{aligned}$$

Therefore, $P^{(R^p)'}_{k,s}(t)$ is still above the threshold p_τ, and t still belongs to the top-k answer of $(R^p)'$ under the function s.

Case 2: Losers.

This case is similar to *Case 1*.

Part II: Score.

Case 1: Winners.

For any winner $t \in A$, we evaluate R^p under a new scoring function s'. Comparing to s, s' only raises the score of t. Use a similar argument as that in (9) Part II

Case 1 but under injective scoring functions, we can show that the Global-Top$k$ probability of t is non-decreasing and is still above the threshold p_τ. Therefore, tuple t still belongs to the top-k answer under the function s'.
Case 2: Losers.
This case is similar to *Case 1*.

(11) U-Top$k$ satisfies *Stability*.
Proof. In the rest of this proof, let A be the set of all winners under U-Top$k$ semantics.
Part I: Probability.
Case 1: Winners.
For any winner $t \in A$, if we only raise the probability of t, we have a new probabilistic relation $(R^p)' = \langle R, p', \mathcal{C} \rangle$, where the new probabilistic function p' is such that $p'(t) > p(t)$ and for any $t' \in R, t' \neq t, p'(t') = p(t')$. In the following discussion, we use superscript to indicate the probability in the context of $(R^p)'$. Note that $pwd(R^p) = pwd((R^p)')$.
Recall that $Q_{k,s}(A_t)$ is the probability of a top-k answer set $A_t \subseteq A$ under U-Top$k$ semantics, where $t \in A_t$. Since $t \in A_t$, $Q'_{k,s}(A_t) = Q_{k,s}(A_t)\frac{p'(t)}{p(t)}$.
For any candidate top-k set B other than A_t, i.e. $\exists W \in pwd(R^p), top_{k,s}(W) = B$ and $B \neq A_t$. By definition,

$$Q_{k,s}(B) \leq Q_{k,s}(A_t).$$

For any world W contributing to $Q_{k,s}(B)$, its probability either increase $\frac{p'(t)}{p(t)}$ times (if $t \in W$), or stays the same (if $t \notin W$ and $\exists t' \in W, t'$ and t are exclusive), or decreases (otherwise). Therefore,

$$Q'_{k,s}(B) \leq Q_{k,s}(B)\frac{p'(t)}{p(t)}.$$

Altogether,

$$Q'_{k,s}(B) \leq Q_{k,s}(B)\frac{p'(t)}{p(t)} \leq Q_{k,s}(A_t)\frac{p'(t)}{p(t)} = Q'_{k,s}(A_t).$$

Therefore, A_t is still a top-k answer to $(R^p)'$ under the function s and $t \in A_t$ is still a winner.
Case 2: Losers.
It is more complicated in the case of losers. We need to show that for any loser t, if we decrease its probability, no top-k candidate set B_t containing t will be a new top-k answer set under the U-Top$k$ semantics. The procedure is similar to that in *Case 1*, except that when we analyze the new probability of any original top-k answer set A_i, we need to differentiate between two cases:
(a) t is exclusive with some tuple in A_i;
(b) t is independent of all the tuples in A_i.
It is easier with (a), where all the worlds contributing to the probability of A_i do not contain t. In (b), some worlds contributing to the probability of A_i contain t, while others do not. And we calculate the new probability for those two kinds of

worlds differently. As we will see shortly, the probability of A_i stays unchanged in either (a) or (b).

For any loser $t \in R, t \notin A$, by applying the technique used in *Case 1*, we have a new probabilistic relation $(R^p)' = \langle R, p', \mathcal{C} \rangle$, where the new probabilistic function p' is such that $p'(t) < p(t)$ and for any $t' \in R, t' \neq t, p'(t') = p(t')$. Again, $pwd(R^p) = pwd((R^p)')$.

For any top-k answer set A_i to R^p under the function s, $A_i \subseteq A$. Denote by S_{A_i} all the possible worlds contributing to $Q_{k,s}(A_i)$. Based on the membership of t, S_{A_i} can be partitioned into two subsets $S^t_{A_i}$ and $S^{\bar{t}}_{A_i}$.

$$S_{A_i} = \{W | W \in pwd(R^p), top_{k,s}(W) = A_i\};$$
$$S_{A_i} = S^t_{A_i} \cup S^{\bar{t}}_{A_i}, S^t_{A_i} \cap S^{\bar{t}}_{A_i} = \emptyset,$$
$$\forall W \in S^t_{A_i}, t \in W \text{ and } \forall W \in S^{\bar{t}}_{A_i}, t \notin W.$$

If t is exclusive with some tuple in A_i, $S^t_{A_i} = \emptyset$. In this case, any world $W \in S^{\bar{t}}_{A_i}$ contains one of t's exclusive tuples, therefore W's probability will not be affected by the change in t's probability. In this case,

$$\begin{aligned} Q'_{k,s}(A_i) &= \sum_{\substack{W \in pwd(R^p) \\ W \in S^{\bar{t}}_{A_i}}} Pr'(W) = \sum_{\substack{W \in pwd(R^p) \\ W \in S^{\bar{t}}_{A_i}}} Pr(W) \\ &= Q_{k,s}(A_i). \end{aligned}$$

Otherwise, t is independent of all the tuples in A_i. In this case,

$$\frac{\sum_{\substack{W \in pwd(R^p) \\ W \in S^t_{A_i}}} Pr(W)}{\sum_{\substack{W \in pwd(R^p) \\ W \in S^{\bar{t}}_{A_i}}} Pr(W)} = \frac{p(t)}{1 - p(t)}$$

and

$$\begin{aligned} Q'_{k,s}(A_i) = &\sum_{\substack{W \in pwd(R^p) \\ W \in S^t_{A_i}}} Pr(W) \frac{p'(t)}{p(t)} \\ &+ \sum_{\substack{W \in pwd(R^p) \\ W \in S^{\bar{t}}_{A_i}}} Pr(W) \frac{1 - p'(t)}{1 - p(t)} \\ = &\sum_{\substack{W \in pwd(R^p) \\ W \in S_{A_i}}} Pr(W) \\ = &\ Q_{k,s}(A_i). \end{aligned}$$

We can see that in both cases, $Q'_{k,s}(A_i) = Q_{k,s}(A_i)$.

Now for any top-k candidate set containing t, say B_t such that $B_t \not\subseteq A$, by definition, $Q_{k,s}(B_t) < Q_{k,s}(A_i)$. Moreover,

$$Q'_{k,s}(B_t) = Q_{k,s}(B_t)\frac{p'(t)}{p(t)} < Q_{k,s}(B_t).$$

Therefore,

$$Q'_{k,s}(B_t) < Q_{k,s}(B_t) < Q_{k,s}(A_i) = Q'_{k,s}(A_i).$$

Consequently, B_t is still not a top-k answer to $(R^p)'$ under the function s. Since no top-k candidate set containing t can be a top-k answer set to $(R^p)'$ under the function s, t is still a loser.

Part II: Score.

Again, $A_i \subseteq A$ is a top-k answer set to R^p under the function s by U-Top$k$ semantics.

Case 1: Winners.

For any winner $t \in A_i$, we evaluate R^p under a new scoring function s'. Comparing to s, s' only raises the score of t. That is, $s'(t) > s(t)$ and for any $t' \in R, t' \neq t, s'(t') = s(t')$. In some possible world such that $W \in pwd(R^p)$ and $top_{k,s}(W) \neq A_i$, t might climb high enough to be in $top_{k,s'}(W)$. Define T to the set of such top-k candidate sets.

$$T = \{top_{k,s'}(W) | W \in pwd(R^p), t \notin top_{k,s}(W) \wedge t \in top_{k,s'}(W)\}.$$

Only a top-k candidate set $B_j \in T$ can possibly end up with a probability higher than that of A_i across all possible worlds, and thus substitute for A_i as a new top-k answer set to R^p under the function s'. In that case, $t \in B_j$, so t is still a winner.

Case 2: Losers.

For any loser $t \in R, t \notin A$. Using a similar technique to *Case 1*, the new scoring function s' is such that $s'(t) < s(t)$ and for any $t' \in R, t' \neq t, s'(t') = s(t')$. When evaluating R^p under the function s', for any world $W \in pwd(R^p)$ such that $t \notin top_{k,s}(W)$, the score decrease of t will not effect its top-k answer, i.e. $top_{k,s'}(W) = top_{k,s}(W)$. For any world $W \in pwd(R^p)$ such that $t \in top_{k,s}(W)$, t might go down enough to drop out of $top_{k,s'}(W)$. In this case, W will contribute its probability to a top-k candidate set without t, instead of the original one with t. In other words, under the function s', comparing to the evaluation under the function s, the probability of a top-k candidate set with t is non-increasing, while the probability of a top-k candidate set without t is non-decreasing[2].

Since any top-k answer set to R^p under the function s does not contain t, it follows from the above analysis that any top-k candidate set containing t will not be a top-k answer set to R^p under the new function s', and thus t is still a loser.

[2] Here, any subset of R with cardinality at most k that is not a top-k candidate set under the function s is conceptually regarded as a top-k candidate set with probability zero under the function s.

(12) U-kRanks violates *Stability*.
The following is a counterexample.
Say $k = 2$, R^p is simple. $R = \{t_1, t_2, t_3\}$, $t_1 \succ_s t_2 \succ_s t_3$. $p(t_1) = 0.3, p(t_2) = 0.4, p(t_3) = 0.3$.

	t_1	t_2	t_3
rank 1	0.3	0.28	0.126
rank 2	0	0.12	0.138
rank 3	0	0	0.036

By U-kRanks, the top-2 answer set is $\{t_1, t_3\}$.
Now raise the score of t_3 such that $t_1 \succ_{s'} t_3 \succ_{s'} t_2$.

	t_1	t_3	t_2
rank 1	0.3	0.21	0.196
rank 2	0	0.09	0.168
rank 3	0	0	0.036

By U-kRanks, the top-2 answer set is $\{t_1, t_2\}$. By raising the score of t_3, we actually turn the winner t_3 to a loser, which contradicts *Stability*.

9.2 Proof for Proposition 1

Proposition 1. *Given a simple probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$ and an injective scoring function s over R^p, if $R = \{t_1, t_2, \ldots, t_n\}$ and $t_1 \succ_s t_2 \succ_s \ldots \succ_s t_n$, the following recursion on Global-Top$k$ queries holds.*

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

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

Proof. By induction on k and i.

– Base case.
  - $k = 0$
    For any $W \in pwd(R^p)$, $top_{0,s}(W) = \emptyset$. Therefore, for any $t_i \in R$, the Global-Top$k$ probability of t_i is 0.
  - $k > 0$ and $i = 1$
    t_1 has the highest score among all tuples in R. As long as tuple t_1 appears in a possible world W, it will be in the $top_{k,s}(W)$. So the Global-Top$k$ probability of t_i is the probability that t_1 appears in possible worlds, i.e. $q(k,1) = p(t_1)$.

– Inductive step.
  Assume the theorem holds for $0 \leq k \leq k_0$ and $1 \leq i \leq i_0$. For any $W \in pwd(R^p)$, $t_{i_0} \in top_{k_0,s}(W)$ iff $t_{i_0} \in W$ and there are at most $k_0 - 1$ tuples with higher score in W. Note that any tuple with score lower than the score of t_{i_0} does not have any

influence on $q(k_0, i_0)$, because its presence/absence in a possible world will not affect the presence of t_{i_0} in the top-k answer of that world.
Since all the tuples are independent,

$$q(k_0, i_0) = p(t_{i_0}) \sum_{\substack{W \in pwd(R^p) \\ |\{t | t \in W \wedge t \succ_s t_{i_0}\}| < k_0}} Pr(W).$$

(1) $q(k_0, i_0 + 1)$ is the Global-Topk_0 probability of tuple t_{i_0+1}.

$$\begin{aligned} q(k_0, i_0 + 1) = & \sum_{\substack{W \in pwd(R^p) \\ t_{i_0+1} \in top_{k_0,s}(W) \\ t_{i_0} \in top_{k_0,s}(W)}} Pr(W) \\ & + \sum_{\substack{W \in pwd(R^p) \\ t_{i_0+1} \in top_{k_0,s}(W) \\ t_{i_0} \in W,\, t_{i_0} \notin top_{k_0,s}(W)}} Pr(W) \\ & + \sum_{\substack{W \in pwd(R^p) \\ t_{i_0+1} \in top_{k_0,s}(W) \\ t_{i_0} \notin W}} Pr(W). \end{aligned}$$

For the first part of the left hand side,

$$\sum_{\substack{W \in pwd(R^p) \\ t_{i_0+1} \in top_{k_0,s}(W) \\ t_{i_0} \in top_{k_0-1,s}(W)}} Pr(W) = p(t_{i_0+1}) q(k_0 - 1, i_0).$$

The second part is zero. Since $t_{i_0} \succ_s t_{i_0+1}$, if $t_{i_0+1} \in top_{k_0,s}(W)$ and $t_{i_0} \in W$, then $t_{i_0} \in top_{k_0,s}(W)$.
The third part is the sum of the probabilities of all possible worlds such that $t_{i_0+1} \in W, t_{i_0} \notin W$ and there are at most $k_0 - 1$ tuples with score higher than the score of t_{i_0} in W. So it is equivalent to

$$\begin{aligned} & p(t_{i_0+1}) \overline{p}(t_{i_0}) \sum_{|\{t | t \in W \wedge t \succ_s t_{i_0}\}| < k_0} Pr(W) \\ = & p(t_{i_0+1}) \overline{p}(t_{i_0}) \frac{q(k_0, i_0)}{p(t_{i_0})}. \end{aligned}$$

Altogehter, we have

$$\begin{aligned} & q(k_0, i_0 + 1) \\ = & p(t_{i_0+1}) q(k_0 - 1, i_0) + p(t_{i_0+1}) \overline{p}(t_{i_0}) \frac{q(k_0, i_0)}{p(t_{i_0})} \\ = & (q(k_0 - 1, i_0) + q(k_0, i_0) \frac{\overline{p}(t_{i_0})}{p(t_{i_0})}) p(t_{i_0+1}). \end{aligned}$$

(2) $q(k_0+1, i_0)$ is the Global-Top(k_0+1) probability of tuple t_{i_0}. Use a similar argument as above, it can be shown that this case is correctly computed by Equation (3) as well.

9.3 Proof for Theorem 2

Theorem 2 (Correctness of Algorithm 1'). *Given a simple probabilistic relation $R^p = \langle R, p, \mathcal{C}\rangle$, a non-negative integer k and an injective scoring function s over R^p, the above TA-based algorithm correctly finds a top-k answer under Global-Top$k$ semantics.*

Proof. In every iteration of Step (2), say $\underline{t} = t_i$, for any unseen tuple t, s' is an injective scoring function over R^p, which only differs from s in the score of t. Under the function s', $t_i \succ_{s'} t \succ_{s'} t_{i+1}$. If we evaluate the top-k query in R^p under s' instead of s, $P_{k,s'}(t) = \frac{p(t)}{\underline{p}}UP$. On the other hand, for any $W \in pwd(R^p)$, W contributing to $P_{k,s}(t)$ implies that W contributes to $P_{k,s'}(t)$, while the reverse is not necessarily true. So, we have $P_{k,s'}(t) \geq P_{k,s}(t)$. Recall that $\underline{p} \geq p(t)$, therefore $UP \geq \frac{p(t)}{\underline{p}}UP = P_{k,s'}(t) \geq P_{k,s}(t)$. The conclusion follows from the correctness of the original TA algorithm and Algorithm 1.

9.4 Proof for Lemma 1

Lemma 1. *Let $R^p = \langle R, p, \mathcal{C}\rangle$ be a probabilistic relation, s an injective scoring function, $t \in R$, and $E^p = \langle E, p^E, \mathcal{C}^E\rangle$ the event relation induced by t. Define $Q^p = \langle E - \{t_{e_t}\}, p^E, \mathcal{C}^E - \{\{t_{e_t}\}\}\rangle$. Then, the Global-Top$k$ probability of t satisfies the following:*

$$P_{k,s}^{R^p}(t) = p(t) \sum_{\substack{W_e \in pwd(Q^p)\\ |W_e| < k}} Pr(W_e).$$

Proof. Given $t \in R$, k and s, let A be a subset of $pwd(R^p)$ such that $W \in A \Leftrightarrow t \in top_{k,s}(W)$. If we group all the possible worlds in A by the set of parts whose tuple in W has higher score than the score of t, then we will have the following partition:

$$A = A_1 \cup A_2 \cup \ldots \cup A_q, A_i \cap A_j = \emptyset, i \neq j$$

and

$$\forall A_i, \forall W_1, W_2 \in A_i, i = 1, 2, \ldots, q,$$
$$\{C_j | \exists t' \in W_1 \cap C_j, t' \succ_s t\} = \{C_j | \exists t' \in W_2 \cap C_j, t' \succ_s t\}.$$

Moreover, denote $CharParts(A_i)$ to A_i's characteristic set of parts.

Now, let B be a subset of $pwd(Q^p)$, such that $W_e \in B \Leftrightarrow |W_e| < k$. There is a bijection $g : \{A_i | A_i \in A\} \rightarrow B$, mapping each part A_i in A to a possible world in B which contains only tuples corresponding to the parts in A_i 's characteristic set.

$$g(A_i) = \{t_{e_{C_j}} | C_j \in CharParts(A_i)\}.$$

The following equation holds from the definition of induced event relation and Proposition 2.

$$\sum_{W \in A_i} Pr(W) = p(t) \prod_{C_i \in CharParts(A_i)} p(t_{e_{C_i}}) \prod_{\substack{C_i \in \mathcal{C} - \{C_{id(t)}\} \\ C_i \notin CharParts(A_i)}} (1 - p(t_{e_{C_i}}))$$
$$= p(t) Pr(g(A_i)).$$

Therefore,

$$\begin{aligned} P_{k,s}^{R^p}(t) &= \sum_{W \in A} Pr(W) = \sum_{i=1}^{q} (\sum_{W \in A_i} Pr(W)) \\ &= \sum_{i=1}^{q} p(t) Pr(g(A_i)) = p(t) \sum_{i=1}^{q} Pr(g(A_i)) \\ &= p(t) \sum_{W_e \in B} Pr(W_e) \\ &= p(t) (\sum_{\substack{W_e \in pwd(Q^p) \\ |W_e| < k}} Pr(W_e)). \end{aligned}$$

9.5 Proof for Proposition 3

Proposition 3 (Correctness of Algorithm 4). *Given a probabilistic relation* $R^p = \langle R, p, \mathcal{C} \rangle$ *and an injective scoring function* s*, for any* $t \in R^p$*, the Global-Top*k *probability of* t *equals the Global-Top*k *probability of* t_{e_t} *when evaluating top-*k *in the induced event relation* $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ *under the injective scoring function* $s^E : E \to \mathbb{R}, s^E(t_{e_t}) = \frac{1}{2}$ *and* $s^E(t_{e_{C_i}}) = i$*:*

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

Proof. Since t_{e_t} has the lowest score under s^E, for any $W_e \in pwd(E^p)$, the only chance $t_{e_t} \in top_{k,s^E}(W_e)$ is when there are at most k tuples in W_e, including t_{e_t}.

$$\begin{aligned} &\forall W_e \in pwd(E^p), \\ &t_{e_t} \in top_{k,s}(W_e) \Leftrightarrow (t_{e_t} \in W_e \wedge |W_e| \le k). \end{aligned}$$

Therefore,

$$P_{k,s^E}^{E^p}(t_{e_t}) = \sum_{t_{e_t} \in W_e \wedge |W_e| \le k} Pr(W_e).$$

In the proof of Lemma 1, B contains all the possible worlds having at most $k-1$ tuples from $E - \{t_{e_t}\}$. By Proposition 2,

$$\sum_{t_{e_t} \in W_e \wedge |W_e| \le k} Pr(W_e) = p(t) \sum_{W'_e \in B} Pr(W'_e).$$

By Lemma 1,

$$p(t) \sum_{W'_e \in B} Pr(W'_e) = P^{R^p}_{k,s}(t).$$

Consequently,

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

9.6 Proof for Proposition 4

Proposition 4 (Correctness of Algorithm 5). *Let $R^p = \langle R, p, \mathcal{C} \rangle$ be a simple probabilistic relation where $R = \{t_1, \ldots, t_n\}$, $t_1 \succeq_s t_2 \succeq_s \ldots \succeq_s t_n$, k a non-negative integer and s a scoring function. For every $t_l \in R$, the Global-Top$k$ probability of t_l can be computed by the following equation:*

$$P^{R^p}_{k,s}(t_l) = \sum_{k'=0}^{k-1} T_{k',[i_l]} \cdot P^{R^p_s(t_l)}_{k-k',s}(t_l)$$

where $R^p_s(t_l)$ is R^p restricted to $\{t \in R | t \sim_s t_l\}$.

Proof. Given a tuple $t_l \in R$, let R_θ be the support relation R restricted to $\{t \in R | t \, \theta \, t_l\}$, and R^p_θ be R^p restricted to R_θ. Similarly, for each possible world $W \in pwd(R^p)$, $W_\theta = W \cap R_\theta$.

Each possible world $W \in pwd(R^p)$ such that $t_l \in all_{k,s}(W)$ contributes $\min(1, \frac{k-a}{b}) Pr(W)$ to $P^{R^p}_{k,s}(t_l)$, where $a = |W_\succ|$ and $b = |W_\sim|$.

$$
\begin{aligned}
P_{k,s}^{R^p}(t_l) &= \sum_{\substack{W \in pwd(R^p), t_l \in W \\ |W_\succ| = a, 0 \le a \le k-1 \\ |W_\sim| = b, 1 \le b \le m}} \min(1, \frac{k-a}{b}) Pr(W) \\
&= \sum_{a=0}^{k-1} \sum_{b=1}^{m} \min(1, \frac{k-a}{b}) (\sum_{\substack{W \in pwd(R^p), t_l \in W \\ |W_\succ| = a \wedge |W_\sim| = b}} Pr(W)) \\
&= \sum_{a=0}^{k-1} \sum_{b=1}^{m} \min(1, \frac{k-a}{b}) (\sum_{\substack{W_\succ \in pwd(R^p_\succ) \\ |W_\succ| = a}} Pr(W_\succ) \sum_{\substack{W_\preceq \in pwd(R^p_\preceq), t_l \in W_\preceq \\ |W_\sim| = b}} Pr(W_\preceq)) \\
&= \sum_{a=0}^{k-1} (\sum_{\substack{W_\succ \in pwd(R^p_\succ) \\ |W_\succ| = a}} Pr(W_\succ) \sum_{b=1}^{m} \min(1, \frac{k-a}{b}) (\sum_{\substack{W_\preceq \in pwd(R^p_\preceq), t_l \in W_\preceq \\ |W_\sim| = b}} Pr(W_\preceq))) \\
&= \sum_{a=0}^{k-1} (T_{a,[i_l]} \sum_{b=1}^{m} \min(1, \frac{k-a}{b}) (\sum_{\substack{W_\sim \in pwd(R^p_\sim), t_l \in W_\sim \\ |W_\sim| = b}} Pr(W_\sim) \sum_{W_\prec \in pwd(R^p_\prec)} Pr(W_\prec))) \\
&= \sum_{a=0}^{k-1} (T_{a,[i_l]} \sum_{b=1}^{m} \min(1, \frac{k-a}{b}) (\sum_{\substack{W_\sim \in pwd(R^p_\sim), t_l \in W_\sim \\ |W_\sim| = b}} Pr(W_\sim))) \\
&= \sum_{a=0}^{k-1} T_{a,[i_l]} \cdot P_{k-a,s}^{R_s^p(t_l)}(t_l)
\end{aligned}
$$

where m is the number of tying tuples with t_l(including), i.e. $m = |R_s^p(t_l)|$.

9.7 Proof for Proposition 5

Proposition 5. *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$ and a scoring function s, for any $t \in R^p$, the Global-Top$k$ probability of t equals the Global-Top$k$ probability of $t_{e_t,\sim}$ when evaluating top-k in the induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$ under the scoring function $s^E : E \to \mathbb{R}$, $s^E(t_{e_t}) = \frac{1}{2}$, $s^E(t_{e_t,\sim}) = \frac{1}{2}$ and $s^E(t_{e_{C_i},\succ}) = i$:*

$$P_{k,s}^{R^p}(t) = P_{k,s^E}^{E^p}(t_{e_t,\sim}).$$

Proof. Similar to what we did in the Proof for Lemma 1. We are trying to create a bijection.

Given $t \in R$, k and s, let A be a subset of $pwd(R^p)$ such that $W \in A \Leftrightarrow t \in all_{k,s}(W)$. If we group all the possible worlds in A by the set of parts whose tuple in W has score higher than or equal to that of t, then we will have the following partition:

$$A = A_1 \cup A_2 \cup \ldots \cup A_q, A_i \cap A_j = \emptyset, i \neq j$$

and

$$\forall A_i, \forall W_1, W_2 \in A_i, i = 1, 2, \ldots, q,$$
$$\{C_{j,\succ} | \exists t' \in W_1 \cap C_j, t' \succ_s t\} = \{C_{j,\succ} | \exists t' \in W_2 \cap C_j, t' \succ_s t\}$$
and
$$\{C_{j,\sim} | \exists t' \in W_1 \cap C_j, t' \sim_s t\} = \{C_{j,\sim} | \exists t' \in W_2 \cap C_j, t' \sim_s t\}.$$

Moreover, denote $CharParts(A_i)$ to A_i's characteristic set of parts. Note that all $W \in A_i$ have the same allocation coefficient $\alpha(t, W)$, denoted by α_i.

Now, let B be a subset of $pwd(E^p)$, such that $W_e \in B \Leftrightarrow t_{e_t,\sim} \in all_{k,s}(W_e)$. There is a bijection $g : \{A_i | A_i \in A\} \rightarrow B$, mapping each part A_i in A to the a possible world in B which contains only tuples corresponding to parts in A_i 's characteristic set.

$$g(A_i) = \{t_{e_{C_j},\succ} | C_{j,\succ} \in CharParts(A_i)\} \cup \{t_{e_{C_j},\sim} | C_{j,\sim} \in CharParts(A_i)\}$$

Furthermore, the allocation coefficient α_i of A_i equals to the allocation coefficient $\alpha(t_{e_t,\sim}, g(A_i))$ under the function s^E.

The following equation holds from the definition of induced event relation under general scoring functions.

$$\begin{aligned}\sum_{W \in A_i} Pr(W) = & \prod_{C_{i,\succ} \in CharParts(A_i)} p(t_{e_{C_i},\succ}) \prod_{C_{i,\sim} \in CharParts(A_i)} p(t_{e_{C_i},\sim}) \\ & \prod_{\substack{C_i \in \mathcal{C} \\ C_{i,\sim} \notin CharParts(A_i) \\ C_{i,\succ} \notin CharParts(A_i)}} (1 - p(t_{e_{C_i},\succ}) - p(t_{e_{C_i},\sim})) \\ = & Pr(g(A_i)).\end{aligned}$$

Therefore,

$$\begin{aligned}P_{k,s}^{R^p}(t) &= \sum_{W \in A} \alpha(t, W) Pr(W) = \sum_{i=1}^{q} (\alpha_i \sum_{W \in A_i} Pr(W)) \\ &= \sum_{i=1}^{q} \alpha_i Pr(g(A_i)) = \sum_{i=1}^{q} \alpha(t_{e_t,\sim}, g(A_i)) Pr(g(A_i)) \\ &= \sum_{W_e \in B} \alpha(t_{e_t,\sim}, W_e) Pr(W_e) \quad (g \text{ is a bijection}) \\ &= P_{k,s^E}^{E^p}(t_{e_t,\sim}).\end{aligned}$$

9.8 Proof for Theorem 5

Theorem 5. *Given a probabilistic relation $R^p = \langle R, p, \mathcal{C} \rangle$, a scoring function s, $t \in R^p$, and its induced event relation $E^p = \langle E, p^E, \mathcal{C}^E \rangle$, where $|E| = 2m$, the following recursion on $u_\succ(k', i, b)$ and $u_\sim(k', i, b)$ holds, where $b_{\max}$ is the number of tuples with positive probability in $E^p_\sim$.*

When $i = 1, 0 \leq k' \leq m$ *and* $0 \leq b \leq b_{\max}$,

$$u_{\succ}(k', i, b) = \begin{cases} 0 & k' = 0 \\ (u_{\succ}(k', i-1, b)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\succ})} & 1 \leq k' \leq m \\ \quad + u_{\succ}(k'-1, i-1, b) & \textit{and } p^E(t_{i-1,\succ}) > 0 \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \\ (u_{\sim}(k', i-1, b+1)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\sim})} & b < b_{\max} \\ \quad + u_{\succ}(k'-1, i-1, b) & \textit{and } 1 \leq k' \leq m \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \textit{and } p^E(t_{i-1,\succ}) = 0 \\ (u_{\succ}(k'-1, i-1, b) & \textit{otherwise} \\ \quad + u_{\sim}(k'-1, i-1, b))p^E(t_{i,\succ}) & \end{cases}$$

$$u_{\sim}(k', i, b) = \begin{cases} 0 & k' = 0 \textit{ or } b = 0 \\ (u_{\sim}(k', i-1, b)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\sim})} & b > 0 \\ \quad + u_{\succ}(k'-1, i-1, b-1) & \textit{and } 1 \leq k' \leq m \\ \quad + u_{\sim}(k'-1, i-1, b-1))p^E(t_{i,\sim}) & \textit{and } p^E(t_{i-1,\sim}) > 0 \\ (u_{\succ}(k', i-1, b-1)\dfrac{1 - p^E(t_{i-1,\succ}) - p^E(t_{i-1,\sim})}{p^E(t_{i-1,\succ})} & \textit{otherwise} \\ \quad + u_{\succ}(k'-1, i-1, b-1) & \\ \quad + u_{\sim}(k'-1, i-1, b-1))p^E(t_{i,\sim}) & \end{cases}$$

The Global-Top$k$ *probability of* $t_{e_t,\sim}$ *in* E^p *under the scoring function* s^E *can be computed by the following equation:*

$$\begin{aligned} P_{k,s^E}^{E^p}(t_{e_t,\sim}) &= P_{k,s^E}^{E^p}(t_{m,\sim}) \\ &= \sum_{b=1}^{b_{\max}}(\sum_{k'=1}^{k} u_{\sim}(k', m, b) + \sum_{k'=k+1}^{k+b-1} \frac{k-(k'-b)}{b} u_{\sim}(k', m, b)) \end{aligned}$$

Proof. Equation 9 follows Equation 7 and Equation 8 as it is a simple enumeration based on Definition 8. We are going to prove Equation 7 and Equation 8 by an induction on i.

– Base case: $i = 1, 0 \leq k' \leq m$ and $0 \leq b \leq b_{\max}$
  When $i = 1$, based on the definition of u, the only non-zero entries are $u_{\succ}(1, 1, 0)$ and $u_{\sim}(1, 1, 1)$. The former is the probability sum of all possible worlds which contain $t_{1,\succ}$ and do not contain $t_{1,\sim}$. The second requirement is redundant since those two tuples are exclusive. Therefore, it is simply the probability of $t_{1,\succ}$. Similarly, the latter is the probability sum of all possible worlds which contain $t_{1,\sim}$ and do not contain $t_{1,\succ}$. Again, it is simply the probability of $t_{1,\sim}$. It is easy to check that no possible worlds satisfy other combinations of k' and b when $i = 1$, therefore their probabilities are 0.

– Inductive step.
Assume the theorem holds for $i \le i_0$, $0 \le k' \le m$ and $0 \le b \le b_{\max}$.
Denote $E_{\succ,[i]}$ and $E_{\sim,[i]}$ to the set of the first i tuples in $E_\succ$ and $E_\sim$ respectively.
For any $W \in pwd(E^p)$, by definition, W contributes to $u_{\succ/\sim}(k', i_0, b)$ iff $t_{i_0,\succ/\sim} \in W$ and $|W \cap (E_{\succ,[i_0]} \cup E_{\sim,[i_0]})| = k'$ and $|W \cap E_{\sim,[i_0]}| = b$. Since $E_{\succ,[i_0]} \cap E_{\sim,[i_0]} = \emptyset$, we have:
W contributes to $u_{\succ/\sim}(k', i_0, b) \Leftrightarrow t_{i_0,\succ/\sim} \in W$ and $|W \cap E_{\succ,[i_0]}| = k' - b$ and $|W \cap E_{\sim,[i_0]}| = b$.
(1) $u_\succ(k', i_0 + 1, b)$ is the probability sum of all possible world W such that $t_{i_0+1,\succ} \in W$, $|W \cap E_{\succ,[i_0+1]}| = k' - b$ and $|W \cap E_{\sim,[i_0+1]}| = b$.

$$
\begin{aligned}
u_\succ(k', i_0+1, b) &= \sum_{\substack{W \in pwd(E^p), t_{i_0+1,\succ} \in W \\ |W \cap E_{\succ,[i_0+1]}| = k'-b \\ |W \cap E_{\sim,[i_0+1]}| = b}} Pr(W) \\
&= \sum_{\substack{W \in pwd(E^p), t_{i_0+1,\succ} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) \quad \text{(Since } t_{i_0+1,\succ} \in W, \ t_{i_0+1,\sim} \notin W\text{)} \\
&= \sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W, t_{i_0,\succ} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) \\
&+ \sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W, t_{i_0,\sim} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) \\
&+ \sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W, t_{i_0,\succ} \notin W, t_{i_0,\sim} \notin W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W)
\end{aligned}
$$

For the first part of the left hand side,

$$
\sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W, t_{i_0,\succ} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) = p(t_{i_0+1}) \sum_{\substack{W \in pwd(E^p), t_{i_0,\succ} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) = p(t_{i_0+1}) u_\succ(k'-1, i_0, b).
$$

For the second part of the left hand side,

$$
\sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W, t_{i_0,\sim} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) = p(t_{i_0+1}) \sum_{\substack{W \in pwd(E^p), t_{i_0,\sim} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) = p(t_{i_0+1}) u_\sim(k'-1, i_0, b).
$$

For the third part of the left hand side, if $p(t_{i_0,\succ}) + p(t_{i_0,\sim}) = 1$, then there is no possible world satisfying this condition, therefore it is zero. Otherwise,

$$\sum_{\substack{W \in pwd(E^p) \\ t_{i_0+1,\succ} \in W \\ t_{i_0,\succ} \notin W, t_{i_0,\sim} \notin W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) = p(t_{i_0+1}) \sum_{\substack{W \in pwd(E^p) \\ t_{i_0,\succ} \notin W, t_{i_0,\sim} \notin W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) \quad (10)$$

Equation 10 can be computed either by Equation 11 when $p(t_{i_0}, \succ) > 0$ or by Equation 12 when $p(t_{i_0}, \sim) > 0$. Notice that at least one of $p(t_{i_0}, \succ)$ and $p(t_{i_0}, \sim)$ is positive, otherwise neither tuple is in the induced event relation E^p according to Definition 11.

$$\begin{aligned} \sum_{\substack{W \in pwd(E^p) \\ t_{i_0,\succ} \notin W, t_{i_0,\sim} \notin W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) &= \frac{1 - p(t_{i_0,\succ}) - p(t_{i_0,\sim})}{p(t_{i_0,\succ})} \sum_{\substack{W \in pwd(E^p), t_{i_0,\succ} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) \\ &= \frac{1 - p(t_{i_0,\succ}) - p(t_{i_0,\sim})}{p(t_{i_0,\succ})} u_{\succ}(k', i_0, b). \end{aligned} \quad (11)$$

$$\begin{aligned} \sum_{\substack{W \in pwd(E^p) \\ t_{i_0,\succ} \notin W, t_{i_0,\sim} \notin W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b}} Pr(W) &= \frac{1 - p(t_{i_0,\succ}) - p(t_{i_0,\sim})}{p(t_{i_0,\sim})} \sum_{\substack{W \in pwd(E^p), t_{i_0,\sim} \in W \\ |W \cap E_{\succ,[i_0]}| = k'-1-b \\ |W \cap E_{\sim,[i_0]}| = b+1}} Pr(W) \\ &= \frac{1 - p(t_{i_0,\succ}) - p(t_{i_0,\sim})}{p(t_{i_0,\sim})} u_{\sim}(k', i_0, b+1). \end{aligned} \quad (12)$$

A subtlety is that when $p(t_{i_0}, \succ) = 0$ and $b = b_{\max}$, simply no possible world satisfies the condition in Equation 10, and Equation 10 equals 0.
Altogether, we show that this case can be correctly computed by Equation 7

(2) $u_{\sim}(k', i_0 + 1, b)$ is the probability sum of all possible world W such that $t_{i_0+1,\sim} \in W$, $|W \cap E_{\succ,[i_0+1]}| = k' - b$ and $|W \cap E_{\sim,[i_0+1]}| = b$. Use a similar argument as above, it can be shown that this case is correctly computed by Equation 8 as well.

References

1. Zhang, X., Chomicki, J.: On the semantics and evaluation of top-k queries in probabilistic databases. In: ICDE Workshops. (2008) 556–563
2. Imielinski, T., Lipski, W.: Incomplete information in relational databases. J. ACM **31**(4) (1984) 761–791
3. Cavallo, R., Pittarelli, M.: The theory of probabilistic databases. In: VLDB. (1987)

4. Halpern, J.Y.: An analysis of first-order logics of probability. Artif. Intell. **46**(3) (1990) 311–350
5. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases : The Logical Level. Addison Wesley (1994)
6. Fuhr, N., Rölleke, T.: A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Trans. Inf. Syst. **15**(1) (1997) 32–66
7. Zimányi, E.: Query evaluation in probabilistic relational databases. Theor. Comput. Sci. **171**(1-2) (1997) 179–219
8. Lakshmanan, L.V.S., Leone, N., Ross, R.B., Subrahmanian, V.S.: Probview: A flexible probabilistic database system. ACM Trans. Database Syst. **22**(3) (1997) 419–469
9. Dalvi, N.N., Suciu, D.: Efficient query evaluation on probabilistic databases. VLDB J. **16**(4) (2007) 523–544
10. Benjelloun, O., Sarma, A.D., Halevy, A.Y., Widom, J.: Uldbs: Databases with uncertainty and lineage. In: VLDB. (2006)
11. Widom, J.: Trio: A system for integrated management of data, accuracy, and lineage. In: CIDR. (2005)
12. http://www.infosys.uni-sb.de/projects/maybms/
13. Olteanu, D., Koch, C., Antova, L.: World-set decompositions: Expressiveness and efficient algorithms. Theor. Comput. Sci. **403**(2-3) (2008) 265–284
14. Fagin, R.: Combining fuzzy information from multiple systems. J. Comput. Syst. Sci. **58**(1) (1999) 83–99
15. Fagin, R., Lotem, A., Naor, M.: Optimal aggregation algorithms for middleware. In: PODS. (2001)
16. Natsev, A., Chang, Y.C., Smith, J.R., Li, C.S., Vitter, J.S.: Supporting incremental join queries on ranked inputs. In: VLDB. (2001)
17. Marian, A., Bruno, N., Gravano, L.: Evaluating top- queries over web-accessible databases. ACM Trans. Database Syst. **29**(2) (2004) 319–362
18. Guha, S., Koudas, N., Marathe, A., Srivastava, D.: Merging the results of approximate match operations. In: VLDB. (2004) 636–647
19. Ilyas, I.F., Aref, W.G., Elmagarmid, A.K.: Joining ranked inputs in practice. In: VLDB. (2002)
20. Ilyas, I.F., Aref, W.G., Elmagarmid, A.K.: Supporting top-k join queries in relational databases. In: VLDB. (2003)
21. Soliman, M.A., Ilyas, I.F., Chang, K.C.C.: Probabilistic top- and ranking-aggregate queries. ACM Trans. Database Syst. **33**(3) (2008)
22. Ré, C., Dalvi, N.N., Suciu, D.: Efficient top-k query evaluation on probabilistic data. In: ICDE. (2007)
23. Hua, M., Pei, J., Zhang, W., Lin, X.: Ranking queries on uncertain data: a probabilistic threshold approach. In: SIGMOD Conference. (2008) 673–686
24. Bruno, N., Wang, H.: The threshold algorithm: From middleware systems to the relational engine. IEEE Trans. Knowl. Data Eng. **19**(4) (2007) 523–537
25. Yi, K., Li, F., Kollios, G., Srivastava, D.: Efficient processing of top-k queries in uncertain databases. In: ICDE. (2008) 1406–1408