Finding Causality and Responsibility for Probabilistic Reverse Skyline Query Non-answers

[Extended Abstract]

Yunjun Gao[#†1], Qing Liu[#2], Gang Chen[#†3], Linlin Zhou[#4], Baihua Zheng[*5]
[#]*College of Computer Science*, *Zhejiang University*, *Hangzhou*, *China*
[†]*The Key Laboratory of Big Data Intelligent Computing of Zhejiang Province*, *Hangzhou*, *China*
[*]*School of Information Systems*, *Singapore Management University*, *Singapore*
{[1]gaoyj, [2]liuq, [3]cg, [4]zlinlin}@zju.edu.cn [5]bhzheng@smu.edu.sg

***Abstract*—This paper explores the *causality and responsibility problem* (CRP) for the non-answers to probabilistic reverse skyline queries (PRSQ). Towards this, we propose an efficient algorithm called CP to compute the causality and responsibility for the non-answers to PRSQ. CP first finds candidate causes, and then, it performs verification to obtain actual causes with their responsibilities, during which several strategies are used to boost efficiency. Extensive experiments using both real and synthetic data sets demonstrate the effectiveness and efficiency of the presented algorithms.**

I. Introduction

Causality and responsibility has become a useful tool in the database community for providing intuitive explanations for answers/non-answers to queries. Specifically, given a query Q over a database P, causality aims to find all the tuples in P that cause the presence of answers or the absence of non-answers to the query Q. Responsibility quantifies the effect that each cause has on the appearance of an answer or the absence of a non-answer, which is defined as a function of the size of the smallest contingency set. In the database literature, the *causality and responsibility problem* (CRP) has been explored in relational databases [4, 5] and probabilistic nearest neighbor search [2]. Nonetheless, CRP is *query-dependent*. None of the existing techniques can efficiently find the causality and responsibility for the answers/non-answers to probabilistic reverse skyline queries (PRSQ), which has a wide range of applications. To be more specific, given a D-dimensional uncertain dataset $\mathcal{P}$, a query object q, and a probability threshold α, a probabilistic reverse skyline query returns the objects in $\mathcal{P}$ whose probabilities to be reverse skyline objects are no less than α. The probabilistic reverse skyline query is a useful tool for multi-criteria decision making [3]. As an example, the coach of a basketball team wants to recruit a new player with some preferred skills, and selects candidates for the position. In this case, the coach can take the new position as a query object, and conduct a probabilistic reverse skyline query on the uncertain dataset formed by all basketball players to find those candidates. In some instances, the returned results may disappoint the users. Continue the aforementioned example. If the basketball player finds out himself absent from the candidate set, he may ask questions such as "*What cause me unqualified for this position? What are the degrees of those causes?*". Intuitively, if the player is not qualified for the position, there must have other more suitable players than him, which constitute the causes for his absence from the query result. Those causes enable the player to understand his competitors better and thus to improve his skills to exceed other players. To this end, we study the problem of finding the causality and responsibility for the non-answers to probabilistic reverse skyline queries, which can enhance the explanation capability for database systems and hence improve the usability of the database.

In order to compute the causality and responsibility for the non-probabilistic-reverse skyline object, we propose an efficient algorithm, termed as CP, which follows a filter-and-refinement framework. Specifically, CP finds the candidate causes for a given non-probabilistic-reverse skyline object, and then, it gets the actual causes and their responsibilities by using several pruning strategies to boost the performance of CP algorithm.

II. Problem Formulation

Let P be a D-dimensional dataset, and Q be a query. The fact that an object $a \in P$ is (is not) an answer to Q over P is denoted as $P \vDash Q(a)$ ($P \nvDash Q(a)$). Next, we formally define the causality and responsibility for non-answers to queries.

Definition 1 (*Causality and Responsibility*). Given a D-dimensional dataset P, a non-answer a_n to a query Q, i.e., $P \nvDash Q(a_n)$, and an object p ($\neq a_n$) $\in P$. Then, for the query Q over P: (i) if $(P - \{p\}) \vDash Q(a_n)$, p is a *counterfactual cause* for a_n; and (ii) if there exists *a contingency set* $\Gamma \subseteq P$ for p such that $(P - \Gamma) \nvDash Q(a_n)$ but $(P - \Gamma - \{p\}) \vDash Q(a_n)$, p is *an actual cause* for a_n. Both counterfactual cause and actual cause constitute the *Causality*. Then, the *Responsibility* of p for a_n, denoted as $\rho(p, a_n)$, is defined as:

$$\rho(p, a_n) = \frac{1}{1 + min_{\Gamma}|\Gamma|} \quad (1)$$

Given three objects p_1, p_2, and p_3 in a D-dimensional dataset P, if p_1 *dominates* p_2 w.r.t. p_3, denoted as $p_1 \prec_{p_3} p_2$, it must hold that (i) $\forall\ i \in [1, D]$, $|p_1[i] - p_3[i]| \leq |p_2[i] - p_3[i]|$, and (ii) $\exists\ j \in [1, D]$, $|p_1[j] - p_3[j]| < |p_2[j] - p_3[j]|$.

Definition 2 (*Probabilistic Reverse Skyline Query*). Given a D-dimensional uncertain dataset $\mathcal{P}$, a query object q, and a

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.33

probability threshold $\alpha \in (0, 1]$, a *probabilistic reverse skyline query* (PRSQ) retrieves those objects $u \in \mathcal{P}$ such that the probability of u being a reverse skyline object, denoted as $Pr(u)$, is no smaller than α, i.e.,

$$Pr(u) = \sum_{i=1}^{l_u} u_i.p \cdot \left(\prod_{\forall u' \in \mathcal{P} - \{u\}} \left(1 - Pr\{u' \prec_{u_i} q\}\right) \right) \geq \alpha \quad (2)$$

where $Pr\{u' \prec_{u_i} q\}$ is the probability of q being dynamically dominated by u' w.r.t. u_i, and

$$Pr\{u' \prec_{u_i} q\} = \sum_{j=1 \wedge u'_j \prec_{u_i} q}^{l_{u'}} u'_j.p \quad (3)$$

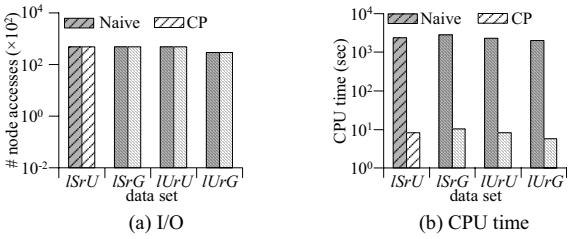
Definition 3 (*Causality and Responsibility Problem on Probabilistic Reverse Skyline Query*). Given an uncertain dataset $\mathcal{P}$, a query object q, a probability threshold $\alpha \in (0, 1]$, and a non-probabilistic-reverse skyline object a_n, the *Causality and Responsibility Problem on Probabilistic Reverse Skyline Query* (CR2PRSQ) needs to (i) find a set $C \subseteq \mathcal{P}$ such that (a) for $\forall\ c \in C$, c is a causality for a_n, and (b) for $\forall\ c' \in (\mathcal{P} - C)$, c' is not a causality for a_n; and (ii) for $\forall\ c \in C$, compute its degree of responsibility $\rho(c, a_n)$ based on Equation (1).

III. CR2PRSQ Algorithm

In this section, we present the algorithm for computing the causality and responsibility for a specified non-probabilistic-reverse skyline object. CR2PRSQ computation involves two aspects, i.e., the computation of the causality and its corresponding responsibility. Towards this, we propose an efficient algorithm, called CP, in order to find the causality and responsibility for the non-answers to probabilistic reverse skyline queries.

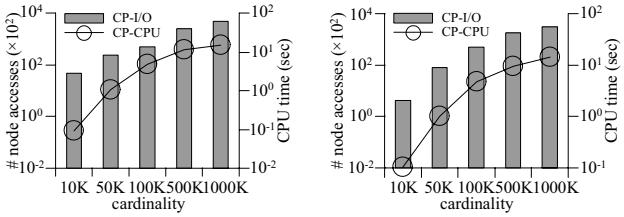
Finding the causality and responsibility for the non-answers to probabilistic reverse skyline queries poses two major challenges. The first one is how to efficiently find the causes for the non-answers. To this end, CP algorithm employs a filter-and-refinement framework to identify the causes for the non-answers to probabilistic reverse skyline queries. Specifically, CP finds the candidate cause set and then refines it to get the actual causes. The second challenge is how to efficiently find the minimum contingency set for every cause, since we define the responsibility of a cause as a function of the size of its smallest contingency set. In view of this, for each cause of a non-answer to the probabilistic reverse skyline query, the minimal contingency set is found by examining the candidate contingency set. In order to reduce the examination cost, CP algorithm utilizes several lemmas to identify the true objects (false objects) that must be present in (absent from) the minimum contingency set.

It is worth mentioning that CP algorithm assumes that the probabilistic dataset $\mathcal{P}$ follows the discrete sample model. In addition, we extend CP to the *continuous pdf* probabilistic dataset model. Moreover, we also extent the CP algorithm to compute the causality and responsibility over reverse skyline queries, which further shows the flexibility of CP algorithm. Interested readers can refer to [1] for the detailed description of the proposed algorithms.



(a) I/O (b) CPU time

Fig. 1. CP cost vs. Naive cost



(a) *lSrU* (b) *lSrG*

Fig. 2. CP cost vs. *cardinality*

IV. Experimental Evaluation

We experimentally evaluate the efficiency of the presented algorithms using both real and synthetic data sets.

CP vs. Naive. First, we compare the performance of CP algorithm with Naive algorithm. The Naive algorithm first finds the candidate causes like CP, and then, it refines them by examining all candidate contingency sets without using any pruning heuristic. Figure 1 depicts the performance of the two algorithms. It is observed that the I/O cost of CP and Naive is the same, while the CPU time of CP algorithm outperforms that of Naive algorithm. This is because the I/O cost mainly comes from the first step of the algorithms, i.e., finding the candidate causes, which is the same for both CP and Naive. Consequently, the I/O cost of them is identical. At the refinement step, CP algorithm utilizes a series of strategies to boost efficiency. Therefore, the CPU time of CP algorithm is smaller than that of Naive algorithm.

Effect of cardinality. Next, we use synthetic datasets to investigate the scalability of our algorithms. Figure 2 plots the corresponding results. As expected, the I/O cost and CPU time of CP ascend as $\mathcal{P}$ grows. The reason is that, the larger the dataset cardinality is, the more intensive the data is. Thus, there are more candidate causes for the non-probabilistic-reverse skyline object, incurring longer processing time.

Acknowledgements. This work was supported in part by the 973 Program of China Grant No. 2015CB352502, the NSFC Grant No. 61522208, 61379033, 61472348, and U1609217. Yunjun Gao is a corresponding author of this work.

References

[1] Y. Gao, Q. Liu, G. Chen, L. Zhou, and B. Zheng, "Finding causality and responsibility for probabilistic reverse skyline query non-answers," *IEEE Trans. Knowl. Data Eng.*, 28(11): 2974–2987, 2016.
[2] X. Lian and L. Chen, "Causality and responsibility: Probabilistic queries revisited in uncertain databases," in *CIKM*, pp. 349-358, 2013.
[3] X. Lian and L. Chen, "Reverse skyline search in uncertain databases," *ACM Trans. Database Syst.*, 35(1), article 3, 2010.
[4] A. Meliou, W. Gatterbauer, K. F. Moore, and D. Suciu, "The complexity of causality and responsibility for query answers and non-answers," in *VLDB*, pp. 34–45, 2011.
[5] B. Qin, S. Wang, X. Zhou, and X. Du, "Responsibility analysis for lineages of conjunctive queries with inequalities," *IEEE Trans. Knowl. Data Eng.*, 26(6): 1532–1543, 2014.