Evaluating Top-k Queries with Inconsistency Degrees
Ousmane Issa, Angela Bonifati, Farouk Toumani

To cite this version:
Ousmane Issa, Angela Bonifati, Farouk Toumani. Evaluating Top-k Queries with Inconsistency Degrees. Proceedings of the VLDB Endowment (PVLDB), VLDB Endowment, 2020, 13 (12), pp.2146-2158. 10.14778/3407790.3407815. hal-02898931

HAL Id: hal-02898931
https://hal.archives-ouvertes.fr/hal-02898931
Submitted on 14 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Evaluating Top-k Queries with Inconsistency Degrees

Ousmane Issa
University Clermont Auvergne
ousmane.issa@uca.fr

Angela Bonifati
Lyon 1 University
angela.bonifati@univ-lyon1.fr

Farouk Toumani
University Clermont Auvergne
farouk.toumani@uca.fr

ABSTRACT
We study the problem of augmenting relational tuples with inconsistency awareness and tackling top-k queries under a set of denial constraints (DCs). We define a notion of inconsistent tuples with respect to a set of DCs and define two measures of inconsistency degrees, which consider single and multiple violations of constraints. In order to compute these measures, we leverage two models of provenance, namely why-provenance and provenance polynomials. We investigate top-k queries that allow to rank the answer tuples by their inconsistency degrees. Since one of our measure is monotonic and the other non-monotonic, we design an integrated top-k algorithm to compute the top-k results of a query w.r.t. both inconsistency measures. By means of an extensive experimental study, we gauge the effectiveness of inconsistency-aware query answering and the efficiency of our algorithm with respect to a baseline, where query results are fully computed and ranked afterwards.

1. INTRODUCTION
Assessing the quality of raw data is crucial in numerous data management tasks, spanning from traditional querying and data integration to cutting-edge in-database analytical and inference/ML processes. Throughout these processes, the quality of the output as well as the trustworthiness of the results might be tremendously affected by the quality of the input tuples [29].

Past work on data curation for relational tuples has mainly addressed the problem of detecting and repairing the violations with respect to a set of constraints [39]. Consistent query answering has also been considered as a mean to reconcile several possible repairs of the original data [11]. However, little or no attention has been paid to leave the database instances intact and quantifying their degrees of inconsistency at different levels of granularity (tuple, sets of tuples, entire relations) as we do in this paper for the first time. Such a characterization enables the users of a DBMS to quantify the level of trust that they shall expect from the data that they query and manipulate.

In our work, we are interested in augmenting relational instances with novel inconsistency measures that can also be propagated to query results. To this end, we focus on top-k rank join queries over an inconsistent relational database in the presence of a set of Denial Constraints (DCs) [17]. The inconsistency degree of an answer tuple of a given query is determined by relying on provenance-based information of the input tuples. We first leverage why-provenance in order to identify the inconsistent base tuples of a relational instance with respect to a set of DCs. Then, we rely on provenance polynomials [26] in order to propagate the annotations of inconsistencies from the base tuples to the answer tuples of Conjunctive Queries (CQs). Building upon the computed annotations, we define two measures of inconsistency degrees, which consider single and multiple violations of constraints. Since one of our measures is a monotone function and the other a non-monotone function, we design an integrated top-k algorithm to rank the top-k results of a query w.r.t. the above inconsistency measures.

We envision several applications of our framework: Inconsistency-conscious queries for analytical tasks, as we expect that our framework enables inconsistency quantification in querying and analytical tasks within data science pipelines. Our annotations are not merely numbers as they also convey provenance-based information about the violated constraints, the latter being ready for user consumption in data science tasks.

External annotations for data cleaning pipelines, as our approach can also ease the tasks of data cleaning tools such as OpenRefine, Wrangler and Tableau by injecting into them the external information of inconsistency indicators and putting upfront the resulting ranking prior to cleaning and curation.

Approximation schemes for integrity constraints that have been employed recently in order to guarantee a polynomial number of samples required by recent probabilistic inference approaches for data cleaning [38]. We believe that an alternative to constraint approximation would be to build samples based on the top-k number of constraints leading to the most consistent (the least consistent, respectively) tuples.

Combined ranking since our quality-informed ranking and query evaluation can be combined with other ranking criteria, such as for instance user preferences in recommender systems and novel unfairness and discrimination detection.
is violated both by the pairs of tuples \((t_2, t_3)\) and \((t_2, t_4)\) while the constraint \(C_2\) is violated by the tuples \((t_2, t_7)\) and the constraint \(C_3\) is violated by the tuples \((t_3, t_7)\). Hence, the inconsistent base tuples of \(I\) w.r.t. the set of denial constraint \(\{C_1, C_2, C_3\}\) are: \(t_2, t_3, t_4, t_7\).

Quantifying the inconsistency degrees of query answers. Evaluating the query \(Q_{as}\) over the inconsistent database \(I\), under the bag answer semantics, returns the answers reported in the first column of Table 1. The second column of Table 1 shows the base tuples of \(I\) that contribute to each answer. One can observe that the tuple \((d_4, d_4)\) is a consistent answer since it is computed using consistent base tuples. This is, however, not the case for \((d_2, d_2)\), which can be derived in three possible ways, either using the base tuples \(\{t_2, t_3, t_7\}\) or \(\{t_2, t_4, t_7\}\) or \(\{t_2, t_6, t_7\}\). Notice that all the above derivations use inconsistent base tuples.

In this paper, we consider two alternatives for quantifying the inconsistency degree of query answers. The first approach quantifies the inconsistency degree of an answer \(t\) by counting the number of constraints violated by the base tuples that contribute to \(t\). The third column of Table 1 shows the constraints violated by the base tuples that contribute to each answer. For example, in line 1 of Table 1 the computation of the answer \((d_2, d_2)\) from the base tuples \(\{t_2, t_3, t_7\}\) leads to the violation of the constraints \(C_1\) (violated by \(t_2\) and \(t_3\)), \(C_2\) (violated by \(t_2\) and \(t_7\)) and \(C_3\) (violated by \(t_7\)). As a consequence, the inconsistency degree of the answer \((d_2, d_2)\) when it is computed from the base tuples \(\{t_2, t_3, t_7\}\) is equal to 3 (i.e., 3 constraints are violated by the base tuples that contribute to this answer).

A second approach to quantify the inconsistency degree of an answer amounts to counting the number of violations of the constraints per each base tuple. The fourth column of Table 1 shows the constraints violated by the base tuples while reporting how many times the same violation occurred (indicated by the exponent of the constraints). For instance, in the first row of Table 1 the computation of the answer \((d_2, d_2)\) from the base tuples \(\{t_2, t_3, t_7\}\) leads to the violation of the constraint \(C_1\) twice (violated by \(t_2\) and \(t_3\)), the violation of \(C_2\) two times (violated by \(t_2\) and \(t_7\)) and the violation of \(C_3\) one time (violated by \(t_7\)). Hence, the corresponding inconsistency degree of \((d_2, d_2)\) is equal to 5 (i.e., 5 violations of the constraints by base tuples).

In this paper, we focus on a constraint-based approach to quantify inconsistency degrees of query answers under bag semantics. As Table 1 (fourth column) shows, the annotations of the base tuples are the richest and the most informative ones since they consider the violations in terms of the constraints and their occurrences. In particular, we consider two alternative measures: (i) counting the number of constraints violated by the base tuples contributing to a given answer (called in the sequel, single occurrence quantification), and (ii) summing up the exponents of the constraints violated by base tuples contributing to a an answer (called in the sequel, multiple occurrence quantification).

Inconsistency-aware query processing. With the inconsistency measures at hand, we are interested in top-k query processing while taking into account the inconsistency degrees of query answers. This problem can be framed in the general context of top-k rank queries where the inconsistency measures are used as scoring functions that can be exploited to rank the answers of queries. In particular, given a query \(Q\), top-k ranking aims at computing the subset of \(Q\)’s
answers with the top $k$ scores, where the scores correspond to the inconsistency degrees of the tuples. Continuing with our example, the top-2 answers of the query $Q_{ex}$ over the database $I$ are: $(d_2, d_4)$, $(d_2, d_2)$, with respective inconsistency degrees 0 and 3 with single occurrence quantification), or, respectively, inconsistency degrees 0 and 5 with multi occurrence quantification, respectively. In the remainder, we will see that both augmenting the instances with inconsistency measures and enabling top-k queries on them are far from being trivial and need special treatment.

### 3. RELATED WORK

**Consistent Query Answering and Repairs.** Central to data quality is data consistency that has been extensively investigated in the past. Starting from the pioneering work of [5], there has been a wealth of research on the problem of **consistent query answering in inconsistent databases** [7, 30, 36, 29, 8]. Most of the existing works make a distinction between consistent and inconsistent answers based on the notion of a database repair (see [10] for a survey). A repair is a database instance obtained by applying some minimal changes on the initial database instance in order to make it compliant with a given set of integrity constraints. Usually, there exist several possible repairs of a given database and for such a reason, following the **certain answers** semantics, consistent answers to a given query are defined as being the intersection of the query answers on all possible repairs of the initial database instance. The problem of **consistent query answering** has been investigated in several settings depending on: (i) the definition of repairs, where a repair is defined in terms of the sets of inserted and/or deleted tuples [11, 10, 54, 13], or updates [39, 32, 6] and (ii) the considered integrity constraints and the query language. The latter has been studied for a variety of queries and constraints such as first-order queries and binary universal constraints or single functional dependencies [5], union of conjunctive queries and key and inclusion dependencies [13], first-order queries and denial constraints [16], to mention a few.

Recent approaches study a counting variant of consistent query answering [36], i.e., counting how many repairs satisfy a given query $q$. The work in [36] establishes a dichotomy for the $\#\text{CERTAINTY}(q)$ problem, where $q$ is a boolean conjunctive query possibly with self-joins and with single-attribute primary-key (PK) constraints. Our work is agnostic to the repairing semantics and computation of repairs and only leverages constraint violations.

**Quantifying Inconsistency in KBs.** Reasoning in Knowledge Bases (KBs) in the presence of inconsistency is also widely studied in the AI community, where it has been shown that an arbitrary consequent can be entailed from an inconsistent set of premises (see the survey by Lang et al. [35]). Similarly, the role of inconsistency tolerance in KBs has been introduced in [9]. They are more concerned about the co-existence of consistency and inconsistency as separate axioms when doing reasoning in KBs rather than using inconsistency measures to yield inconsistency-aware query answering, as we do in our work.

Another line of research deals with the definition of inconsistency measures in Knowledge Bases (KBs) [29, 24]. Numerous measures of inconsistency in KBs have been proposed based on various paradigms such as information theory [35], possibility theory [19] or quasi-classical logic [24], to mention a few. The ultimate goal is, however, to provide measures to quantify the inconsistency of a KBs in order, for example, to allow comparison of full KBs based on their inconsistency degrees. Extensions of such an approach to the database field has been recently investigated in [10] where inconsistency measures based on various classes of repair semantics are studied. However, their notion of inconsistency measures (despite the name) relies on restoring the consistency of a database, thus on the class of repairs admitted by the database. As such, it is quite different from our notion of inconsistency.

**Top-k Query Processing.** Top-k query processing algorithms address the efficient computation of the first top-k answers of selection [22], join [27, 30, 31] and aggregate queries [15]. There exists a vast literature on top-k join query processing, see [28] for a survey. However, a basic assumption of several previous papers on this topic is that the scoring function $f$ used to aggregate the scores of the underlying tuples is a monotonic function [15, 28, 30, 31]. Non-monotonic scoring functions have been considered in [40] for top-k query processing, while assuming that attribute values are indexed by tree-structured indexes (e.g., B-tree, R-tree). However, none of the existing top-k ranking algorithms would be optimal to handle both monotonic and non-monotonic functions. Additionally, our novel top-k algorithm is tailored for inconsistency measures and leverages a suitable cost model for them. Our algorithm is thus generic and designed in a way that it incorporates both monotonic and non-monotonic inconsistency measures.

Analytical studies of the performance of existing top-k join algorithms have been conducted mostly for the class of algorithms, introduced initially in [22], that use a sorted data access strategy, i.e., the input relations are scanned sequentially in base score order [30, 31]. Such a framework, which imposes a sequential data access strategy, is not suitable for non-monotonic functions. This motivated our work on a new notion of optimality based on: (i) a more general class of algorithms, called Semi-Blind Algorithms (SBA),
that alleviate the assumptions on the permitted data access strategies, and (ii) a new cost model that enable to deal with nondeterministic choices due to the assumption that SBA algorithms do not exploit any specific knowledge on join attributes. Using this framework, we show the optimality of our algorithm w.r.t. the SBA class.

4. PRELIMINARIES

We introduce some basic notions used throughout the paper. Table 2 summarizes the notation. We consider $S = \{R_1, \ldots, R_n\}$ as a database schema with $R_i (i \in [1, n])$ a predicate and with arity$(R_i)$ denoting the arity of $R_i$. We consider $D$, $I$ as the domain and a database instance over $S$ and $D$, respectively. Let $\Gamma$ be an infinite set of identifiers distinct from the domain $D$. We denote by $id$ a function that associates to each tuple $t \in I(R_i)$ an unique identifier $id(t)$ from the set $\Gamma$.

**Conjunctive Queries (CQ).** A CQ is in form $Q(u) \leftarrow R_1(u_1), \ldots, R_n(u_n)$, where each $R_i$ is a relation symbol in $S$ and $Q$ is a relation symbol in the output schema, $u$ is a tuple of either distinguished variables or constants and each $u_i$ is a tuple of either variables or constants having the same arity as $R_i$. The set of variable in a query $Q$ is denoted $Vars(Q)$.

**Denial Constraints.** A denial constraint is of the form:

$$\forall \alpha \in \{0, 1\} \exists a \in K : a \odot \alpha = 0$$

and $\odot$ (product) such that $(K, \odot, 0)$ and $(K, \odot, 1)$ are commutative monoids with identities 0 and 1 respectively, $\odot$ is distributive over $+$ and $0 \odot a = a \odot 0 = 0$ holds $\forall a \in K$.

An $n$-ary $K$-relation is a function $R : D^n \rightarrow K$ such that its support, defined by $supp(R) \triangleq \{ t \in D^n, R(t) \neq 0 \}$, is finite. Let $R$ be an $n$-ary $K$-relation and let $t \in D^n$, the value $R(t) \in K$ assigned to the tuple $t$ by the $K$-relation $R$ is called the annotation of $t$ in $R$. Note that $R(t) = 0$ means that $t$ is "out of" $R$ [23]. A $K$-instance is a mapping from relations symbols in a database schema $S$ to $K$-relations (i.e., a finite set of $K$-relations over $S$). If $J$ is a $K$-instance over a database schema $S$ and $R_i \in S$ is a relation symbol in $S$, we denote by $\mathcal{J}(R_i)$ the $K$-relation corresponding to the value of $R_i$ in $\mathcal{J}$.

**Conjunctive queries on $K$-instances.** Let $Q(u) \leftarrow R_1(u_1), \ldots, R_n(u_n)$, be a conjunctive query and let $\mathcal{J}$ be a $K$-instance over the same schema than $Q$, with $(K, \odot, \oplus, 0, 1)$ a semiring.

A valuation of $Q$ over a domain $D$ is a function $v : Vars(Q) \rightarrow D$, extended to be the identity on constants. The result of executing a query $Q$ over a $K$-instance $\mathcal{J}$, using the semiring $(K, \odot, \oplus, 0, 1)$, is the $K$-relation $Q(\mathcal{J})$ defined as follows: $Q(\mathcal{J}) \triangleq \{ (v(u), \Pi^n_{i=1} R_i(v(u_i))) | v \text{ is a valuation over } Vars(Q) \}$.

The $K$-relation $Q(\mathcal{J})$ associates to each tuple $t = v(u)$, which is in the answer of the query $Q$ over the $K$-instance $\mathcal{J}$, an annotation $\Pi^n_{i=1} R_i(v(u_i))$ obtained from the product, using the $\odot$ operator, of the annotations $R_i(v(u_i))$ of the base tuples contributing to $t$. Since there could exist different ways to compute the same answer $t$, the complete annotation of $t$ is obtained by summing the alternative ways (i.e., the various valuations) to derive a tuple $t$ using the $\oplus$ operator. Consequently, the provenance of an answer $t$ of $Q$ over a $K$-instance $\mathcal{J}$ is given by: $Q(\mathcal{J})(t) = \sum_{v \in t} \Pi^n_{i=1} R_i(v(u_i))$.

**Example 1.** Consider the $K$-instance $hdb^5$ in Figure 3 with $K = \mathbb{N}[\mathbb{Y}]$, the set of polynomials with variables from $\mathbb{Y}$ and coefficients in $\mathbb{N}$. The annotation of the tuple $(01, d2, 1)$ of relation $S$ is given by: $hdb^5(S)((01, d2, 1)) = C_1$. Evaluation of $Q_x$ over $hdb^5$ uses $(\mathbb{N}[\mathbb{Y}], +, \times, 0, 1)$ semiring and, for example, leads to an answer $(d2, d2)$ annotated as follows: $Q_x(hdb^5)((d2, d2)) = C_2^2 C_2^2 C_3 + C_2^2 C_2^2 C_2^2 + C_2^2 C_2^2 C_3$.

5. QUANTIFYING INCONSISTENCY DEGREES OF QUERY ANSWERS

In this section, we focus on the problem of quantifying the inconsistency degree of query answers. Let $I$ be an instance over a database schema $S$, let $DC$ be a set of denial constraints over $S$ and let $\mathcal{Y}$ be the set of identifiers of the constraints in $DC$. We proceed in three steps:

(i) **Identifying inconsistent base tuples.** We first start by identifying the inconsistent tuples of an instance $I$ over $S$ w.r.t a set of denial constraints $DC$. To achieve this task, we turn each constraint $C \in DC$ into a conjunctive query $Q^{C, id}$ and, by exploiting why-provenance, we consider the lineage of $Q^{C, id}$ as the set of tuples of $I$ that violates the constraint $C$. As a consequence, we are able to define a function $VC(I, DC, t)$ that associates to each tuple $t \in I$, the set of constraints violated by $t$.

$$\mathcal{N}(P) \quad \text{Set of monomials of a polynomial } P$$
$$\text{Vars}(M) \quad \text{Set of variables of a monomial } M$$
$$W(M) \quad \text{weight of a monomial } M$$
$$DC \quad \text{Set of denial constraints}$$
$$\mathcal{T} \quad \text{Set of constraints identifiers}$$
$$\text{Vars}(M) \quad \text{Set of tuple identifiers}$$
$$\text{supp} \quad \text{Set of constraints violated by a tuple } t$$
$$\mathcal{T}^+ \quad \text{A } \mathcal{T}-\text{instance obtained from } \mathcal{T}$$
$$\mathcal{T}^+(R_i) \quad \text{a } \mathcal{T}-\text{relation of } \mathcal{T}^+$$
$$\Pi^\mathcal{T}\mathcal{K} \quad \text{top-k query w.r.t. measure } \alpha$$
$$\text{ind}(R) \quad \text{Index associated with } R$$
$$\text{label}(B) \quad \text{Label of a bucket } B \text{ of an index}$$

Table 2: Summary of the notation used in the paper.
(ii) Annotating the initial database instance. Using the set \( VC(I, DC, t) \), we convert the instance \( I \) into a \( Y \)-instance, denoted \( \mathcal{T} \), obtained by annotating each tuple in \( I \) by a monomial with variables from \( Y \). The variables record the constraints violated by the annotated tuple.

(iii) Defining inconsistency degrees of query answers. Given a query \( Q \) over the instance \( I \), we use the provenance semirings to annotate the query answers. The latter provenance is the most informative provenance of formulation provenance \([25]\) and hence is exploited in our setting in order to define two inconsistency degrees for query answers.

We shall detail in the sequel the proposed approach.

5.1 Identifying inconsistency degrees of base tuples

Let \( I \) and \( DC \) be respectively an instance and a set of DCs over \( S \). We first convert each constraint \( C \) of \( DC \) as follows:

\[ Q^{C_{id}}(I) \leftarrow R_1(u_1) \wedge \ldots \wedge R_n(u_n) \wedge C \]

\[ Q^{C_{id}} \] is defined as follows:

\[ Q^{C_{id}}(I) \leftarrow R_1(u_1) \wedge \ldots \wedge R_n(u_n) \wedge C \]

**Example 2.** The set DC of the constraints depicted in Figure 1 are converted into the following boolean queries:

\[ Q^{C_1}(I) \leftarrow D(x, y, z) \wedge S(x, y, u) \wedge z > u \]

\[ Q^{C_2}(I) \leftarrow D(x, y, z) \wedge V(x, y, u) \wedge z > u \]

\[ Q^{C_3}(I) \leftarrow S(x, y, z) \wedge V(x, v, z) \]

It is easy to verify that an instance \( I \) violates the set of denial constraints \( DC \) iff the boolean UCQ query \( Q^{DC}(I) \) evaluates to \( \bot \) over \( I \), i.e., at least one of the conjunctive queries \( Q^{C_{id}} \), for \( C \in DC \), returns false when evaluated over \( I \). The base tuples of \( I \) which are inconsistent w.r.t. \( DC \) are exactly those tuples of \( I \) that contribute to the computation of the empty answer \( <> \) (i.e., true) for query \( Q^{DC} \) when evaluated over \( I \). In particular, a base tuple \( t \) of \( I \) violates a constraint \( C \in DC \) iff \( t \) contributes to the derivation of the answer true when the query \( Q^{C_{id}} \) is evaluated over \( I \).

We shall use below the why-provenance \([12]\) to compute the inconsistent base tuples while keeping track of the constraints violated by each inconsistent tuple. Indeed, the why-provenance of an answer \( t \) in a query output is made of the set of all contributing input tuples.

We first formulate the why-provenance in the provenance semirings framework as proposed in \([25]\). Let \( P(G) \) be the powerset of the set of all identifiers \( G \). Consider the following provenance semiring: \( P(G) \cup \{ \bot \} \), where: \( \forall S, T \in P(G) \cup \{ \bot \} \), we have \( S \cup = S \), \( S \cup \bot = S \), \( \bot \cup = \bot \), and \( S \cup T = S \cup T = S \cup T \) if \( S \neq \bot \) and \( T \neq \bot \). This semiring consists of the powerset of \( G \) augmented with the distinguished element \( \bot \) and equipped with the set union \( (\cup) \) operation which is used both as addition and multiplication. The distinguished element \( \bot \) is the neutral element of the addition and the annihilating element of the multiplication.

In order to compute the why-provenance, we convert the instance \( I \) over the schema \( S \) into a \( K \)-instance, denoted \( I^{LP} \), with \( K = P(\Gamma) \cup \{ \bot \} \). The \( K \)-instance \( I^{LP} \) is defined as follows:

\[ I^{LP}(R_i)(t) = \begin{cases} 0 & \text{if } t \notin I^{LP}(R_i) \\ \prod_{C_{id} \in T} C_{id}^{l} & \text{otherwise} \end{cases} \]

with \( l = 1 \) if \( C_{id} \in VC(I, DC, t) \) or \( l = 0 \) otherwise.

**Example 3.** Figure 2 shows the provenance database \( hdb^{LP} \) obtained from the hospital database \( hdb \) of our motivating example by annotating each tuple \( t \in hdb \) with a singleton set \( \{id(t)\} \) (column \( iprov \)) containing the tuple identifier of \( t \).

Using the provenance semirings \( (P(\Gamma) \cup \{ \bot \}, +, \cdot, \bot) \), we define below the function \( VC \) that associates with each base tuple \( t \) of \( I \) the set of constraints violated by \( t \).

**Definition 2 (Violated constraints).** Given an instance \( I \) and a set of denial constraints \( DC \), the set \( VC(I, DC, t) \) of constraints of \( DC \) violated by a tuple \( t \) in \( I \) is defined as follows:

\[ VC(I, DC, t) = \{ C_{id} \in T \mid t \in I^{LP}(C_{id})(<>)) \} \]

**Example 4.** Consider the boolean conjunctive queries \( Q^{C_1}, Q^{C_2} \) and \( Q^{C_3} \) of Example 1. Hence, the lineage of the output true of each query gives the set of tuples that violate the corresponding constraint:

\[ VC(Q^{C_1}) = \{ t_2, t_3, t_4 \} \]

\[ VC(Q^{C_2}) \]

These lineages enable us to compute the set of constraints violated by each tuple \( t \) given by the function \( VC(I, DC, t) \). For example, we have \( VC(I, DC, t_1) = \emptyset \), \( (i.e., t_1 \) is a consistent tuple) while \( VC(I, DC, t_2) \) and \( VC(I, DC, t_3) \) are \( \{ C_1, C_2 \} \) and \( \{ C_1 \} \) respectively.

Note that, even if why-provenance computes richer annotations \([12]\), we choose to not keep the information about combinations of tuples that cause a violation. Hence, our approach is currently agnostic to the inconsistency reasons.

5.2 Annotating the database instance

In the next section, we will show how to use the provenance polynomials to define the inconsistency degrees of query answers. In order to obtain that, we first need to convert the instance \( I \) into a \( N[Y] \)-instance, denoted \( \mathcal{T} \). As shown in the following definition, an instance \( \mathcal{T} \) is derived from \( I \) by tagging each tuple \( t \) in \( I \) with a monomial with variables in \( Y \).

**Definition 3 (\( \mathcal{T} \)-instance).** Let \( I \) be an instance over a database schema \( S \) and let \( DC \) be a set of denial constraints over \( S \). Let \( K = N[Y] \). The \( K \)-instance \( \mathcal{T} \) is constructed as follows:

\[ \mathcal{T}(R_i)(t) = \begin{cases} 0 & \text{if } t \notin \mathcal{T}(R_i) \\ \prod_{C_{id} \in T} C_{id}^{l} & \text{otherwise} \end{cases} \]

with \( l = 1 \) if \( C_{id} \in VC(I, DC, t) \) or \( l = 0 \) otherwise.

Hence, an annotation \( \mathcal{T}(R_i)(t) \) of a tuple \( t \) is equal to \( 1 \) if the base tuple \( t \) is consistent (i.e., \( VC(I, DC, t) = \emptyset \)), otherwise it is equal to a monomial expression that uses as variables the identifiers of the constraints violated by \( t \) (i.e., the elements of \( VC(I, DC, t) \)).
The annotations of the tuples $t_1$ (a consistent tuple) and $t_2$ (an inconsistent tuple). From the previous example, we have $\text{VC}(I, DC, t_1) = \emptyset$ and hence the annotation of $t_1$ is computed as follows: $\text{hdb}^T(R_i)(t_1) = \prod_{C \in \text{ann}(t_1)} C_i^0 = 1$. For the tuple $t_2$, we have $\text{VC}(I, DC, t_2) = \{C_1, C_2\}$ and hence: $\text{hdb}^T(R_i)(t_2) = C_1^2 \times C_2^2 \times C_3^2 = C_1 C_2$.

In the sequel, we assume that the relations of an annotated instance $I$ are augmented with an attribute prov that stores the annotations of the base tuples. As an example, Figure 2 shows the annotated relations of the instance $\text{hdb}^T$ together with their respective prov columns.

**5.3 Computing inconsistency degrees of query answers**

Given a query $Q$, we evaluate $Q$ over the $\mathbb{N}[\mathcal{I}]+$-instance $I^\mathcal{T}$ and use the provenance polynomials summing in order to annotate each answer $t$ of $Q$. The computed annotations, expressed as polynomials with variables from the set $\mathcal{T}$ of constraint, are fairly informative as they allow to fully record how the constraints are violated by base tuples that contribute to each answer. Such annotations are hence exploited to compute the various inconsistency measures needed for query answers.

**Example 6.** Continuing with the example, evaluating the query $Q_{ex}$ over $\text{hdb}^T$ and computing its polynomial provenance leads to the following annotated answers:

- $Q_{ex}(\text{hdb}^T)(\langle d2, d2 \rangle) = C_1^2 C_2^2 C_3 + C_1^2 C_2^2 C_3 + C_1^2 C_2^2 C_3$
- $Q_{ex}(\text{hdb}^T)(\langle d4, d4 \rangle) = 1$.

The monomial $C_1^2 C_2^2 C_3$ that appears in the annotation of the answer $\langle d2, d2 \rangle$ of $Q_{ex}$ encodes the fact that this answer can be computed from inconsistent base tuples that lead to the violation of the constraints $C_1$ and $C_2$ twice and to the violation of the constraint $C_3$ once.

Hence, the polynomial expression $Q(I^\mathcal{T})(t)$ fully records the inconsistency of an output $t$ in terms of violations of constraints and therefore can be used to quantify the inconsistency degrees of a query outputs. Consider a polynomial $P = Q(I^\mathcal{T})(t)$ of an output $t$ of a given query $Q$. Each monomial $M$ from $P$ gives an alternative way to derive the output $t$. In the sequel, we consider bag semantics of query answers. Although our approach is extensible to set semantics of query answers, we do not discuss this further in the paper.

Under bag semantics, each query answer will be annotated with a monomial corresponding to the unique derivation of the considered answer.

Two different measures can be defined in order to quantify the inconsistency degree of an answer $t$ depending on how one deals with multiple occurrences of the same variable in monomials. This situation occurs when a constraint is violated by more than one base tuple that contribute to an answer $t$. We define two possible quantifications to deal with this issue: single occurrence quantification, in which a variable that appears in a monomial is counted exactly once, and multi-occurrence quantification, where the exact number of occurrences of a variable in a monomial is taken into account when quantifying the inconsistency degree of a given answer. As a consequence, we obtain two different inconsistency measures: the CBM (Constraint-based, Bag semantics, Multiple occurrence) measure and the CBS (Constraint-based, Bag semantics, Single occurrence) measure.

**Definition 4.** *(Inconsistency measures).* Let $\mathcal{I}$, $Q$ and $DC$ be defined as previously. Let $M = Q(I^\mathcal{T})(t)$ be the monomial annotating an output $t$ of $Q$. We define inconsistency measures of $t$ as: $\text{CBM}(t, Q, I, DC) \stackrel{\text{def}}{=} \text{weight}(M)$ and $\text{CBS}(t, Q, I, DC) \stackrel{\text{def}}{=} |\text{Var}(M)|$.

It follows that a single occurrence quantification of a monomial $M$ amounts to counting the number of distinct variables that occur in $M$ while a multi occurrence semantics sums the total number of occurrence of each variable in $M$ (given by the weight $\text{weight}(M)$ of $M$).

**Example 7.** Consider the annotation $Q_{ex}(\text{hdb}^T)(\langle d2, d2 \rangle) = C_1^2 C_2^2 C_3 + C_1^2 C_2^2 C_3 + C_1^2 C_2^2 C_3$. The monomial $C_1^2 C_2^2 C_3$ that appears in the annotation of the answer $\langle d2, d2 \rangle$ of $Q_{ex}$ encodes that this answer can be computed from inconsistent base tuples that lead to the violation of the constraints $C_1$ and $C_2$ twice and to the violation of the constraint $C_3$ once.

Hence, the polynomial expression $Q(I^\mathcal{T})(t)$ fully records the inconsistency of an output $t$ in terms of violations of constraints and therefore can be used to quantify the inconsistency degrees of a query outputs. Consider a polynomial $P = Q(I^\mathcal{T})(t)$ of an output $t$ of a given query $Q$. Each monomial $M$ from $P$ gives an alternative way to derive the output $t$. In the sequel, we consider bag semantics of query answers. Although our approach is extensible to set semantics of query answers, we do not discuss this further in the paper.

Under bag semantics, each query answer will be annotated with a monomial corresponding to the unique derivation of the considered answer.

Two different measures can be defined in order to quantify the inconsistency degree of an answer $t$ depending on how one deals with multiple occurrences of the same variable in monomials. This situation occurs when a constraint is violated by more than one base tuple that contribute to an answer $t$. We define two possible quantifications to deal with this issue: single occurrence quantification, in which a variable that appears in a monomial is counted exactly once, and multi-occurrence quantification, where the exact number of occurrences of a variable in a monomial is taken into account when quantifying the inconsistency degree of a given answer. As a consequence, we obtain two different inconsistency measures: the CBM (Constraint-based, Bag semantics, Multiple occurrence) measure and the CBS (Constraint-based, Bag semantics, Single occurrence) measure.

**Lemma 1.** Let $\mathcal{I}$, $Q$, $DC$ be defined as previously. $\forall t \in Q(I)$ we have:

- $\text{CBM}(t, Q, I, DC) = \text{CBS}(t, Q, I, DC) \Rightarrow t$ is a CQA.

**Lemma 6** is trivial to prove as any answer that has inconsistency degree equal to 0 is computed from tuples that do not violate any constraint. As these tuples do not involve any violation, they belong to all the repairs of database regardless of the repair semantics. Clearly, the set of CQA can be larger by including the tuples that involve violations and leveraging a given repair semantics. More differences between CQA and our method are addressed in Section 7.
6. INCONSISTENCY-AWARE RANKING

In this section, we study top-k ranking of inconsistency-aware tuples as defined above. To this end, we leverage the introduced inconsistency measures. We consider as input an annotated instance $I^t$, where each base tuple $t$ of a relation $R$ is annotated with the monomial $I^t (R)(t)$ (c.f. Definition 3). Let $\alpha$ be either $CBM$ or $CBS$, the main idea is to use $\alpha$ as a scoring function over the results of a query $Q$ where the score of an output $t$ of $Q$ is given by $CBM(t, Q, I, DC) = W(I)$ (respectively, $CBS(t, Q, I, DC) = |Var(M)|$), with $M = Q(I^t)(t)$. The goal is then to rank the answer tuples while taking into account the inconsistency degrees of the base tuples contributing to the answers. The fundamental computation problem is then to be able to efficiently enumerate (part of) query answers in a specific order w.r.t. their inconsistency degrees. We focus on one particular instance of this problem where the goal is to return the query results with the top $k$ scores, hereafter called inconsistency-aware top-k ranking.

Definition 5 (Top-k queries). Let $I$, $Q$ and $DC$ be respectively a database instance, a conjunctive query and a set of denial constraints over the same instance. Let $k$ be an integer, let $\alpha \in \{CBM, CBS\}$. The top-k query answers of a query $Q$ using the inconsistency measure $\alpha$, denoted by $Q^{k,\alpha}(I)$, is defined as follows:

(i) $Q^{k,\alpha}(I) \subseteq Q(I)$,

(ii) $Q^{k,\alpha}(I) = \text{Min}(k, |Q(I)|)$, and

(iii) $\forall (t_1, t_2) \in Q^{k,\alpha}(I) \times (Q(I) \setminus Q^{k,\alpha}(I))$, we have: $\alpha(t_1, Q, I, DC) \leq \alpha(t_2, Q, I, DC)$

Condition (i) and (ii) ensure that a top-k query $Q^{k,\alpha}(I)$ computes at most $k$ answers of $Q(I)$ while condition (iii) ensures that the computed answers are the best answers in $Q(I)$ w.r.t. the inconsistency measure $\alpha$. The following example illustrates a top-k query over our running example.

Example 8. Assume $k = 1$ and $\alpha = CBM$. Continuing with the database $hdb$ and the query $Q_{ex}$ as in Figure 7, we have: $Q^{1,\alpha}_{ex}(hdb) = \{(d4, d4)\}$.

A naive approach to solve a top-k query $Q^{k,\alpha}(I)$, with $\alpha$ being one of our inconsistency measures, would be to entirely compute and sort, w.r.t. $\alpha$, the set $Q(I)$ and then filter out the $k$ best answers. Such a naive approach is clearly suboptimal in particular when $k \ll |Q(I)|$.

The problem of computing $Q^{k,\alpha}(I)$ answers falls in the general setting of rank join problems [28]. Performance of rank join algorithms have been deeply studied in the case of monotonic scoring functions [28, 31] while very few works deal with nonmonotonic functions.

Let us first recall the monotonicity property. Let $f$ be a scoring function. The function $f$ is monotonic w.r.t. an operator $\circ$ iff: $f(x) \leq f(y) \land f(z) \leq f(v) \Rightarrow f(x \circ z) \leq f(y \circ v)$, $\forall x, y, z, v$. Otherwise, $f$ is said to be non-monotonic w.r.t. $\circ$.

The following lemma says that inconsistency degrees based on multi occurrence quantification are monotonic w.r.t. the join operator $\oplus$ while the inconsistency measures based on the single occurrence quantification are nonmonotonic w.r.t. $\times$.

Lemma 2. A scoring function that associates to each tuple $t$ a score computed using CBM (respectively, using CBS) is monotonic (respectively, non-monotonic) w.r.t. $\times$.

We present in the next section an integrated algorithm to handle both monotonic and non-monotonic inconsistency measures in the case of bag semantics and we prove the optimality of the algorithm.

6.1 The TopINC algorithm

In this section, we present our top-k ranking algorithm, called TopINC, for top-k queries under both inconsistency measures $CBM$ and $CBS$. A general idea behind existing rank join algorithms [28] is to process the tuples of the relations involved in a given query in a specific order by considering at each step the most promising tuples at first. However, when the scoring function is not monotonic with respect to the join operator, which is the case for $CBS$ due to single-occurrences, it is not straightforward to identify the order in which tuples should be processed.

The core intuition of our top-k ranking algorithm consists of using an index based on the inconsistency measures to access the most promising tuples at each step of query processing. More precisely, we build for each relation $R$, an associated index, denoted $ind(R)$, whose nodes are labeled by a subset of the constraints. More precisely, each node $B$ of $Ind(R)$ is labeled with $label(B) \subseteq DC$. A node $B$ stores the set of tuple’s ids of $R$, denoted $B(R)$, that violate exactly the set of constraints $label(B)$, i.e., the set $B(R) = \{t \in R \mid VC(I, DC, t) = B\}$. We call $B$ a bucket of the index $Ind(R)$ and the set $B(R)$ the bucket content.

Example 9. In our example, the buckets of the index are labeled with subsets of the constraints $DC = \{C_1, C_2, C_3\}$. For instance, a bucket $B_0$, with $label(B_0) = \emptyset$, will store the consistent tuples $B_0(D) = \{t_1\}, B_0(S) = \{t_5, t_6\}$ and $B_0(V) = \{t_8\}$ of the relations $D, S$ and $V$. The labels of the buckets and the content of the buckets are given below:

<table>
<thead>
<tr>
<th>$B$</th>
<th>$B_0$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
<th>$B_5$</th>
<th>$B_6$</th>
<th>$B_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$label(B)$</td>
<td>$\emptyset$</td>
<td>${C_1}$</td>
<td>${C_2}$</td>
<td>${C_3}$</td>
<td>${C_1, C_2}$</td>
<td>${C_1, C_3}$</td>
<td>${C_2, C_3}$</td>
<td>${C_1, C_2, C_3}$</td>
</tr>
<tr>
<td>$B(D)$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>${t_2}$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$B(S)$</td>
<td>${t_5, t_6}$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>${t_8}$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$B(V)$</td>
<td>${t_8}$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Each index $ind(R)$ is defined as an ordered list of the nodes containing fragments of $R$, where nodes are ordered with respect to the cardinality set of their labels, i.e., $B \leq B'$ iff $|label(B)| \leq |label(B')|$. 

Example 10. The following indexes are associated with the relations $D, S$ and $V$ of our example database: $Ind(D) = \{B_0, B_3\}$, $Ind(V) = \{B_0, B_6\}$ and $Ind(S) = \{B_0, B_1, B_5\}$.

We now describe the pseudocode of the Algorithm in detail. Let $Q$ be a query $Q(u) \leftarrow R_1(u_1), \ldots, R_m(u_m)$. In order process the query $Q^{\alpha,\circ}$, with $\alpha \in \{CBS, CBM\}$, the algorithm TopINC (c.f. Algorithm 1) takes as input: the query $Q^{\alpha,\circ}$, the indexes $ind(R_i)$, with $i \in [1, m]$, one for each input relation and the set $ViolC$ of the violated constraints obtained by unioining the labels of all the buckets in the indexes. The algorithm uses as many temporary buffers $H[R_i]$, $i \in [1, m]$ as the number of indexes. Each temporary buffer contains the bucket contents. In addition, the algorithm uses a vector $jB$ of size $m$ storing the ids of the buckets that need to be joined during the course of the algorithm. The algorithm TopINC follows a level-wise sequencing of the iterations, where each level denotes the inconsistency degree of the answers computed at this level (c.f., lines 8-10 of algorithm 1). Level 0 means that consistent tuples are processed.
while the subsequent level $l$ indicates the number of constraint that are considered for the violated tuples. When processing a given level $l$, the algorithm resets the variable $curVSet$, used to keep track of the violated constraints while exploring the input relations at level $l$, and makes a recursive call to the $I terateJoin$ procedure (line 6). For each level $l$, the $I terateJoin$ procedure (Algorithm 2) explores the input relations sequentially from $R_1$ to $R_n$ (lines 11 to 14). For each input relation $R_p$, $I terateJoin$ uses the index $ind(R_p)$ to identify the buckets of $R_p$ that are worthwhile to consider for the join (i.e., the buckets to be loaded in $jB[p]$), i.e., those buckets whose label size’s is less than $l$ (line 5). The relevant bucket ids of input relations are loaded in the $jB$ buffer (line 13 of algorithm 2) and when $R_p$ is the last input relation (i.e., $p = m$) (line 15), a join is performed between the buffers of $jB$ (lines 17 to 20) of the algorithm 2 in order to compute the answers with inconsistency degree equal to the current level $l$. Note that the variable $curVSet$ will contain duplicate occurrences if $\alpha = CBM$ (line 10) and single occurrences if $\alpha = CBS$ (line 7). It enables us to keep track of the current level of inconsistency while exploring the inputs. When intermediate inputs are explored, the $I terateJoin$ algorithm ensures that $|curVSet|$ does not exceed the current inconsistency level $l$ (line 5 and line 11). When the last input is processed, a join is performed between the buckets in $jB$ only if $|curVSet| = 1$ (line 15) which ensures that the computed answers have $l$ as inconsistency degree.

Example 11. We assume that the input relations are processed in this order: $D,S$ and $V$. We illustrate the processing of query $Q^{CBS}_x(I)$ by $TopINC$. Figure 3 exemplifies the iterations of the algorithm. The gray cells, in each step, denote the newly read tuples in that step. The final result (Level 3) is reported at the bottom of the Figure 3. Starting from level 0, the selected buckets are $jB = [B_0, B_0, B_0]$, thus leading to join the contents of these buckets only at the very beginning. The contents of $HV(B_0)$, $HS(B_0)$ and $HD(B_0)$ are shown in Figure 3. The first answer corresponding to the join of the above buffers is found (i.e., $res = (\{d,A\})$). The $TopINC$ continues to read in selected buckets in $jB$. It then tries to read in $B_0$ of $V$ but no additional tuples, then it moves to read in $B_0$ of $V$. Next, the tuple $(02,2,5)$ is loaded in $HV(B_0)$ but no additional answer is found. As there is no additional answers (because $k=2$) and all tuples are read in selected buffers, $TopINC$ jumps to the next level, i.e., the level 1. At this level, the first selected buckets are $jB = [B_0, B_1, B_0]$, thus leading to read the only one tuple of bucket $B_1$ of relation $V$ into $HV(B_1)$. But no additional answer is found and no others buckets in level 1 can be selected bringing $TopINC$ to level 2. Finally, $TopINC$ processes level 2 and then level 3 as shown in Figure 3 and halts when it reaches $|res| = 2$.

The following two lemmas state the correctness and the worst case complexity of $TopINC$, respectively.

**Lemma 3.** Let $\alpha \in \{CBM,CBS\}$. The algorithm $TopINC$ computes correctly $Q^{k,\alpha}(I)$.

**Lemma 4.** Let $Q(X) \leftarrow R_1(X_1),\ldots,R_m(X_m)$ be a conjunctive query, let $s = |X|$ and let $k$ be an integer. A query $Q^{k,\alpha}$ is evaluated by the $TopINC$ in time $O(n^m)$ and in space $O(I[K] + k \times s)$.

Unsurprisingly, and like most of state of the art top-k join algorithms that do not use specific knowledge related to the join attributes (e.g., see [28, 31]), the worst case time complexity of $TopINC$ is in $O(n^m)$, with $m$ the number of input relations in $Q$. Interestingly, the previous lemma also provides a tighter upper bound regarding the space complexity. This lemma is a consequence of the level-wise approach followed by $TopINC$ which ensures that query answers are computed in the ascending order of their inconsistency degree. Hence, $TopINC$ strictly computes the answers that belong
to the output (and the algorithm stops when $k$ answers are computed). All the proofs are reported in appendix section. **Optimality of TopINC**. As discussed in Section 3, the presence of the non-monotonic function $CBS$ prevents us from exploiting the existing frameworks [30, 31] to analyze the performance of TopINC. We define hereafter a new class of algorithms, called Semi-Blind Algorithms (SBA), which includes the algorithms that exploit only the $prov$ column (i.e., the scoring function) without imposing any predefined data access strategy. We show that in this general class, and modulo non-deterministic choices, TopINC is optimal w.r.t. the number of tuples read from the inputs.

We start by defining the family of SBA algorithms as follows. We define $S_A$ as a database schema where each table $R$ has a set of attributes $A$. Let $I_1$ and $I_2$ be two relational instances over $S_A$, we define equivalent instances under attributes $A$: $I_1 \equiv_A I_2$ iff $\forall R \in S_A$, $\pi_A(I_1(R)) = \pi_A(I_2(R))$ with $\pi_A(R)$ the projection over table $R$ on attributes $A$.

Let $Q(X)$ $\leftarrow$ $R_1(X_1)$, ..., $R_n(X_n)$ and $\mathcal{I}$ be respectively a conjunctive query and an instance over schema $S_{\text{prov}}$. Let $AL$ be an algorithm that allows to compute the answers of $Q^{k,\alpha}(\mathcal{I})$, with $\alpha \in \{CBS, CBM\}$. A join test $J$ of $AL$ when processing $Q^{k,\alpha}$ over an instance $\mathcal{I}$ has the following form $J = t_1 \bowtie \ldots \bowtie t_m$ where $t_i \in \mathcal{I}(R_i)$ with $i \in [1, m]$. A score of a join test is defined as follows: $s(J) = s(t_1 \bowtie \ldots \bowtie t_m)$. We define $jPath(AL, Q, \mathcal{I})$ as the sequence of join tests performed by the algorithm $AL$ when evaluating $Q^{k,\alpha}$ over $\mathcal{I}$. The intuition is that $jPath(AL, Q, \mathcal{I})$ enables to capture the behavior of the algorithm $AL$ when it evaluates $Q^{k,\alpha}$ over $\mathcal{I}$. Formally, $jPath(AL, Q, \mathcal{I}) = [J_1, ..., J_n]$, where $J_i = t_1^{(i)} \bowtie \ldots \bowtie t_m^{(i)}$, with $t_l^{(i)} \in \mathcal{I}(R_l)$ and $j \in [1, m]$. We define $label(J_i) = (\pi_{\text{prov}}(t_1^{(i)}), \ldots, \pi_{\text{prov}}(t_m^{(i)}))$. We are now ready to give the formal definition of the class of Semi-Blind Algorithms (SBA), the algorithms that use only information from the $prov$ column when processing inconsistency-aware queries.

**Definition 6 (SBA algorithms).** Let $Q^{k,\alpha}$ be a top-$k$ query, let $I_1, I_2$ be two instances, and let $AL$ be an algorithm such that: $jPath(AL, Q, I_1) = [J_1, ..., J_n]$ and $jPath(AL, Q, I_2) = [J_1, ..., J_n]$. The algorithm $AL$ belongs to the class SBA iff: $I_1 \equiv_{\text{prov}} I_2 \Rightarrow label(J_i) = label(J'_i)$, for $i \in [1, \max(n_1, n_2)]$.

According to this definition, an algorithm $AL \in SBA$ will have a similar behavior when evaluating a query $Q^{k,\alpha}$ over different instances that are equivalent under $prov$; i.e., $AL$ will explore the same sequence of join tests but may stop earlier or later depending on the success or failure of the join tests. The outcome of this latter test is related to the content of the join attributes of each specific input instance and remains independent from the $prov$ column. As one can easily notice, TopINC belongs to SBA. Indeed, it only relies on the information given by the $prov$ column without exploiting any auxiliary information. A natural metric to measure the performance of an algorithm $AL$ is to compute the number of tuples of the input relations accessed by $AL$, denoted $cost(AL, Q, \mathcal{I})$.

Let $J = t_1 \bowtie \ldots \bowtie t_m$, we denote by $\text{tuple}(J) = \{t_1, \ldots, t_m\}$ the set of tuples involved in a join test $J$. Consider an algorithm $AL$ with $jPath(AL, Q, \mathcal{I}) = [J_1, ..., J_n]$.  

Then, the cost of $AL$ is given by: $cost(AL, Q, \mathcal{I}) = | \bigcup_{i=1}^{n} \text{tuple}(J_i) |$. The following lemma states that there exists no algorithm in the class SBA that is optimal w.r.t. the cost metric defined above.

**Lemma 5.** Let $Q^{k,\alpha}$ be a top-$k$ query. Then, we have: $\forall AL_1 \in SBA$, $\exists AL_2 \in SBA$ and an instance $\mathcal{I}$ such that, $cost(AL_1, Q, \mathcal{I}) > cost(AL_2, Q, \mathcal{I})$.

The intuition behind the above lemma is that the SBA algorithms need to make a non-deterministic choice among the join tests that have equivalent score. We illustrate this case by relying on a corner case of an instance $\mathcal{I}$ containing exclusively consistent tuples. Consider a query $Q$ over $n$ inputs $R_1, \ldots, R_n$ such that $|Q(\mathcal{I})| = 1$. Consider now the evaluation of the query $Q^{1,CBS}$ by SBA algorithms. Since all the join tests among the tuples of the input relations will have the same score, an SBA algorithm needs to make a non-deterministic choice among the elements of $R_1 \times \ldots \times R_n$ to decide in which order the join tests will be performed. Hence, the best algorithm $AL$ would luckily pick the right tuples in the first round of choice which leads to an optimal cost: $cost(AL, Q, \mathcal{I}) = n$. The worst-case algorithm $AL'$ might end up with the least cost after exploring the entire cartesian product of the inputs, which leads to $cost(AL', Q, \mathcal{I}) = \sum_{i=1}^{n} |R_i|$ (i.e., the algorithms $AL'$ needs to read the entire inputs). Consequently, we argue that it is not worthwhile to distinguish between SBA algorithms w.r.t. to the order of exploring join tests with equivalent score (since this is a non-deterministic choice). We formalize this notion using regions, defined as maximal subsequences of join tests with equivalent score, and we define a new metric based on the number of regions explored by the algorithm. The region-based cost enables us to get rid of lucky choices when comparing the performances of SBA algorithms.

Below, we define the notion of a region. Let $jPath(AL, Q, \mathcal{I}) = [J_1, ..., J_n]$. A region of $jPath(AL, Q, \mathcal{I})$ is a maximal subsequence of $jPath(AL, Q, \mathcal{I})$ made of join tests with equal inconsistency degree. More formally, a sequence $[J_l, ..., J_n]$ with $l \leq p$, is a region of $jPath(AL, Q, \mathcal{I})$ if $l, p \in [1, n]$, and: (i) $\alpha(J_i) = \alpha(J_l), \forall i \in [l, p]$ and (ii) $\alpha(J_{i+1}) \neq \alpha(J_i)$. We define $\text{Regs}(AL, Q, \mathcal{I})$ to be the set of regions of $jPath(AL, Q, \mathcal{I})$. We define the cost model $\text{cost}^\mathcal{V}(AL, Q, \mathcal{I})$ as the number of regions explored by the algorithm $AL$ during processing of query $Q$ over $\mathcal{I}$: $\text{cost}^\mathcal{V}(AL, Q, \mathcal{I}) = |\text{Regs}(AL, Q, \mathcal{I})|$.

The introduced cost model $\text{cost}^\mathcal{V}(AL, Q, \mathcal{I})$ conveniently prevents an algorithm to compute an answer that can be dropped after the top-$k$ answers, thus avoiding more useless I/O operations, as proved in the following theorem.

**Theorem 1.** For any instance $\mathcal{I}$ and any top-$k$ conjunctive query $Q^{k,\alpha}$, we have: $\text{cost}^\mathcal{V}(\text{TopINC}, Q, \mathcal{I}) \leq \text{cost}^\mathcal{V}(AL, Q, \mathcal{I})$, $\forall AL \in SBA$.

Notice that the TopINC algorithm is only sensitive to the class of queries and unsensitive to the class of constraints, the latter being hardcoded in the inconsistency annotations. TopINC and Theorem are established for conjunctive queries, which are by definition monotonic queries (i.e. for which a partial output can be computed starting
Table 3: Datasets used in our empirical evaluation.

7. EMPIRICAL EVALUATION

Our experimental assessment has been carried out with two main objectives in mind. In the first part, we aimed at measuring the time required to transform a database instance into a N[T]-instance. The latter transformation is considered as the pre-processing time for our approach. In second part, we empirically evaluate the TopINC algorithm to perform top-k query answering. We have implemented our framework in PostgreSQL 10 by leveraging PSQL and JDK 11. All the experiments have been executed on a DELL Core i7 2.5 GHz laptop with 16 GB RAM running Linux OS. The datasets used in our study are summarized in Table 3. As real-life datasets, we employed the Hospital, Tax and Pstock datasets from [17] as they are the only datasets in the literature that are equipped with denial constraints.

In addition, we generated our own synthetic datasets and companion denial constraints.

In Table 3, the column Inc denotes the percentage of inconsistency per relation, whereas #Top and #Rel denote the number of tuples and relations in the dataset, respectively. Finally, #Cons indicates the number of denial constraints per dataset. The column #atom gives interval of number of atoms for the constraints of a given dataset. The column Syn gives the type of dataset (synthetic or not). All the queries used are described as follows: Q1 to Q5 are on the Hospital dataset and they contain one join; Q6 to Q9 are on the Tax dataset and they contain one join; Q10 to Q14 are on the synthetic dataset; Q10, Q11 have a join across three tables, Q12, Q13 have a join across four tables and Q14 has a join across five tables.

Table 3 shows the constraints in each dataset. Columns #Equal and #!(In-)equal provide the number of constraints that use exclusively equality predicates and those using a mixture of predicates from \{≤, ≥, ≠, <, >, =\}, respectively. As one can notice, the latter are in greater number.

Runtimes of the instance transformation. In this section, we measure the runtimes of the transformation of a database instance into a N[T]-instance. Figure 4 shows the runtimes for each dataset while varying the number of constraints. We can observe that the runtimes of the K-instance transformation linearly scale with the number of constraints for all datasets. They range between tens and thousands of seconds, depending on the dataset. We observed the highest runtimes only with one dataset (Tax), which has fifty denial constraints and took approximately 1h to transform 100000 tuples. Such a transformation is part of the pre-processing and only done once for the annotated instances, thus it remains quite reasonable. One can also notice that there is a huge gap between runtimes of transformation of Tax (Figure 4a) and Hospital (Figure 4b) despite the fact that these two datasets have similar characteristics. The observed gap is due to the fact that the constraints in the dataset Tax are more sophisticated than the constraints in the dataset Hospital (i.e., with larger built-in atoms).

Overhead of inconsistency measures on query execution. In order to gauge the overhead of running a query Q with inconsistency degrees, we have employed our 14 queries and measured the overhead per each tuple in the answer (Table 5a) as well as the total overhead for the complete output of the query (Table 5b). The obtained results are reported in Table 5. The greater is the size of the output of a query the larger is the overhead of query execution with inconsistency degrees. The columns query and #answers in Table 5a are the total query runtime of the original query on an inconsistency-free database instance and the size of query answer, respectively. Depending to the size of output of the queries, the overhead ranges between 2ms and 6m, respectively for queries Q4 and Q12. (respectively, yellow and red cells in Table 5a). The difference can be explained by looking at the size of the answer set of Q4 and Q1 that are 50 tuples and 15M tuples, respectively. These overhead are, however, not trustworthy to understand the overhead of our approach, since they concern the entire output set of queries, whereas our algorithms returns the top-k results tuple by tuple. Thus, one should look at the overhead per tuple in Table 5b. We can observe that the overheads per tuple are reasonable in all cases, and range between 5 microseconds and 293 microseconds (yellow and red cells in Table 5b).

TopINC Performance vs. baseline. We have implemented a baseline algorithm leveraging PostgreSQL, where all answers of a query are computed beforehand and then sorted (ORDER BY) and filtered (LIMIT k). Figure 7
shows the performance of Top INC (with $k$ varying from 10 to 300) as opposed to the aforementioned baseline algorithm. We have chosen five queries as representatives of different datasets and join sizes ranging from one join ($Q_1, Q_2, Q_8$) and three joins ($Q_{11}$) to five joins ($Q_{14}$). The algorithm Top INC (blue bar) can be up to $2^8$ times faster than the baseline approach as shown in Figure 7a. There is only one query, i.e. the most complex query $Q_{14}$, for which Top INC has lower performance compared to the baseline, starting from a value of $k \geq 200$. The reason for that is the fact that Top INC for higher values of $k$ and higher number of joins in the query will inspect more buckets and try to perform more joins that will likely produce no answers. Furthermore, notice that in all these experiments the baseline turns to have advantageous with respect to Top INC, as it leverages the query planning of Postgres and opportunistically picks the most efficient join algorithm. Despite these advantages, our approach is still superior in terms of performance in the majority of the cases. In Figure 7 from 7a to 7j, we measure the memory consumption of our approach for the same queries. We observe that Top INC always consumes less memory than the baseline.

We also ran another experiment on synthetic datasets to study the impact of other parameters on the performance of Top INC. The results are reported in Figure 6. Precisely, we wanted to study the dependency of Top INC on the following parameters: the number of answers to be returned ($k$), the number of violated constraints ($DC$), the size of search space formed by $DC$ (i.e. the number of subsets of constraints violated by base tuples) and the exact size of output of $Q$ (i.e. $|Q(I)|$). In these results, we focused on a relatively simple query $Q$ containing a join between two synthetic relations (of size 1000 each one) and kept constant the value of $k$ (equal to 20). We ran this experiment with varying number of denial constraints $DC$ from 10 to 30, respectively in Figures 6a, 6b and 6c. In each of these plots, we vary the selectivities of $Q$ and the size of the search space of the algorithm. One can see that Top INC outperforms the baseline in all cases and is particularly advantageous with larger query outputs and smaller search space. The underlying reason is that the greater is the output size of the query, the larger is the probability to find answers within the first combinations scanned within the search space.

**Qualitative evaluation.** We have designed an experiment devoted to show the utility of our inconsistency measures on real-world inconsistent data. We chose two real-life datasets, namely Adult, used in [1,37] and containing census data, along with and Food Inspection [2] including information about inspections of food establishments in Chicago. The features of the two datasets are shown in Table 4a. Table 4b reports the the constraints of Adult, namely $A_1$ and $A_2$, that have been derived using Holoclean [38]. While $A_1$ indicates that men who have ‘married’ as marital status are husbands, $A_2$ expresses the dual constraint for women. In addition, we handcrafted a third constraint $A_3$ establishing that adults who are not in a family should not have ‘married’ as marital status. This third constraint allows to capture violated tuples that overlap with the tuples violated by the two former constraints. We also built meaningful denial constraints for the second dataset as shown in Table 4c. The constraint $F_1$ (respectively, $F_2$) states that a licence number, which is a unique number assigned to an establishment, uniquely identifies the legal name of the establishment (respectively, its risk category). The constraint $F_3$ states that if a given establishment has been inspected for ‘consultation’ at a date $d$, one cannot expect to have an inspection of a different type prior to $d$ for the same establishment. This is because the attribute Inspection type takes the value ‘consultation’ when the inspection “is done at the request of the owner prior to the opening of the establishment.”

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>0</td>
<td>276.9</td>
<td>0</td>
<td>1</td>
<td>99.9</td>
</tr>
<tr>
<td>$F_1$</td>
<td>1</td>
<td>1</td>
<td>28.0</td>
<td>2</td>
<td>2</td>
<td>10.4</td>
</tr>
<tr>
<td>$F_2$</td>
<td>1</td>
<td>1</td>
<td>13.0</td>
<td>2</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>$F_3$</td>
<td>1</td>
<td>1</td>
<td>23.0</td>
<td>2</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>$F_1F_2F_3$</td>
<td>1</td>
<td>1</td>
<td>23.0</td>
<td>2</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>$F_1F_2F_3$</td>
<td>1</td>
<td>1</td>
<td>23.0</td>
<td>2</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>(a) Data Inconsistency.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Obs.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6239</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>495</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>(c) Constraints on Food Inspection.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Obs.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>114150.0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3183</td>
</tr>
<tr>
<td>(c) Top-100 A1 Ans.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Obs.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>104289.0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3183</td>
</tr>
<tr>
<td>(d) Distrib. of A1 Ans.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Obs.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6239</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>495</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>(e) Comparison with CQA.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs.</th>
<th>Obs.</th>
<th>Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>114150.0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3183</td>
</tr>
<tr>
<td>(d) Distrib. of A1 Ans.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Real-world datasets with their denial constraints.
surface”, followed by an inspection issued as a response to a complaint and then a non-failing normal inspection.

Table 5.a. shows the violations of constraint for the two datasets. We can notice that there are different kinds of tuples returned by AQ1 as illustrated in Table 5.b. The majority of the results (276M tuples) are consistent, thus both CBS and CBM are equal to 0, while the remaining answers exhibit 1 or 2 as inconsistency degrees. Most of the inconsistent tuples violate one constraint at a time (in the order A3, A1 and A2) while the rest of the tuples violate two constraints. For this dataset, in the majority of the cases a constraint is violated at most once and hence CBS and CBM measures are not discriminating, except for the 23 answers where violations occur twice (5th line of Table 5.b). These tuples are captured when running TopInc for top-100 tuples starting from the most inconsistent ones as illustrated in Table 5.c. Note that, while CBM does not distinguish between the 71 most inconsistent answers of AQ1 (answers with CBM = 2 corresponding to the first four rows of Table 5.c), the CBS measure provides a different ranking for the 23 answers of the 4th row of this Table.

We show in Table 5.d the inconsistency degrees of the answers when evaluating query FQ1 on the second dataset. Note that for this query the CBS degrees vary from 0 up to 3 while the CBM degrees range from 0 up to 9. We observe now that many answers are not distinguishable under CBS while they exhibit a wider range of CBM degrees (e.g., CBM varies from 1 to 3 for answers having CBS = 1). This again shows that CBS and CBM provide the user with two different types of information, both being useful to carry out the ranking. Furthermore, by looking at the tuples that contribute to the last row of Table 5.d (CBM = 9) violating the three constraints F1, F2 and F3 in tandem, we made an interesting discovery. These violations (and also those corresponding to CBM = 8) are all related to the same erroneous license numbers (all set to 0 in the corresponding tuples).

Finally, we show by means of examples the remarkable difference between our approach and CQA. Note that as we already pinpointed in Section 9 these are complementary approaches. Table 5.e shows three CQA-consistent answers for query FQ1. First, we note that all our consistent answers (i.e., with CBS = CBM = 0) are also CQA-consistent as stated in Lemma 1. The converse is not true as it can be observed in Table 5.e where the answer (34183) is CQA-consistent but not consistent in our framework (with CBS and CBM ≠ 0). On another note, we can notice that the CQA approach does not distinguish between the three CQA-consistent answers of Table 5.e. In particular, the information that the CQA-consistent answer (34183) is computed using inconsistent base tuples (violation of F1) is not conveyed by CQA.

8. CONCLUSION AND OUTLOOK

We have presented a novel framework for inconsistency-aware query answering leveraging two different measures of inconsistency. We have grounded the computation of inconsistency degrees in why-provenance annotations and we have designed a novel top-k rank algorithm suitable for the above measures while proving its optimality.

As future work, we plan to investigate the extension of our approach to other classes of constraints and queries, such as universal constraints and aggregate queries. Techniques developed in the literature e.g., [18] for tracing lineage of first order queries could be employed to extend our approach to the upper class of constraints beyond DCs (e.g., universal constraints). However, in the presence of negation, the connection between provenance and inconsistency degrees of base tuples remains unclear and raises intriguing research questions. On the other hand, while recent literature, e.g., [4], provides foundations to study the computation of inconsistency measures for aggregate queries using polynomial provenance, the problem of efficiently computing top-k aggregate queries in presence of non-monotonic scoring functions such as CBS remains open and unsolved up to date.

Handling updates of tuples or denial constraints in our framework is also in our research agenda. One has to resort to classical incremental view maintenance mechanisms. However, handling constraint modifications without fully recomputing both the constraint-based annotations and the TopInc’s index, remains an open challenge.

9. ACKNOWLEDGEMENTS

Our research is supported by the French National Agency (ANR) through the grant nr. 18-CE23-0002 QualiHealth.

---

Footnote:
3We consider repair by deletion and symmetric set difference as measure of minimality 5.
10. REFERENCES

11. APPENDIX

Lemma 6. Let $I, Q, DC$ be defined as previously. ∀$t ∈ Q(I)$ we have: $CBM(t, Q, I, DC) = CBS(t, Q, I, DC) = 0 ⇒ t$ is a CQA.

Proof. Given an instance $I$, a query $Q$ and a set of denial constraint $DC$. An answer $t$ of $Q$ over $I$ such that $CBM(t, Q, I, DC) = CBS(t, Q, I, DC) = 0$ is a CQA because all the tuples used to compute it do not involve in any violation of constraints $DC$. So, trivially any king of answer as $t$ is CQA under any semantic of repair.

Lemma 7. A scoring function that associates to each tuple $t$ a score computed using $CBM$ (respectively, using $CBS$) is monotonic (respectively, non-monotonic) w.r.t. $\forall t \in I$.

Proof. Recall that in the context of bag semantics, the scoring functions $CBM$ and $CBS$ are applied over monomials. Let $M_1, \ldots, M_m$ be monomials corresponding, respectively, to the annotations of $m$ base tuples $t_1, \ldots, t_m$ in an $\mathbb{N}[T]$-instance $I^T$. Let $t = t_1 \times \ldots \times t_m$ be an answer to the query $Q$ over $I^T$. The annotation of $t$ using provenance polynomials semiring is given by the monomial $Q(I^T)(t) = M_1 \times \ldots \times M_m$. The monotonicity of $CBM$ is derived from the following property of the weight function $W$ (i.e., the sum of powers of variables in a monomial): $W(Q(I^T)(t)) = W(M_1 \times \ldots \times M_m) = W(M_1) + \ldots + W(M_m)$. This makes $W$ monotonic w.r.t. the product (i.e., $\times$) which implies that $CBM$ is monotonic w.r.t. $\forall t$.

We show the non-monotonicity of $CBS$ by counterexample. Assume $I^T(t_1) = C_1 C_2$, $I^T(t_2) = C_1 C_3 C_4$, $I^T(t_3) = C_2 C_4$ and $I^T(t_4) = C_1 C_2 C_3 C_4$. So, $Q(I^T(t'_1)) = C_1 C_2 C_3 C_4$ and $Q(I^T(t'_2)) = C_1 C_2 C_3 C_4$. We have $\text{Var}(I^T(t_1)) = 2 ≤ \text{Var}(I^T(t_2)) = 3$ and $\text{Var}(I^T(t'_1)) = 2 ≤ \text{Var}(I^T(t'_2)) = 3$ but $\text{Var}(Q(I^T(t'_1))) = 4 > \text{Var}(Q(I^T(t'_2))) = 3$. We can conclude that $CBS$ is a non-monotonic function.

Lemma 8. Let $α ∈ \{CBM, CBS\}$. The algorithm TopINC computes correctly $Q^{k,α}(I)$.

Proof. The TopINC algorithm proceeds one level at a time. It starts from inconsistency degree $0$ up to upper inconsistency degree $d$. At each inconsistency degree $d$, it looks for all the joins of buckets (from join relations) where the union of their labels has cardinality $d$; the join is only performed among the tuples within the buckets found. It stops processing when $k$ answers are found otherwise TopINC continues to look for the remaining joins of buckets for the same inconsistency degree $d$; if there exists no other join of buckets that leads to inconsistency degree $d$, the TopINC moves forward to inconsistent degree $d + 1$. Hence, this processing ensures that the $K$ answers output by TopINC are correct.

Lemma 9. Let $Q(X) ← R_1(X_1), \ldots, R_m(X_m)$ be a conjunctive query, let $s = |X|$ and let $k$ be an integer. A query $Q^{k,α}$ is evaluated by the TopINC in time $O(n^m)$ and in space $O(|I| + k × s)$.

Proof. This lemma is a consequence of the level-wise approach followed by TopINC which ensures that query answers are computed in the ascending order of their inconsistency degree. Hence, TopINC strictly computes the answers that belong to the output (and the algorithm stops when $k$ answers are computed).

Lemma 10. Let $Q^{k,α}$ be a top-$k$ query. Then, we have: $∀ Al_1 ∈ SBA, ∀ Al_2 ∈ SBA$ and an instance $I$ such that, $\text{cost}(Al_1, I, Q) > \text{cost}(Al_2, I, Q)$.

Proof. W.l.o.g., assume that $m = 2$ (i.e., $Q$ is a join between two relations $R_1$ and $R_2$). Take two integers $l > 1, p > 1$ such that $k ≤ l + p$. We build an instance $I_2$ as follows:

- The relation $I_2(R_1)$ contains two subsets of tuples (the $l$-tuples and the $p$-tuples): $l$ tuples $t'_1, \ldots, t'_l$ with $\pi_{\text{prov}}(t'_1) = \pi_{\text{prov}}(t'_2)$, $\forall i, j ∈ [1, l]$ and $p$ tuples $t''_1, \ldots, t''_p$ with $\pi_{\text{prov}}(t''_1) = \pi_{\text{prov}}(t''_p)$, $\forall i, j ∈ [1, p]$. We take $\pi_{\text{prov}}(t'_1) ≠ \pi_{\text{prov}}(t'_1)$.
- The relation $I_2(R_2)$ contains two subsets of tuples (the $l$-tuples and the $p$-tuples): $l$ tuples $t'_1, \ldots, t'_l$ with $\pi_{\text{prov}}(t'_1) = \pi_{\text{prov}}(t'_2)$, $\forall i, j ∈ [1, l]$ and $p$ tuples $t''_1, \ldots, t''_p$ with $\pi_{\text{prov}}(t''_1) = \pi_{\text{prov}}(t''_p)$, $\forall i, j ∈ [1, p]$.

The $\text{prov}$ values are chosen such that they satisfy the following condition: $α(\pi_{\text{prov}}(t'_1) × \pi_{\text{prov}}(t'_2)) < \min(α(\pi_{\text{prov}}(t''_1) × \pi_{\text{prov}}(t''_2)), α(\pi_{\text{prov}}(t''_1) × \pi_{\text{prov}}(t''_2)))$. Consider an algorithm $Al_1 ∈ SBA$ with $j\text{Path}(Al_1, Q, I) = [J_1, J_2]$ and $Al_2$ reads at least one useless tuple (the $p$-tuple $t''_p$ which, by construction of $I_2$ do not contribute to any answer). It is easy to see that a round robin algorithm that alternate reading $l$-tuples from $R_1$ and $p$-tuples from $R_2$ and performing join tests between the tuples loaded in the memory is optimal. Indeed, to maximize the generated answers for a given cost $s$ (i.e., reading $s$ base tuples), the best strategy is to read $[s/2]$ base $l$-tuples from $R_1$ and $[s/2]$ base $p$-tuples from $R_2$ (or inversely) which enables to compute the maximal number of $[s/2] × [s/2]$ answers. The behavior of such an algorithm $Al_2$ is given by $j\text{Path}(Al_2, Q, I) = [J'_1, \ldots, J'_k]$ with $J'_1 = t'_1 \bowtie t''_1$, $J'_2 = t'_2 \bowtie t''_2$, $J'_3 = t'_1 \bowtie t''_2$, $J'_4 = t'_2 \bowtie t''_1$, $J'_5 = t'_1 \bowtie t''_1$, $J'_6 = t'_2 \bowtie t''_2$.

Since, $Al_2$ reads the minimal number of base tuples to compute $k$ answers from $I_2$ and $Al_1$ reads at least one useless base tuple, we can conclude that $\text{cost}(Al_1, I, Q) > \text{cost}(Al_2, I, Q)$.

- Case when $J'_1 = t'_1 \bowtie t''_1$ or $J'_1 = t'_1 \bowtie t''_2$ (i.e., $Al_1$ starts reading an $l$ tuple from $R_1$). This case is the dual of the first one and can be proved following the same reasoning while inverting the roles of $p$-tuples and $l$-tuples when building $I_2$ and $Al_2$.

5This is because the optimum of the function $f(x) = sx - x^2$, for a constant $s$, is given by $x = s/2$. 

\[ \text{Cost}(Al_1, I, Q) > \text{Cost}(Al_2, I, Q) \]
Theorem 2. For any instance $I$ and any top-k conjunctive query $Q^{k,\alpha}$, we have: $\text{cost}^\mathcal{V}(\text{TopINC}, Q, I) \leq \text{cost}^\mathcal{V}(AL, Q, I)$, $\forall AL \in SBA$.

Proof. Let $Q^{k,\alpha}$, with $\alpha \in \{CBM, CBS\}$, be a top-k query over an instance $I$. Assume that there exists an algorithm $AL \in SBA$ that outputs $Q^{k,\alpha}(I)$ with $\text{cost}^\mathcal{V}(AL, Q, I) < \text{cost}^\mathcal{V}(\text{TopINC}, Q, I)$. The intuition behind our proof is that if TopINC explores a region $Z$ that is not explored by $AL$ then $AL$ is not correct (i.e., it do not correctly compute the top-k answers). Note that, when TopINC starts performing join tests with inconsistency degree $d$, it fully explores the region (i.e., performs all the possible join tests with inconsistency degree $d$) before moving to another region. In addition, TopINC processes the regions sequentially in increasing order of their inconsistency degrees (i.e, starting from region of inconsistency degree 0 to upper degrees of inconsistency). Let $Z$ be the region having the biggest inconsistency degree, noted $d_z$, among the regions explored by $AL$. Two cases occur: (i) either $AL$ has exhaustively explored all the regions with inconsistency degree $< d_z$ and in this case $\text{cost}^\mathcal{V}(AL, Q, I) \geq \text{cost}^\mathcal{V}(\text{TopINC}, Q, I)$ (because TopINC will also explore the same regions or a subset of them), or (ii) $AL$ skips some regions with inconsistency degree $< d_z$. In this case, one can prove that $AL$ is incorrect. Indeed, since $AL \in SBA$, one can build an instance $I' \equiv_{prov} I$ such that $AL$ outputs incorrect answers when it evaluates $Q^{k,\alpha}$ over $I'$. $\square$