Efficient Computation of Polynomial Explanations of Why-Not Questions
Nicole Bidoit, Melanie Herschel, Aikaterini Tzompanaki

To cite this version:
Efficient Computation of Polynomial Explanations of Why-Not Questions

Nicole Bidoit
Université Paris Sud / Inria
91405 Orsay Cedex, France
nicole.bidoit@lri.fr

Melanie Herschel
Universität Stuttgart
70569 Stuttgart, Germany
melanie.herschel
@ipvs.uni-stuttgart.de

Aikaterini Tzompanaki
Université Paris Sud / Inria
91405 Orsay Cedex, France
katerina.tzompanaki@lri.fr

ABSTRACT

Answering a Why-Not question consists in explaining why a query result does not contain some expected data, called missing answers. This paper focuses on processing Why-Not questions in a query-based approach that identifies the culprit query components. Our first contribution is a general definition of a Why-Not explanation by means of a polynomial. Intuitively, the polynomial provides all possible explanations to explore in order to recover the missing answers, together with an estimation of the number of recoverable answers. Moreover, this formalism allows us to represent Why-Not explanations in a unified way for extended relational models with probabilistic or bag semantics. We further present an algorithm to efficiently compute the polynomial for a given Why-Not question. An experimental evaluation demonstrates the practicality of the solution both in terms of efficiency and explanation quality, compared to existing algorithms.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

General Terms

Algorithms

Keywords

Why-Not question, explanation, provenance

1. INTRODUCTION

The increasing load of data produced nowadays is coupled with an increasing need for complex data transformations that developers design to process these data in everyday tasks. These transformations, commonly specified declaratively, may result in unexpected outcomes. For instance, given the sample query and data of Fig. 1 on airlines and destination countries, a developer (or traveller) may wonder why Emirates does not appear in the result. Traditionally, she would repeatedly manually analyze the query to identify a possible reason, fix it, and test it to check whether the missing answer is now present or if other problems need to be fixed.

Figure 1: Example query and data

<table>
<thead>
<tr>
<th>Airline</th>
<th>ccode</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLM</td>
<td>1919</td>
<td>1</td>
</tr>
<tr>
<td>Qatar</td>
<td>1993</td>
<td>1</td>
</tr>
<tr>
<td>Aegean</td>
<td>1987</td>
<td>2</td>
</tr>
<tr>
<td>Emirates</td>
<td>1985</td>
<td>3</td>
</tr>
</tbody>
</table>

Answering a Why-Not question, that is, understanding why some data are not part of the result, is valuable in a series of applications, such as query debugging and refinement, data verification or what-if analysis. To help developers explain missing answers, different algorithms have recently been proposed for relational and SQL queries and other types of queries (top-k, reverse skyline).

For relational queries, Why-Not questions can be answered for example based on the data (instance-based explanations), the query (query-based explanations), or both (hybrid explanations). We focus on solutions producing query-based explanations, as these are generally more efficient while providing sufficient information for query analysis and debugging. Essentially, a query-based explanation is a set of query conditions that are responsible for pruning out data relevant to the missing answers. Existing methods producing query-based explanations are not satisfactory as they are designed over query trees, making the explanations depending on the topology of a given tree. Consequently, they return different explanations for the same SQL query and may miss explanations.

Example 1.1. Consider again the SQL query and data of Fig. 1 and assume that a developer wants an explanation for the absence of Emirates from the query result. Fig. 2 shows two possible query trees. It also shows the tree operators that Why-Not [7] (◦) and NedExplain [3] (⋆) return as query-based explanations as well as the tree operators returned as part of hybrid explanations by Conseil [13, 14] (♦). Each algorithm returns a different result for each of the two query trees, and in most cases, it is only a partial result as the true explanation of the missing answer is that both the selection is too strict for the tuple (Emirates, 1985, 3) from table Airline and this tuple does not find join partners in table Country.
The above example clearly shows the shortcomings of existing algorithms. Indeed, the developer first has to understand and reason at the level of query trees instead of reasoning at the level of the declarative SQL query she is familiar with. Second, she always has to wonder whether the explanation is complete, or if there are other explanations that she could consider instead. In this paper, we make the following contributions:

**Extended formalization of Why-Not explanation polynomial.**

We have recently introduced polynomials as Why-Not explanations in the context of the relational model under set semantics [2]. A polynomial provides a complete explanation and is independent of a specific query tree representation, solving the problems illustrated by Ex. 1.1. We now extend Why-Not explanations to the relational model under bag and probabilistic semantics. This confirms the robustness of the chosen polynomial representation, making it a good fit for a unified framework for representing Why-Not explanations. In parallel, we considerably simplify our initial framework by eliminating the formerly used notion of query tableaux.

**Efficient Ted++ algorithm.** We show that a naive algorithm computing Why-Not explanations, as presented in [2], is impractical. We thus propose a novel algorithm, Ted++, capable of efficiently computing the Why-Not explanation polynomial, based on techniques like schema and data partitioning (allowing for a distributed computation) and advantageous replacement of expensive database evaluations by mathematical calculations.

**Experimental validation.** We validate our solutions both in terms of efficiency and effectiveness. Our experiments include a comparative evaluation to existing algorithms computing query-based explanations for SQL queries (or sub-languages thereof) as well as a thorough study of Ted++ performance w.r.t. different parameters. Note that such an evaluation was missing from [2].

The remainder of this paper is structured as follows. Sec. 2 covers related work. Sec. 3 defines in detail our problem setting and the Why-Not explanation polynomials. Next, we discuss the Ted++ algorithm in Sec. 4. Finally, we present our experimental setup and evaluation in Sec. 5 and conclude in Sec. 6.

## 2. RELATED WORK

The work presented in this paper falls in the category of data provenance research and specifically on explaining missing answers from query results. Due to the lack of space, we focus here on this sub-problem, thus on algorithms answering Why-Not questions, summarized in Tab. 1. This table classifies the algorithms according to the type of explanation they generate and reports the class of query and Why-Not question (simple or complex)\(^1\) they support.

**Query-based and hybrid explanations.** Why-Not [7] takes as input a simple Why-Not question and a selection, projection, join, and union (SPJU) query and returns the erroneous query operators as query-based explanations. Similarly, NedExplain [3] considers selection, projection, join, and aggregation and unions thereof (SPJUA queries) and simple Why-Not questions as well. The common drawback of the two algorithms is that their design is dependent on a specific query tree representation, thus the explanations proposed are tied to this tree. Moreover, the generated explanations are incomplete. To address this problem, Ted [2] proposes explanations in the form of a polynomial. The shortcomings of this work have already been presented in Sec. 1.

Conseil [13, 14] produces hybrid explanations that include an instance-based and a query-based component. The latter consists in a set of picky query operators. However, as Conseil considers

\(^1\)A *simple* Why-Not question involves conditions that impact one relation only, otherwise it is *complex* (see Sec. 3).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Why-Not question format</th>
<th>Query-based explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why-Not [7]</td>
<td>simple query operators</td>
<td>SPJ</td>
</tr>
<tr>
<td>NedExplain [3]</td>
<td>simple query operators</td>
<td>SPJUA</td>
</tr>
<tr>
<td>Ted [2]</td>
<td>complex polynomial</td>
<td>conj. queries with inequalities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hybrid explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conseil [13, 14]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instance-based explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MA [16]</td>
</tr>
<tr>
<td>Artemis [15]</td>
</tr>
<tr>
<td>Meliou et al. [19]</td>
</tr>
<tr>
<td>Calvanese et al. [1]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ontology-based explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cate et al. [6]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Refinement-based explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConQueR [20]</td>
</tr>
<tr>
<td>TALOS [21]</td>
</tr>
<tr>
<td>FlexIQ [17]</td>
</tr>
<tr>
<td>Zhang et al. [12]</td>
</tr>
<tr>
<td>Islam et al. [18]</td>
</tr>
<tr>
<td>WQRTQ [9]</td>
</tr>
<tr>
<td>Chen et al. [8]</td>
</tr>
</tbody>
</table>

Table 1: Algorithms for answering Why-Not questions

The above example clearly shows the shortcomings of existing algorithms. Indeed, the developer first has to understand and reason at the level of query trees instead of reasoning at the level of the declarative SQL query she is familiar with. Second, she always has to wonder whether the explanation is complete, or if there are other explanations that she could consider instead. In this paper, we make the following contributions:

**Extended formalization of Why-Not explanation polynomial.**

We have recently introduced polynomials as Why-Not explanations in the context of the relational model under set semantics [2]. A polynomial provides a complete explanation and is independent of a specific query tree representation, solving the problems illustrated by Ex. 1.1. We now extend Why-Not explanations to the relational model under bag and probabilistic semantics. This confirms the robustness of the chosen polynomial representation, making it a good fit for a unified framework for representing Why-Not explanations. In parallel, we considerably simplify our initial framework by eliminating the formerly used notion of query tableaux.

**Efficient Ted++ algorithm.** We show that a naive algorithm computing Why-Not explanations, as presented in [2], is impractical. We thus propose a novel algorithm, Ted++, capable of efficiently computing the Why-Not explanation polynomial, based on techniques like schema and data partitioning (allowing for a distributed computation) and advantageous replacement of expensive database evaluations by mathematical calculations.

**Experimental validation.** We validate our solutions both in terms of efficiency and effectiveness. Our experiments include a comparative evaluation to existing algorithms computing query-based explanations for SQL queries (or sub-languages thereof) as well as a thorough study of Ted++ performance w.r.t. different parameters.

The work presented in this paper falls in the category of data provenance research and specifically on explaining missing answers from query results. Due to the lack of space, we focus here on this sub-problem, thus on algorithms answering Why-Not questions, summarized in Tab. 1. This table classifies the algorithms according to the type of explanation they generate and reports the class of query and Why-Not question (simple or complex)\(^1\) they support.

**Query-based and hybrid explanations.** Why-Not [7] takes as input a simple Why-Not question and a selection, projection, join, and union (SPJU) query and returns the erroneous query operators as query-based explanations. Similarly, NedExplain [3] considers selection, projection, join, and aggregation and unions thereof (SPJUA queries) and simple Why-Not questions as well. The common drawback of the two algorithms is that their design is dependent on a specific query tree representation, thus the explanations proposed are tied to this tree. Moreover, the generated explanations are incomplete. To address this problem, Ted [2] proposes explanations in the form of a polynomial. The shortcomings of this work have already been presented in Sec. 1.

Conseil [13, 14] produces hybrid explanations that include an instance-based and a query-based component. The latter consists in a set of picky query operators. However, as Conseil considers

\(^1\)A *simple* Why-Not question involves conditions that impact one relation only, otherwise it is *complex* (see Sec. 3).

The remainder of this paper is structured as follows. Sec. 2 covers related work. Sec. 3 defines in detail our problem setting and the Why-Not explanation polynomials. Next, we discuss the Ted++ algorithm in Sec. 4. Finally, we present our experimental setup and evaluation in Sec. 5 and conclude in Sec. 6.

## 2. RELATED WORK

The work presented in this paper falls in the category of data provenance research and specifically on explaining missing answers from query results. Due to the lack of space, we focus here on this sub-problem, thus on algorithms answering Why-Not questions, summarized in Tab. 1. This table classifies the algorithms according to the type of explanation they generate and reports the class of query and Why-Not question (simple or complex)\(^1\) they support.

**Query-based and hybrid explanations.** Why-Not [7] takes as input a simple Why-Not question and a selection, projection, join, and union (SPJU) query and returns the erroneous query operators as query-based explanations. Similarly, NedExplain [3] considers selection, projection, join, and aggregation and unions thereof (SPJUA queries) and simple Why-Not questions as well. The common drawback of the two algorithms is that their design is dependent on a specific query tree representation, thus the explanations proposed are tied to this tree. Moreover, the generated explanations are incomplete. To address this problem, Ted [2] proposes explanations in the form of a polynomial. The shortcomings of this work have already been presented in Sec. 1.

Conseil [13, 14] produces hybrid explanations that include an instance-based and a query-based component. The latter consists in a set of picky query operators. However, as Conseil considers

\(^1\)A *simple* Why-Not question involves conditions that impact one relation only, otherwise it is *complex* (see Sec. 3).

The above example clearly shows the shortcomings of existing algorithms. Indeed, the developer first has to understand and reason at the level of query trees instead of reasoning at the level of the declarative SQL query she is familiar with. Second, she always has to wonder whether the explanation is complete, or if there are other explanations that she could consider instead. In this paper, we make the following contributions:

**Extended formalization of Why-Not explanation polynomial.**

We have recently introduced polynomials as Why-Not explanations in the context of the relational model under set semantics [2]. A polynomial provides a complete explanation and is independent of a specific query tree representation, solving the problems illustrated by Ex. 1.1. We now extend Why-Not explanations to the relational model under bag and probabilistic semantics. This confirms the robustness of the chosen polynomial representation, making it a good fit for a unified framework for representing Why-Not explanations. In parallel, we considerably simplify our initial framework by eliminating the formerly used notion of query tableaux.

**Efficient Ted++ algorithm.** We show that a naive algorithm computing Why-Not explanations, as presented in [2], is impractical. We thus propose a novel algorithm, Ted++, capable of efficiently computing the Why-Not explanation polynomial, based on techniques like schema and data partitioning (allowing for a distributed computation) and advantageous replacement of expensive database evaluations by mathematical calculations.

**Experimental validation.** We validate our solutions both in terms of efficiency and effectiveness. Our experiments include a comparative evaluation to existing algorithms computing query-based explanations for SQL queries (or sub-languages thereof) as well as a thorough study of Ted++ performance w.r.t. different parameters.

## 3. WHY-NOT EXPLANATION POLYNOMIAL

This section introduces a polynomial formalization of query-based Why-Not explanations. We assume the reader familiar with the relational model [1], and we only briefly revisit some relevant notions in Sec. 3.1 while we formalize Why-Not questions. In Sec. 3.2, we define the explanation of a Why-Not question as a polynomial. In Sec. 3.3 we provide a unified general framework for Why-Not explanations in the context of set, bag, and probabilistic semantics databases.
The notion of complex and simple conditions is extended to complex and simple Why-Not questions in a straightforward manner.

Due to lack of space, we do not provide here more real-world examples of Why-Not questions and refer the reader to scenarios in [4]. As we said, a Why-Not question \( WN \) summarizes a set of (missing) tuples that the user expected to find in the query result. To be able to obtain these missing tuples as query results, data from the input relation instances that satisfy \( WN \) need to be combined by the query. The candidate data combinations are what we call compatible tuples and can be computed using \( WN \) as in Def. 3.3.

**Definition 3.3 (Compatible Tuples).** Consider the query \( Q_{WN} = (S, A(S), WN) \), where \( S \) is also the input schema of \( Q \). The set \( CT \) of compatible tuples is the result of the query \( Q_{WN} \) over \( I \).

We further introduce the notion of a well-founded Why-Not question. Intuitively, a Why-Not question can be answered under a query-based approach, only if some data in \( I \) match the Why-Not question (otherwise instance-based explanations should be sought for). Moreover, a Why-Not question is meaningful if it tracks data not already returned by the query.

**Definition 3.4 (Well-founded Why-Not Question).** A Why-Not question \( WN \) is said to be well founded if \( CT \neq \emptyset \) and \( \pi_I[CT] \cap Q[I] = \emptyset \).

**Example 3.2.** Continuing Ex. 3.1, we may wonder why there is not a tuple for which \( R.B < S.D \) and \( T.C \leq 9 \). According to Def. 3.2, this Why-Not question can be seen as the conjunction of the conditions \( R.B < S.D \) and \( T.C \leq 9 \) (Fig. 3(c)). Since \( R.B < S.D \) is a complex condition, \( WN \) is a complex Why-Not question. The compatible tuples set \( CT \) is the result of the query \( Q_{WN} = \{ (R.B < S.D), T.C \leq 9 \} \), which contains 12 tuples.

Each tuple in \( CT \) could have led to a missing tuple, if it was not eliminated by some of the query conditions. Thus, explaining \( WN \) amounts to identifying these blocking query conditions.

**3.2 Why-Not Explanation.**

To build the query-based explanation of \( WN \), we start by specifying what explains that a compatible tuple \( \tau \) did not lead to an answer. Intuitively, the explanation consists of the query conditions pruning out \( \tau \).

**Definition 3.5 (Explanation for \( \tau \)).** Let \( \tau \in CT \) be a compatible tuple w.r.t. \( WN \), given \( Q \). Then, the explanation for \( \tau \) is the set of conditions \( E_\tau = \{ \epsilon \in C \mid \tau \not\models \epsilon \} \).

**Example 3.3.** Consider the compatible tuple \( \tau_1 \) in Ex. 3.2. The conditions of \( Q \) (see Ex. 3.1), not satisfied by \( \tau_1 \) are: \( c_1, c_3 \), and \( c_4 \). So, the explanation for \( \tau_1 \) is \( E_\tau_1 = \{ c_1, c_3, c_4 \} \).

Having defined the explanation w.r.t. one compatible tuple, the explanation for \( WN \) is obtained by simply summing up the explanations for all the compatible tuples in \( CT \), leading to the expression \( \bigcup_{\tau \in CT} E_\tau \). We justify modelling the explanation of \( \tau \) with a product (meaning conjunction) of conditions by the fact that in order for \( \tau \) to ‘survive’ the query conditions and give rise to a missing tuple, every single condition in the explanation must be ‘repaired’. The sum (meaning disjunction) of the products for each \( \tau \in CT \) means that if any explanation is ‘correctly repaired’, the associated \( \tau \) will produce a missing tuple.

Of course, several compatible tuples can share the same explanation. Thus, the final Why-Not explanation is a polynomial having as variables the query conditions and as integer coefficients the number of compatible tuples sharing an explanation.
**Definition 3.6.** (Why-Not Explanation). With the same assumption as before, the Why-Not explanation for \( WN \) is defined as the polynomial

\[
PEX = \sum_{E \in E} coe_E \prod_{\tau \in \tau} c
\]

where \( E = 2^\tau \) and \( coe_E \in \{0, \ldots, |CT|\} \).

Intuitively, \( E \) contains all potential explanations, and each of these explanations prunes from zero to at most \(|CT|\) compatible tuples. Moreover, an important property of \( PEX \) is the fact that \( coe_E = |\{\tau \in CT|E\} \) is the explanation for \( \tau \) \}, meaning that \( coe_E \) equals the number of compatible tuples with the same explanation. Each term of the polynomial provides an alternative explanation to be explored by the user who wishes to recover some missing tuples. Additionally, the polynomial offers, through its coefficients, some useful hints to users interested in the recovery of some missing tuples. Moreover, an important property of \( PEX \) is the fact that \( coe_E \) equals the number of compatible tuples with the same explanation.

The naive Ted algorithm \cite{dust} implements the definition of \cite{dust} in a straightforward manner. Essentially, Ted first enumerates the set of compatible tuples. Then, it computes the explanation for each compatible tuple, leading to the computation of the final Why-Not explanation. However, both of these steps make Ted computationally prohibitive. Not only is the computation of the set of compatible tuples time and space consuming as it often requires cross product executions, but the same holds for the iteration over this (potentially very large) set. Ted’s time complexity is \( O(n^{3}) \), \( n=\max(|\{t| R \})), R \in S \). As experiments in Sec. 5 confirm, this complexity renders Ted impractical.

To overcome Ted’s poor performance, we propose Ted++. The main feature of Ted++ is to completely avoid enumerating and iterating over the set \( CT \), thus it significantly reduces both space and time consumption. Instead, Ted++ opts for (i) iteratively over the space of possible explanations, which is expected to be much smaller, (ii) computing partial sets of passing compatible tuples, and (iii) computing the number of eliminated compatible tuples for each explanation. Intuitively, passing tuples w.r.t. an explanation are tuples satisfying the conditions of the explanation. Finally, we compute the polynomial based on mathematical calculations.

Alg. 1 provides an outline of Ted++. The input includes the query \( Q= (S, \Gamma, C) \), the Why-Not question \( WN \) and the input instance \( I \). First, in Alg. 1, line 1, all potential explanations (combinations of the conditions in \( C \)) are enumerated \( (E=2^\tau) \). The remaining steps, discussed in the next subsections, aim at computing the coefficient of each explanation. To illustrate the concepts
introduced in the detailed discussions, we will rely on our running example, for which Fig. 4 shows all relevant intermediate results. It should be read bottom-up. For convenience, in our examples, we use subscript $i$ instead of $c_i$.

The subsequent discussion on $Ted^+$ can be considered as a proof sketch of the following theorem.

**THEOREM 4.1.** Given a query $Q$, a Why-Not question $WN$ and an input instance $I$, $Ted^+$ computes exactly PEX.

### 4.1 Partial Compatible Tuples Computation

Using the conditions in $WN$, $Ted^+$ partitions the schema $S$ (Alg. 1, line 2) into components of relations connected by the conditions in $WN$ (Def. 4.1).

**DEFINITION 4.1.** (*Valid Partitioning of $S$*). Given $WN$, the partitioning of a database schema $S$ into $k$ partitions, denoted $\mathcal{P} = \{Part_1, \ldots, Part_k\}$, is valid if each $Part_i$, $i \in \{1, \ldots, k\}$ is minimal w.r.t. the following property: if $R \in Part_i$ and $R' \in S$ s.t. $\exists c \in WN$ with $A(c) \cap A(R') \neq \emptyset$ and $A(c) \cap A(R') \neq \emptyset$ then $R' \in Part_i$.

The partitioning of $S$ allows for handling compatible tuples more efficiently, by ‘cutting’ them in distinct meaningful ‘chunks’. We refer to chunks of compatible tuples as partial compatible tuples and group them in sets depending on the partition they belong to.

The set $CT_{|Part_i}$, in which $Part_i \in \mathcal{P}$ is obtained by evaluating the query $Q_{Part_i} = \{Part, A(Part), WN_{|Part_i}\}$ over $I_{|Part_i}$ (Alg. 1, line 4), $WN_{|Part_i}$ and $I_{|Part_i}$ denote the restriction of $WN$ and $I$ over the relations in $Part_i$, respectively.

**EXAMPLE 4.1.** The valid partitioning of $S$ is $Part_1 = \{R, S\}$ (because of the condition $R.B < S.D$) and $Part_2 = \{T\}$. The sets of partial compatible tuples $CT_{|Part_1}$ and $CT_{|Part_2}$ are given in the bottom line of Fig. 4.

It is easy to prove that the valid partitioning of $S$ is unique and that the set $CT$ can be computed from the sets $CT_{|Part_i}$.

**LEMMA 4.1.** Let $\mathcal{P}$ be the valid partitioning of $S$. Then, $CT = \times_{Part_i \in \mathcal{P}} CT_{|Part_i}$.

Lemma 4.1 makes it clear how to compute $CT$ from partial compatible tuples. Our algorithm is designed in a way that avoids computing $CT$ and relies on the computation of $CT_{|Part_i}$ only.

Next, we compute the number of compatible tuples pruned by each potential explanation, using the partial compatible tuple sets. In this way we calculate the coefficient of the terms in the polynomial. Since from this point on we are only handling compatible tuples, we omit the word ‘compatible’ to lighten the discussion.

#### 4.2 Polynomial Coefficient Estimation

Each set $E$ in the powerset $E$ is in fact a potential explanation that is further processed. This process is meant to associate with $E$ (i) the set of partitions $Part_E$ on which $E$ is defined, (ii) the view definition $V_E$ meant to store the passing partial tuples w.r.t. $E$, and (iii) the number $\alpha_E$ of tuples eliminated by $E$.

Alg. 2 describes how we process $E$ in ascending order of explanation size. This enables us to reuse results obtained for sub-explanations and in combination with mathematics, avoid cross product computations.

We first determine the set of partitions for an explanation $E$ as $Part_E = \cup_{c \in E} \{Part_c\}$, where $Part_c$ contains at least one relation over which $c$ is specified.

**EXAMPLE 4.2.** Consider $E_1 = \{c_1\}$ and $E_2 = \{c_2\}$. From Fig. 3(b) and the partitions in Fig. 4, we can see that $c_1$ impacts only $Part_1$, whereas $c_2$ spans over $Part_1$ and $Part_2$. Hence, $Part_{E_1} = \{Part_1\}$ and $Part_{E_2} = \{Part_1, Part_2\}$. Then, $E = \{c_1, c_2\}$ is impacted by the union of $Part_{E_1}$ and $Part_{E_2}$, thus $Part_E = \{Part_1, Part_2\}$.

We use Eq. (A) to calculate the number $\alpha_E$ of eliminated tuples, using the number $\beta_E$ of eliminated partial tuples and the cardinality of the partitions not in $part_E$.

**Intuitively, this formula extends the partial tuples to “full” tuples over $CT$’s schema.**

$$\alpha_E = \beta_E \times \prod_{Part_c \in Part_E} |CT_{|Part_c}|,$$

where $Part_E = \mathcal{P} \setminus part_E$. Note that when $part_E$ is empty, we abusively consider that $\prod = 1$.

The presentation now focuses on calculating $\beta_E$. Two cases arise depending on the size of $E$.

**Atomic explanations.** We start with atomic explanations $E$ containing only one condition $c$ (Algorithm 2 lines 3-5). We firstly compute the set of passing partial tuples w.r.t. $c$, i.e., the tuples that satisfy $c$, which we store in the view $V_c$:

$$V_c = \begin{cases} 
\pi_{(R_{id}, R \in Part_c)}(\sigma_c(CT_{|Part})) & \text{if } Part_c = \{Part\} \\
\pi_{(R_{id}, R \in Part_c \cup Part_2)}(\sigma_c(CT_{|Part_c}) \cap \neg \sigma_c(CT_{|Part_2})) & \text{if } Part_c = \{Part_1, Part_2\}
\end{cases}$$

We choose to store passing rather than eliminated tuples as they are usually less numerous. In an optimized version this decision could be made dynamically based on view cardinality estimation.
Then, the number of eliminated partial tuples by $E$ is

$$
\beta_E = \prod_{c \in \text{part}_E} |CT_{\text{part}_c} | - |V_c|
$$

**Example 4.3.** For $c_2$, we have $\text{part}_{c_2} = \{ \text{Part}_1, \text{Part}_2 \}$, so $V_{c_2} = \text{Part}_1 \cup \text{Part}_2$. For $V_{c_2}$, and by Eq. (B) we obtain $\beta_{c_2} = |CT_{\text{part}_{c_2}}| - |V_{c_2}| = 3 \times 4 = 12$. Since all partitions of $\mathcal{P}$ are in part$_{c_2}$, applying Eq. (A) results in $\alpha_{c_2} = \beta_{c_2} = 8$. For $c_3$, $\beta_{c_3} = |CT_{\text{part}_{c_3}}| - |V_{c_3}| = 2 - 2 = 0$, so $\alpha_{c_3} = \beta_{c_3} = 0$. Fig. 4 (second level) displays the process for all atomic explanations.

**Non atomic explanations.** Now, assume that $E = \{ c_1, \ldots, c_n \}$, $n > 1$ (Alg. 2, lines 6-16). For the moment, we assume that the conditions in $E$ share the same schema, so the intersection and union of $V_{c_i}$ for $i = 1, \ldots, n$ are well-defined. Firstly, we compute the view $V_E$ of passing partial tuples w.r.t. $E$ as $V_E = V_{c_1} \cap \ldots \cap V_{c_n}$. To compute the number of partial tuples pruned by $E$, we need to find the number of partial tuples pruned out by $c_1$ and ... and $c_n$, i.e., $\beta = |V_{c_1} \cap \ldots \cap V_{c_n}|$. By the well-known Demorgan law [22], we have $\beta = |V_{c_1} \cup \ldots \cup V_{c_n}|$, which spares us from computing the complements of $V_{c_i}$.

To compute the cardinality of the union of the $V_{c_i}$, we rely on the Principle of Inclusion and Exclusion for counting (11):

$$
\bigcup_{i=1}^{n} V_{c_i} = \sum_{\emptyset \neq J \subseteq \{1, \ldots, n \}} (-1)^{|J|+1} \left| \bigcap_{j \in J} V_{c_j} \right|
$$

We further rewrite the previous formula to re-use results obtained for sub-combinations of $E$, obtaining Eq. (C).

$$
\bigcup_{i=1}^{n} V_{c_i} = \bigcup_{i=1}^{n-1} \left( V_{c_i} + V_{c_{n}} \right)
+ \sum_{\emptyset \neq J \subseteq \{1, \ldots, n-1 \}} (-1)^{|J|} \left| \bigcap_{j \in J} V_{c_j} \cap V_{c_n} \right|
$$

At this point, we can compute $\beta_E$. However, so far we assume that the conditions in $E$ have the same schema. In the general case, this does not hold and we have to "extend" the schema of a view $V_E$ to the one of $V_c$, in order to ensure set operations to be well-defined. The cardinality of an extended $V_{c}^{\text{ext}}$ is given by Eq. (D).

$$
| V_{c}^{\text{ext}} | = \prod_{\text{part} \in \text{part}_E \setminus \text{part}_c} |CT_{\text{part}} | - |V_c|
$$

Based on Eq. (D) we obtain Eq. (E) that generalizes Eq. (C).

$$
| \bigcup_{i=1}^{n} V_{c_i}^{\text{ext}} | = \left| \bigcup_{i=1}^{n-1} V_{c_i}^{\text{ext}} \right| + | V_{c_n}^{\text{ext}} |
+ \sum_{\emptyset \neq J \subseteq \{1, \ldots, n-1 \}} (-1)^{|J|} \left| \left( \bigcap_{j \in J} V_{c_j} \right)^{\text{ext}} \right| \cap V_{c_n}^{\text{ext}}
$$

In Eq. (E) we have replaced the intersection with natural join. The cardinalities of the views $V_{c}^{\text{ext}} = (\bigcap_{j \in J} V_{c_j})^{\text{ext}}$ associated with $E'$ for $J \subset J_n$ have already been computed by previous steps and have only to be extended to the schema of $V_E$. When $|J| = n - 1$, then $V_E = V_c$. A detailed discussion on how and when we materialize the view $V_c$ is given shortly after.

Now, we trivially compute the number $\beta_E$ of eliminated partial tuples as the complement of $| \bigcup_{i=1}^{n} V_{c_i}^{\text{ext}} |$ (Alg. 2, line 17). The number of 'full' eliminated tuples is then calculated by Eq. (A).

**Example 4.4.** To illustrate the concepts introduced above, please follow on Fig. 4 below discussion.

For the explanation $c_2 c_3$, Eq. (E) gives:

$$
(V_2 \cup V_3)\text{ext} = |V_2^{\text{ext}}| + |V_3^{\text{ext}}| - |V_2 \cap V_3|.
$$

The schema of $\text{part}_{c_2} = \{ \text{Part}_1, \text{Part}_2 \}$ is $\Gamma_{c_2} = (R_1, S_1, T_1, T_2)$. The view $V_2$ has already a matching schema, thus $V_2^{\text{ext}} = |V_2| = 4$. For $V_3$, $\Gamma_3 = (T_3, T_4)$. We thus apply Eq. (D) and obtain $|V_3^{\text{ext}}| = |V_{\text{part}_c} \cap V_3| = 3 \times 2 = 6$. Still, $|V_3^{\text{ext}}| = |V_2 \cap V_3^{\text{ext}}|$ remains to be calculated. Intuitively, because $V_2$ and $V_3$ target schemas share attribute $T_1$, $V_2 = V_2 \cap V_3^{\text{ext}}$ is materialized and contains 2 tuples (as shown in Fig. 4). So, finally, from Eq. (E) we obtain $|(V_2 \cup V_3)^{\text{ext}}| = 4 + 6 = 10$. Since $|\text{Part}_1| \times |\text{Part}_2| = 12$ then $\beta_{c_2 c_3} = 12 - 8 = 4$, and by Eq. (A) $\alpha_{c_2 c_3} = 4$.

We now focus on the explanation $c_2 c_3$. The schemas of $V_3$ and $V_2$ are disjoint and intuitively $V_{c_2 c_3} = V_2 \cap V_3$. Here, $V_{c_2 c_3}$ is not materialized, we simply calculate $|V_{c_2 c_3}| = |V_2| = 6$. Then, $\beta_{c_2 c_3} = 12 - 12 + 6 - 6 = 0$. As we will see later, these steps are never performed in our algorithm. The fact that $c_2$ does not eliminate any tuple (see $\alpha = 0$ in Fig. 4) implies that neither do any of its super-combinations. Thus, a priori we know that $\alpha_{c_2} = \alpha_{c_2 c_3} = \ldots = 0$.

Finally, we illustrate the case of a bigger size combination, for example $c_2 c_3 c_4$ of size 3. Eq. (E) yields:

$$
(V_2 \cup V_3 \cup V_4)^{\text{ext}} = |V_2 \cup V_3| + |V_2 \cap V_3| + |V_2 \cap V_3| - |V_2 \cap V_3| - |V_2 \cap V_3|.
$$

All terms of the right side of the equation are available from previous iterations, except for $|(V_2 \cap V_3 \cap V_4)|$. As before, we check the common attributes of the views and obtain $V_{c_2 c_3 c_4} = V_2 \cap V_3 \cap V_4$. So, $|(V_2 \cup V_3 \cup V_4)^{\text{ext}}| = 6 + 6 - 2 - 1 + 0 - 9 = 6$. The number of eliminated tuples is then calculated by Eq. (D).

**View Materialization: when and how.** To decide when and how to materialize the views for the explanations, we partition the set of views associated with the conditions in $E$. Consider the relation $\sim$ defined over these views by $V_1 \sim V_2$ if the target schemas of $V_1$ and $V_2$ have at least one common attribute. Consider the transitive closure $\sim^*$ of $\sim$ and the induced partitioning of $V_E$ through $\sim^*$.

When this partitioning is a singleton, $V_E$ needs to be materialized (Alg. 2, line 9). The materialization of $V_E$ is specified by joining the views associated with the sub-CONDITIONS, which may be done in more than one way, as usual. For example, for the combination $c_2 c_3 c_4$, $V_{c_2 c_3}$ can either be computed through $V_2 \cap V_3$ or $V_2 \cap V_3$ or $V_2 \cap V_3 \cap V_4$... because all these views are known from previous iterations. The choice of the query used to materialize $V_E$ is done based on a cost function. This function gives priority to materializing $V_E$ by means of one join, which is always possible: because $V_E$ needs to be materialized, we know that at least one view associated with a sub-combination of size $n - 1$ has been materialized. In other words, priority is given to using at least one materialized view associated with one of the largest sub-combinations. For our example, it means that either $V_2 \cap V_3$ or $V_2 \cap V_3$ or $V_2 \cap V_3 \cap V_4$... is considered. In order to choose among the one-join queries computing $V_E$, we favour a one-join query $V_i \cap V_j$ minimal w.r.t. $|V_i| + |V_j|$. For the example, and considering also Fig. 4 we find that $|V_2| + |V_3| = |V_2| + |V_3| = 5$ and $|V_2| + |V_2| = 3$. So, the query used for the materialization is $V_2 \cap V_3$ (its result being empty in our example). Nevertheless, we avoid the materialization of $V_E$ if the partitioning is a singleton (Alg. 2, line 9).
We have found in Ex. 4.4 that the coefficients of its super-combinations from line 6) is about calculating the coefficient of \( \alpha \).

To make things even, the last step of \( \mathcal{E} \) may be eliminated by some super-combinations of \( \mathcal{E} \).

When computing the explanations, \( \mathcal{E} \) divides into the phases of (i) partitioning \( \mathcal{S} \), (ii) materializing a view for each partition, (iii) computing the explanations, and (iv) computing the exact coefficients. When computing the explanations, according to Alg. 2, \( \mathcal{E} \) iterates through \( 2^{|\mathcal{C}|} \) condition combinations and for each, it decides upon view materialization (again through partitioning) before materializing it, or simply calculates \( \frac{|V|}{\mathcal{E}} \) before applying equations to compute \( \alpha \).

Overall, we consider that all mathematical computations are negligible so, the worst case complexities of steps (i) through (iv) sum up to \( O(|S|+|W|N)+O(|S|)+O(2^{|C|}(|S|+|C|))+O(2^{|C|}) \). For sufficiently large queries, we can assume that \( |S|+|C|<<2^{|C|} \), in which case the complexity simplifies to \( O(2^{|C|}) \).

The complexity analysis above does not take into account the cost of actually materializing views; in its simplified form, it only considers how many views need to be materialized in the worst case. Assume that \( n=\max\{|I_\mathcal{R}|:R \in \mathcal{S}\} \). The materialization of any view is bound by the cost of materializing a cross product over the relations involved in the view - in the worst case \( O(n^{|C|}) \). This yields a combined complexity of \( O(2^{|C|}n^{|S|}) \). However, \( \mathcal{E} \) in the general case (more than one induced partitions), has a tighter upper bound: \( O(n^{k_\mathcal{E}}+n^{k_2}+\ldots+n^{k_{|C|}}) \), where \( k_\mathcal{E}=|\text{part}_\mathcal{E}| \), for all combinations \( \mathcal{E} \) and \( N=2^{|C|} \).

5. EXPERIMENTAL EVALUATION

We perform an experimental evaluation of \( \mathcal{E} \) on real and synthetic datasets. In Sec. 5.1, we compare \( \mathcal{E} \) to \( \mathcal{E} \) [2], NedExplain [3], and Why-Not [7]. Sec. 5.2 studies the runtime of \( \mathcal{E} \) w.r.t. various parameters that we vary in a controlled manner. All Java implementations of the algorithms ran on MAC OS X 10.9.5 with 1.8 GHz Intel Core i5, 4GB memory, and 120GB SSD. PostgreSQL 9.3 was used as database system.
5.1 Comparative Evaluation

The comparative evaluation to Why-Not and NedExplain considers both efficiency (runtime) and effectiveness (explanation quality). When considering efficiency, we also include Ted in the comparison (Ted producing the same Why-Not explanation as Ted++).


For each dataset, we have created a series of scenarios (crime1-gov5 in Tab. 3 - ignore remaining scenarios for now). Each scenario consists of a query further defined in Tab. 2 (Q1-Q7) and a simple Why-Not question, as Why-Not and NedExplain support only this type of Why-Not question. We have designed queries with a small set of conditions (Q6) or a larger one (Q1, Q3, Q5, Q7), containing self-joins (Q3, Q4), having empty intermediate results (Q2), as well as containing inequalities (Q2, Q4, Q5, Q6).

5.1.1 Why-Not Explanation Evaluation

Table 1 states that the explanations returned by Why-Not and NedExplain consist of sets of query conditions, whereas Ted++ returns a polynomial query conditions. For comparison purposes, we trivially map Ted++'s Why-Not explanation to sets of conditions, e.g., \(3 \cdot a + b\) maps to \(\{a, c_d, c_j\}\). For conciseness, we abbreviate condition sets, e.g., to \(c_{d4}, c_{d6}\).

Table 2 summarizes the Why-Not explanations of the three algorithms. These scenarios make apparent that the explanations by NedExplain or Why-Not are incomplete, in two senses. First, they produce only a subset of the possible explanations, failing to provide alternatives that could be useful to the user when she tries to fix the query. Second, even the explanation they provide may lack parts, which can drive the user to fruitless fixing attempts. On the contrary, Ted++ produces all the possible, complete explanations.

For the first argument, consider the scenario gov2. Why-Not and NedExplain return \(c_1\) and \(c_3\) respectively, but they both fail to indicate that both the explanations are valid, as opposed to Ted++. Then, consider crime2. NedExplain returns the join \(c_2\) (S\(_{\text{price}}\)P) - Why-Not does not produce any explanations. Ted++ indicates that except for this join, the selection \(c_3\) (S\(_{\text{name}}\)P) for instance is also an explanation. From a developer's perspective, selections are typically easier or more reasonable to change. So, having the complete set of explanations potentially provides the developer with useful alternatives.

5.1.2 Runtime Evaluation

Ted++ vs. NedExplain and Why-Not. Fig. 5 shows the runtimes in logarithmic scale for each algorithm and scenario. We observe that Ted++ and NedExplain are comparable and that in some cases, Ted++ is significantly faster than Why-Not.

For the second argument consider crime1. NedExplain returns the join \(c_1\) (C\(_{\text{sector}}\)W). The explanation of Ted++ does not contain the atomic explanation \(c_1\), but there exist combinations including \(c_1\) as a part, like \(c_{15}\). This means that the explanation by NedExplain is incomplete; a repair attempt of \(c_1\) alone will never yield the desired results. Similarly, crime7 illustrates a case, where the Why-Not algorithm produces an explanation \(c_7\) that misses some parts. Then, in gov3 NedExplain and Why-Not both return \(c_2\). However, let us now assume the developer prefers to not change this condition. Keeping in mind that those algorithms' answers may change when changing the query tree, she may start trying different trees to possibly obtain a Why-Not explanation without \(c_2\). Knowing the explanation of Ted++ prevents her from spending any effort on this, as it shows that all explanations contain \(c_2\) as a part.

By mapping the explanation of Ted++ to sets of explanations, the usefulness of the coefficients of the polynomial has been neglected. For example, the Why-Not explanation polynomial of crime8 is \(2384 + c_{23} + 20 + c_1 + 4 + c_1 + 8 + c_2\). Assume that the developer would like to recover at least 5 missing tuples, by changing as few conditions as possible. The polynomial implies to change either \(c_3\) or \(c_4\); they both require one condition change and provide the possibility of obtaining up to 20 and 8 missing tuples, respectively. \(c_3\) can recover up to 4 tuples, whereas \(c_2\) requires two condition changes. Clearly, the results of NedExplain or Why-Not are not informative enough for such a discussion.

Table 2: Queries for the scenarios in Table 3

<table>
<thead>
<tr>
<th>Query Expression</th>
<th>Scenario</th>
<th>Query</th>
<th>Why-Not question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>P sector W</td>
<td>S(<em>{\text{name}})S(</em>{\text{hair}})</td>
<td>crime1 Q1 {P.Name=Hank, C.type=Car theft}</td>
</tr>
<tr>
<td>Q2</td>
<td>P sector C2</td>
<td>S(<em>{\text{name}})S(</em>{\text{hair}})</td>
<td>crime2 Q1 {P.Name=Roger, C.type=Car theft}</td>
</tr>
<tr>
<td>Q3</td>
<td>L(_{\text{extprice}}) &gt; 100000</td>
<td></td>
<td>crime3 Q2 {P.Name=Roger, C.type=Car theft}</td>
</tr>
<tr>
<td>Q4</td>
<td>L(_{\text{extprice}}) &gt; 100000</td>
<td></td>
<td>crime4 {P.Name=Hank, C.type=Car theft}</td>
</tr>
<tr>
<td>Q5</td>
<td>S(_{\text{name}})</td>
<td></td>
<td>crime5 Q2 {P.Name=Hank}</td>
</tr>
<tr>
<td>Q6</td>
<td></td>
<td></td>
<td>crime6 Q3 {C2.Type=kidnapping}</td>
</tr>
<tr>
<td>Q7</td>
<td></td>
<td></td>
<td>crime7 Q3 {W.Name=Susan, C2.Type=kidnapping}</td>
</tr>
</tbody>
</table>

Table 3: Scenarios

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Query</th>
<th>Why-Not question</th>
</tr>
</thead>
<tbody>
<tr>
<td>crime1</td>
<td>Q1</td>
<td>{P.Name=Hank, C2.Name=Car theft}</td>
</tr>
<tr>
<td>crime2</td>
<td>Q1</td>
<td>{P.Name=Roger, C2.Name=Car theft}</td>
</tr>
<tr>
<td>crime3</td>
<td>Q2</td>
<td>{P.Name=Roger, C2.Name=Car theft}</td>
</tr>
<tr>
<td>crime4</td>
<td></td>
<td>{P.Name=Hank, C2.Name=Car theft}</td>
</tr>
<tr>
<td>crime5</td>
<td>Q2</td>
<td>{P.Name=Hank}</td>
</tr>
<tr>
<td>crime6</td>
<td>Q3</td>
<td>{C2.Type=kidnapping}</td>
</tr>
<tr>
<td>crime7</td>
<td>Q3</td>
<td>{W.Name=Susan, C2.Type=kidnapping}</td>
</tr>
</tbody>
</table>

For the second argument consider crime1. NedExplain returns the join \(c_1\) (C\(_{\text{sector}}\)W). The explanation of Ted++ does not contain the atomic explanation \(c_1\), but there exist combinations including \(c_1\) as a part, like \(c_{15}\). This means that the explanation by NedExplain is incomplete; a repair attempt of \(c_1\) alone will never yield the desired results. Similarly, crime7 illustrates a case, where the Why-Not algorithm produces an explanation \(c_7\) that misses some parts. Then, in gov3 NedExplain and Why-Not both return \(c_2\). However, let us now assume the developer prefers to not change this condition. Keeping in mind that those algorithms' answers may change when changing the query tree, she may start trying different trees to possibly obtain a Why-Not explanation without \(c_2\). Knowing the explanation of Ted++ prevents her from spending any effort on this, as it shows that all explanations contain \(c_2\) as a part.

By mapping the explanation of Ted++ to sets of explanations, the usefulness of the coefficients of the polynomial has been neglected. For example, the Why-Not explanation polynomial of crime8 is \(2384 + c_{23} + 20 + c_1 + 4 + c_1 + 8 + c_2\). Assume that the developer would like to recover at least 5 missing tuples, by changing as few conditions as possible. The polynomial implies to change either \(c_3\) or \(c_4\); they both require one condition change and provide the possibility of obtaining up to 20 and 8 missing tuples, respectively. \(c_3\) can recover up to 4 tuples, whereas \(c_2\) requires two condition changes. Clearly, the results of NedExplain or Why-Not are not informative enough for such a discussion.
### Table 4: Ted++, Why-Not, NedExplain answers per scenario

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Ted++</th>
<th>Why-Not</th>
<th>NedExplain</th>
</tr>
</thead>
<tbody>
<tr>
<td>crime1</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td>crime2</td>
<td>$c_3$</td>
<td>$c_4$</td>
<td>$c_5$</td>
</tr>
<tr>
<td>crime3</td>
<td>$c_5$</td>
<td>$c_6$</td>
<td>$c_7$</td>
</tr>
<tr>
<td>crime4</td>
<td>$c_7$</td>
<td>$c_8$</td>
<td>$c_9$</td>
</tr>
<tr>
<td>crime5</td>
<td>$c_9$</td>
<td>$c_{10}$</td>
<td>$c_{11}$</td>
</tr>
<tr>
<td>crime6</td>
<td>$c_{11}$</td>
<td>$c_{12}$</td>
<td>$c_{13}$</td>
</tr>
<tr>
<td>crime7</td>
<td>$c_{13}$</td>
<td>$c_{14}$</td>
<td>$c_{15}$</td>
</tr>
<tr>
<td>imdb1</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td>imdb2</td>
<td>$c_3$</td>
<td>$c_4$</td>
<td>$c_5$</td>
</tr>
<tr>
<td>gov1</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$c_3$</td>
</tr>
<tr>
<td>gov2</td>
<td>$c_3$</td>
<td>$c_4$</td>
<td>$c_5$</td>
</tr>
<tr>
<td>gov3</td>
<td>$c_5$</td>
<td>$c_6$</td>
<td>$c_7$</td>
</tr>
<tr>
<td>gov4</td>
<td>$c_7$</td>
<td>$c_8$</td>
<td>$c_9$</td>
</tr>
<tr>
<td>gov5</td>
<td>$c_9$</td>
<td>$c_{10}$</td>
<td>$c_{11}$</td>
</tr>
</tbody>
</table>

---

**Figure 5: Runtime for Ted++, Ted, NedExplain, and Why-Not**

Figure 5 reports runtimes for Ted on 6 out of 15 scenarios as for the others, Ted runs out of time. To examine this behavior, we compare the time distribution in Ted and Ted++ (Fig. 6). The algorithms are divided in four common phases. Note that in scenarios crime7, gov1 – gov3 the diagram bars for Ted are not totally displayed as the execution time is much higher compared to the other scenarios and to the runtime of Ted++ (the runtime of the coefficientEstimation phase is the label on the respective bars).

As said in Sec. 4, Ted’s main issue is its dependence on the number of compatible tuples. This is experimentally observed in Fig. 6: with the growth of the set of compatible tuples, the time dedicated to coefficientEstimation also grows (the scenarios are reported in an ascending order of number of compatible tuples). This can be seen in crime8 and crime7, or gov3 and gov1; while the number of tuples grows, Ted++’s runtime remains roughly steady.

#### 5.2 Ted++ Analysis

We now study Ted++’s behavior w.r.t. the following parameters: (i) the type (simple or complex) of the input query $Q$ and the number of $Q$’s conditions, (ii) the type of the Why-Not question (simple or complex) and the number and selectivity of conditions the Why-Not question involves, and (iii) the size of the database instance $I$. Note that (ii) and (iii) are tightly connected with the number of compatible tuples, which is one of the main parameters influencing the performance. Another important factor is the selectivity of the query conditions over the database.

To reduce the runtime of Ted++, we add attribute-constant comparisons (i.e., simple conditions) to the query conditions over the compatible data. This is experimentally observed in Fig. 6: the runtime of Ted++ is slower than - but still comparable to - Ted, for example in gov1–gov3. In NedExplain all compatible tuples are pruned out by conditions very close to the leaf level of the query tree, so the bottom-up traversal of the tree can stop very early. Ted++ always “checks” all conditions so cannot benefit from such an early termination. However, this runtime improvement of NedExplain often comes at the price of incomplete explanations (e.g., gov1).

**Ted++ vs. Ted.** Fig. 5 reports runtimes for Ted on 6 out of 15 scenarios as for the others, Ted runs out of time. To examine this behavior, we compare the time distribution in Ted and Ted++ (Fig. 6). The algorithms are divided in four common phases. Note that in scenarios crime7, gov1 – gov3 the diagram bars for Ted are not totally displayed as the execution time is much higher compared to the other scenarios and to the runtime of Ted++ (the runtime of the coefficientEstimation phase is the label on the respective bars).

As said in Sec. 4, Ted’s main issue is its dependence on the number of compatible tuples. This is experimentally observed in Fig. 6: with the growth of the set of compatible tuples, the time dedicated to coefficientEstimation also grows (the scenarios are reported in an ascending order of number of compatible tuples). This can be seen in crime8 and crime7, or gov3 and gov1; while the number of tuples grows, Ted++’s runtime remains roughly steady.

**5.2 Ted++ Analysis**

We now study Ted++’s behavior w.r.t. the following parameters: (i) the type (simple or complex) of the input query $Q$ and the number of $Q$’s conditions, (ii) the type of the Why-Not question (simple or complex) and the number and selectivity of conditions the Why-Not question involves, and (iii) the size of the database instance $I$. Note that (ii) and (iii) are tightly connected with the number of compatible tuples, which is one of the main parameters influencing the performance. Another important factor is the selectivity of the query conditions over the database.

For the parameter variations (i) and (ii), we use again the crime, imdb, and gov databases. To adjust the database instance size for case (iii), we use data produced by the TPC-H benchmark data generator (http://www.tpc.org/tpch/). We have generated instances of 1GB and 10GB and further produced smaller data sets of 10MB and 100MB to obtain a series of datasets whose size differs by a factor of 10. In this paper, we report results for the original query Q3 of the TPC-H set of queries. It includes two complex and three simple conditions, two of which are inequality conditions. Since the original TPC-H query Q3 is an aggregation query, we have changed the projection condition. The queries used in this section are Q1- Q12 (Tab. 2) and the scenarios are crime7-tpch (Tab. 3).

**Adjusting the query.** Given a fixed database instance and Why-Not question, we start from query Q1 and gradually add simple conditions, yielding the series of queries Q1, Q2, Q3, Q4. The evolution of Ted++ runtime for these queries is shown in Fig. 7 (a). Similarly, starting from query Q3, we introduce step by step complex conditions, yielding Q3, Q4. Corresponding runtime results are reported in Fig. 7 (b).

As expected, in both cases, increasing the number of query conditions (either complex or simple) results in increasing runtime. The incline of the curve depends on the selectivity of the introduced condition; the less selective the condition the steeper the line becomes. This is easy to explain, as the view for the explanations involving a low selective condition contains more tuples (spassing partial tuples). This, leaves space for further optimization by dynamically deciding on passing vs eliminated tuples materialization.

**Adjusting the Why-Not question.** The scenarios considered for Fig. 8 (a) have as starting point the simple Why-Not question of crime5 (see Tab. 3). Keeping the same input instance and query, we add attribute-constant comparisons (i.e., simple conditions) to WN, resulting in fewer tuples in each step. As expected, the more conditions (the less tuples) the faster the Why-Not explanation is returned, until we reach a certain point (here from crime5 on).
6. CONCLUSION AND OUTLOOK

This paper provides a framework for Why-Not explanations based on polynomials, which enables to consider relational databases under set, bag and probabilistic semantics in a unified way. To efficiently compute the Why-Not explanation polynomial under set semantics we have designed a new algorithm Ted++, whose main feature is to completely avoid enumerating and iterating over the set of compatible tuples, thus significantly reducing both space and time consumption. Our experimental evaluation showed that Ted++ is at least as efficient as existing algorithms while providing useful insights in its Why-Not explanation for a developer. Also, we show that Ted++ scales well w.r.t various parameters, making it a practical solution.

Why-Not explanation polynomials are easy to extend for unions of conjunctive queries, whereas an extension for aggregation queries is subject to future work. Currently, we have been working on exploiting the Why-Not explanation polynomial to efficiently rewrite a query in order to include the missing answers in its result set. As there are many rewriting possibilities, we plan to select the most promising ones based on a cost function, built with the polynomial. For instance, we may rank higher rewritings with minimum condition changes (i.e., small combinations), minimum side-effects (i.e., small coefficients), etc.

7. REFERENCES