Summarized Causal Explanations For Aggregate Views (Full Version)

ABSTRACT

SQL queries with group-by and average are frequently used and plotted as barcharts in several data analysis applications. Understanding the reasons behind the results in such an aggregate view may be a highly non-trivial and time-consuming task, especially for large datasets with multiple attributes. Hence, generating automated explanations for aggregate views can allow users to gain better insights into the results while saving time in data analysis. When providing explanations for such views, it is paramount to ensure that they are succinct yet comprehensive, reveal different types of insights that hold for different aggregate answers in the view, and, most importantly, they reflect reality and arm users to make informed data-driven decisions, i.e., the explanations do not only consider correlations but are *causal*. In this paper, we present CauSumX, a framework for generating summarized causal explanations for the entire aggregate view. Using background knowledge captured in a causal DAG, CauSumX finds the most effective causal treatments for different groups in the view. We formally define the framework and the optimization problem, study its complexity, and devise an efficient algorithm using the Apriori algorithm, LP rounding, and several optimizations. We experimentally show that our system generates useful summarized causal explanations compared to prior work and scales well for large high-dimensional data.

1 INTRODUCTION

As database interactions grow in popularity and their user base broadens to data analysts and decision-makers with varied backgrounds, it becomes important to generate insightful and automated explanations for results of the queries users run on the data. One simple yet important class of queries used in data analysis is the class of *SQL queries with group-by and average*, which are frequently used and plotted as barcharts in data analysis applications. These queries show how the average varies in different sub-populations in the data by creating an *aggregate view* over the input database (e.g., average salary per country, occupation, race, or gender; average severity of car accidents per major city in the USA, etc.). Understanding the *causal reasons* behind the high/low values of the average in different groups for such queries can enable sound data-driven decision-making to address unwarranted situations. For instance, if a policymaker knows the possible causal reasons behind lower average salary of a certain race or gender in a certain region in the USA, they can try to improve the situation with corrective measures, which may not be possible with insights that are based on non-causal associational factors. Here we give a running example that we will frequently use in the paper.

Example 1.1. *Consider the Stack Overflow annual developer survey [6], where respondents from around the world answer questions about their job. We consider a subset of the data with 38090 tuples from 20 countries and 5 continents appearing the most in the dataset and augmented the data with additional attributes that describe the economy of each country:* `HDI` *(Human Development Index, higher values mean more human development),* `Gini` *(measures income inequity, higher values imply more inequity),* `GDP` *(Gross Domestic Product per capita, a measure of country's economic health, higher is better). Table 1 shows a few sample tuples with a subset of the attributes. The other attributes are* `SexualOrientation`, `EducationParents`, `Dependents`, `Student`, `Hobby`, `HoursComputer`, *and* `Exercise`.

Now consider the following group-by SQL query measuring the average salary in different countries:

```
SELECT Country, AVG(Salary)  FROM Stack-Overflow
GROUP BY Country
```

The results are plotted as a barchart in Figure 1 (the colors will be explained later). There is a huge variation in average salary (converted to USD) in different countries. The user may wonder (i) what are the main factors for this variation across countries, and also (ii) within each country, what is **causing** *developers to earn more or less. However, the dataset is too big, both in terms of the number of tuples and attributes, to look for a succinct yet informative explanation by manual inspection. While tools like Tableau [3] give highly sophisticated visualizations by slicing and dicing the data across several dimensions, they return the aggregates for these dimensions and do not differentiate between causal and non-causal reasons behind Figure 1.*

Understanding the importance of generating insightful explanations for aggregated query results, several approaches have been proposed in database research on *explanations for aggregated query answers*. A simple form is given by the *provenance* for aggregate query answers that show how the output was computed using the input tuples [13]. However, an aggregate answer over a large dataset uses many input tuples, hence several approaches have focused on providing high-level explanations as *predicates* on input tuples that are responsible for producing query answers of interest [46, 64, 84] or provide other types of insights explaining them (e.g., the counterbalance approach in [53]). While these approaches provide predicates as explanations, which are easy to comprehend, they aim to explain certain answers in the view (e.g., outliers [84], high/low values of an answer [46, 53], or comparisons of a set of answers [64]), and do not provide a summarized explanation for the entire view. Further, although the explanations returned by these approaches reveal many interesting insights, they are not causal.

Causal inference, nevertheless, has been studied for several decades in Artificial Intelligence (AI) by *Pearl's Graphical Causal Model* [58], and in Statistics by *Rubin's Potential Outcome Framework* [66]. Causal analysis is a vital tool in determining the effect of a treatment on an outcome, and has been used in decision-making in medicine [62], economics [17], biology [74], and in critical applications like understanding the efficacy of a new vaccine using *randomized controlled trials.* While randomized trials cannot be performed In many applications due to ethical or feasibility issues, fortunately, the

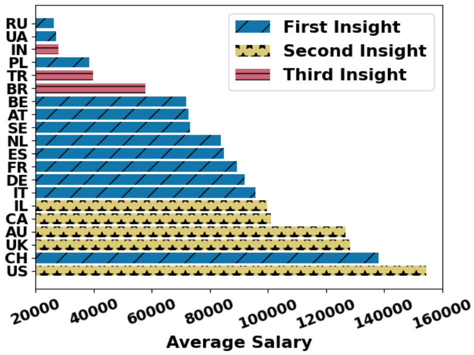


Figure 1: A visualization of the Stack Overflow query results.

above causal models provide ways to do sound causal analysis on *observed* datasets under some assumptions (ref. Section 3).

Recent works have introduced causality to the field of database research [35, 50, 68, 69, 88], allowing users to benefit from this well-founded approach and infer solid causal conclusions from their data and queries. In particular, there has been prior work on extending Pearl's causal model for relational databases [69], providing causal hypothetical reasoning for what-if and how-to queries [36], and providing explanations for aggregate queries using causal analysis that focused on revealing unobserved factors influencing the results [68, 88]; however, [68, 88] provide a single explanation of the entire view, and do not offer fine-grained explanations for individual groups. A recent work [50] introduces a framework that searches for predicates that explain the difference in two average outcomes, and marks the patterns as either causal or not, by proposing a new causal discovery algorithm that extends the PC algorithm [75]. However, they do not search for important treatments affecting the outcomes and do not give causal explanations summarizing the entire view. On the other hand, *summarization* techniques for data and query answers form another active topic in database research [18, 22, 32, 43, 44, 70, 83, 87], often with *diversity* and *coverage* factors, however these summaries are also not causal.

Our contributions. In this work, *we present a novel framework called CauSumX (Causal Summarized EXplanations) to explain the entire aggregate view* from a query with group-by-average. Given a database D, causal background knowledge in the form of a causal DAG by Pearl's graphical causal model [58], a group-by-average query Q, and parameters k and θ, CauSumX generates a set of k *explanation patterns* (predicates) that explain at least θ fraction of groups in $Q(D)$. An explanation pattern contains a *grouping pattern* capturing a subset of output groups covered by the explanation, and a *treatment pattern* with a high or low value of *conditional average treatment effect (CATE)* (ref. Section 3) on the average attribute A_{avg} as the outcome. In standard causal analysis, the goal is to estimate the causal effect of a given treatment on a given outcome, whereas in CauSumX we search for treatments with high and low causal effects for different subsets of groups defined by the grouping pattern. CauSumX combines the features of (i) causal inference, (ii) explanation, and (iii) summarization to provide succinct yet comprehensive and causal explanations for group-by-average queries, helping save time and effort in data analysis.

- For countries in Europe (blue bars with '/'), the most substantial effect on high salaries (effect size of 36K, $p<$ 1e-3) is observed for individuals under 35 with a Master's degree. Conversely, being a student has the greatest adverse impact on annual income (effect size: -39K, $p<$ 1e-3).
- For countries with a medium GDP level (yellow bars with '*'), the most substantial effect on high salaries (effect size of 41K, $p<$ 1e-3) is observed for C-level executives. Conversely, being over 55 with a bachelor's degree has the greatest adverse impact on annual income (effect size: -35K,$p<$ 1e-4).
- For countries with a high Gini coefficient (pink bars with '-'), the most substantial effect on high salaries (effect size of 29K, $p<$ 1e-4) is observed for white individuals under 45. Conversely, being having no formal degree has the greatest adverse impact on annual income (effect size: -28K, $p<$ 1e-3).

Figure 2: Causal explanation summary by CauSumX.

Example 1.2. *Reconsider the dataset and query from Example 1.1. The user runs CauSumX to search for an explanation for her query with no more than three insights while covering all groups, and receives the answers shown in Figure 2. The mapping between countries and insights is visualized using the bars' color and texture in Figure 1. Each country can be mapped to more than one insight, but for simplicity, only one color/texture is visualized. CauSumX uses a* causal DAG *(a partial DAG is shown in Figure 3), explores multiple patterns, and evaluates their causal effect on the salary across different countries. There are three parts in each insight: (a) A* **grouping pattern** *(first underlined text in magenta), illustrates a property or predicate on the groups or* `countries` *(group-by attribute in the query) for which this insight holds. (b) A* **positive treatment pattern** *(second underlined text in blue) is a predicate on the individuals from the above groups with a high positive treatment effect. (c) A* **negative treatment pattern** *(third underlined text in red) is a predicate on the individuals from the above groups with a high negative treatment effect.*

Without having to manually explore this large dataset by running many subsequent queries, the user learns the main reasons for high and low salaries in different countries. These reasons are not just predicates summarizing tuples in the dataset, they have high and low causal effects *as determined by the causal model. Moreover, the user knows that these explanations not only hold for one country but hold for several countries that share the same grouping pattern. The user can continue the exploration by varying parameters in CauSumX.*

Here are our main technical contributions.

(1) We **develop a framework CauSumX** that generates a summarized causal explanation to explain an aggregate view $Q(D)$ for Q with group-by and average. We define explanation patterns that comprise a grouping pattern and a treatment pattern, define an optimization problem to maximize the causal explainability of these explanations subject to a size constraint on the number of explanations and a coverage constraint on the number of output groups covered by them, and show its NP-hardness.

(2) We **design a three-step algorithm** named CauSumX. The first step mines frequent grouping patterns using the seminal Apriori algorithm [11]. The second step uses a greedy lattice-based algorithm for mining promising treatment patterns for each grouping pattern from the previous step. In the third step, we model the optimization problem as an Integer Linear Program (ILP) and solve

Table 1: A subset of the Stack Overflow dataset.

ID	Country	Continent	Gender	Ethnicity	Age	Role	Education	YrsCoding	Undergrad Major	Salary
1	US	North America	Male	White	26	Data Scientist	PhD	10	Computer Science	180k
2	US	North America	Non-binary	White	32	QA developer	Bachelor's degree	5	Mechanical Eng.	83k
3	India	Asia	Male	South Asian	29	C-suite executive	Bachelor's degree	8	Computer Science	24k
4	India	Asia	Female	South Asian	25	Back-end developer	Master's degree	9	Mathematics	7.5k
5	China	Asia	Male	East Asian	21	Back-end developer	Bachelor's degree	7	Computer Science	19k

Table 2: Positioning of CauSumX w.r.t. Query Result Explanation, Interpretable Pred. Models, and Data Summarization.

Related Work		Causal	Entire View	Supports groups
Query Result Explanation	[20, 51–54, 63, 64, 84]	✗	✗	✗
	[46]	✗	✗	✓
	[50]	✓	✗	✓
	[68, 88]	✓	✗	✗
Interpretable Pred. Models	[24, 42]	✗	✓	✗
Data Summarization	[32]	✗	✓	✗
	[70, 83, 87]	✗	✓	✗
CauSumX		✓	✓	✓

it by randomized rounding of its LP relaxation using the grouping and treatment patterns from previous steps.

(3) We provide a thorough **experimental analysis** and **multiple case study** that include five datasets, four baselines, and two variations of our solution as additional comparison points. We show that the explanations generated by CauSumX are of high quality compared to existing approaches and may provide different (and more justifiable) explanations from those given by associational approaches. Additionally, we analyze the runtime, explainability, and coverage of the obtained explanations, and show that CauSumX is both efficient and useful in providing explanations.

2 RELATED WORK

Table 2 summarizes the differences between CauSumX and previous work. Columns in bold highlight our novelty, namely: CauSumX generates a **summarized** and **causal** explanation to the **entire aggregated view** generated by a SQL query while **accounting for variations among the groups**.

Query Result Explanation. A substantial body of research has been dedicated to query result explanations. Multiple works used *data provenance* to obtain explanations for (possibly missing) query results [21, 23, 29, 45, 46, 51, 52, 54, 78]. Other forms of explanations include (non-causal) interventions [31, 63, 64, 77, 84], entropy [33], Shapley values [48, 61], and counterbalancing patterns [53]. Those works are orthogonal to our work, as we aim to explain an entire aggregated view via a small set of causal explanations. Recent works [68, 88] propose using causal inference to explain query results. They identify *confounding attributes* that influence both the outcome and grouping attributes in aggregated SQL queries. However, unlike our approach, they provide *a single explanation* for the entire view, and rather than searching for causes for the outcome, they focused on revealing factors that may have influenced the results. Another work [50] introduced a framework that identifies causal and non-causal patterns to explain the differences between two groups of tuples. Unlike our solution, they do not search for treatments that affect the outcome but rather aim to find patterns that distinguish between groups of interest.

Causal Inference. There is an extensive body of literature on causal inference over observational data in AI and Statistics [38, 58, 66, 79]. We employ standard techniques from this literature to compute causal effects. In related research, estimating heterogeneous treatment effects has been explored [82, 85]. This refers to variations in treatment effects across different population subgroups. However, this research differs from our framework. They assume known treatment and outcome variables and focus on identifying subpopulations with varying treatment effects. In contrast, we assume only the outcome variable is given and aim to identify treatments that influence the outcome for each subpopulation, potentially leading to different treatments for each subgroup.

This work assumes that a *causal DAG* for the input dataset is provided. Though it is well-known that background knowledge is required to determine causal relations [59], a causal DAG can be inferred from the data under some assumptions [27, 37] (e.g., sufficiency, faithfulness). Multiple works have proposed algorithms to discover a causal DAG based on data properties [37, 73, 75]. This line of work is orthogonal to ours, as we expect to receive a causal DAG as part of the input for our framework.

Interpretable Prediction Models. Previous work developed models that offer both high predictive accuracy and interoperability [41, 49, 67, 71]. Rule-based interpretable prediction models [24, 42, 86] often utilize association rule mining processes to produce predictive rules. For instance, [42] introduced IDS, a model for binary classification that aims to optimize both the accuracy and interpretability of the chosen rules. This model generates a short, non-overlapping rule set encompassing the entire feature space and classes. Similarly, [24] devised FRL, an ordered rule list that includes probabilistic if-then rules. We compare against both [24, 42] in Section 6.

Data Summarization. Data summarization is the process of condensing an input dataset into interpretable and representative subsets [40, 83, 90]. A broad spectrum of approaches have been proposed for data and view summarization [18, 22, 43, 44, 70, 83, 87]. Unlike our work, none of these methods specifically aim to uncover causal explanations for aggregated views. In [32], the authors use *explanation tables*, which is one of the baselines in Section 6.

3 BACKGROUND ON CAUSAL INFERENCE

We use Pearl's model for *observational causal analysis* on collected datasets [58] and present the following concepts according to it.

Causal inference, Treatment, ATE, and CATE. The broad goal of *causal inference* is to estimate the effect of a *treatment variable* T on an outcome variable Y (e.g., what is the effect of higher `Education` on `Salary`). The gold standard of causal inference is by doing *randomized controlled experiments*, where the population is randomly divided into a *treated* group that receives the treatment (denoted by $\texttt{do}(T = 1)$ for a binary treatment) and the *control* group ($\texttt{do}(T = 0)$). One popular measure of causal estimate is *Average Treatment Effect* (ATE). In a randomized experiment, ATE is the difference in the average outcomes of the treated and control groups [58, 66]

$$ATE(T, Y) = \mathbb{E}[Y \mid \mathrm{do}(T = 1)] - \mathbb{E}[Y \mid \mathrm{do}(T = 0)] \quad (1)$$

The above definition assumes that the treatment assigned to one unit does not affect the outcome of another unit (called the Stable Unit Treatment Value Assumption (SUTVA) [66][1].

In our work on causal explanations for SQL group-by-average queries, where the treatment with maximum effect may vary among different tuples in the query answer, we are interested in computing the *Conditional Average Treatment Effect* (CATE), which measures the effect of a treatment on an outcome on *a subset of input units* [39, 65]. Given a subset of units defined by (a vector of) attributes B and their values b, we can compute $CATE(T, Y \mid B = b)$ as:

$$\mathbb{E}[Y \mid \mathrm{do}(T = 1), B = b] - \mathbb{E}[Y \mid \mathrm{do}(T = 0), B = b] \quad (2)$$

However, randomized experiments where treatments are assigned at random cannot be done in many practical scenarios due to ethical or feasibility issues (e.g., effect of higher education on salary). In these scenarios, *Observational Causal Analysis* still allows sound causal inference under additional assumptions. Randomization in controlled trials mitigates the effect of *confounding factors* or *covariates*, i.e., attributes that can affect the treatment assignment and outcome. Suppose we want to understand the causal effect of `Education` on `Salary` from the SO dataset. We no longer apply Eq. (1) since the values of `Education` were not assigned at random in this data, and obtaining higher education largely depends on other attributes like `Gender`, `EducationParents`,and `Country`.

$$\textbf{Unconfoundedness}: \quad Y \perp\!\!\!\perp T | Z{=}z \quad (3)$$

$$\textbf{Overlap}: \quad 0 < Pr(T{=}1|Z{=}z) < 1 \quad (4)$$

The unconfoundedness assumption, Eq. (3), states that if we condition on Z, then treatment T in the dataset and the outcome Y are independent. In SO, assuming that only Z ={`Gender`, `EducationParents`, `Country`} affects T = `Education`, if we condition on a fixed set of values of Z, i.e., consider people of a given gender, from a given country, and with a given education level of parents, then T = `Education` and Y = `Salary` are independent. For such confounding factors Z, Eq. (1) and (2) respectively reduce to the following form (Eq. (4) gives feasibility of the expectation difference):

$$ATE(T, Y) = \mathbb{E}_Z [\mathbb{E}[Y \mid T = 1, Z = z] - \mathbb{E}[Y \mid T = 0, Z = z]] \quad (5)$$

$$CATE(T, Y \mid B = b) =$$
$$\mathbb{E}_Z [\mathbb{E}[Y \mid T = 1, B = b, Z = z] - \mathbb{E}[Y \mid T = 0, B = b, Z = z]] \quad (6)$$

The above equations no longer have the $\mathrm{do}(T = b)$, that can be estimated from an observed dataset. Pearl's model gives a systematic way to find such a dataset when a causal DAG is available.

Causal DAG. Pearl's Probabilistic Graphical Causal Model model [58] can be written as a tuple $(\mathcal{E}, \mathcal{N}, Pr_{\mathcal{E}}, \psi)$, where $\mathcal{E}$ is a set of *unobserved exogenous (noise)* variables or attributes, $\Pr_{\mathcal{E}}$ is the joint distribution of $\mathcal{E}$, and $\mathcal{N}$ is a set of *observed endogenous variables*. Here ψ is a set of structural equations that encode dependencies among variables. The equation for $A \in \mathcal{N}$ takes the following form:

$$\psi_A : \mathrm{dom}(Pa_{\mathcal{E}}(A)) \times \mathrm{dom}(Pa_{\mathcal{N}}(A)) \rightarrow \mathrm{dom}(A)$$

Here $Pa_{\mathcal{E}}(A) \subseteq \mathcal{E}$ and $Pa_{\mathcal{N}}(A) \subseteq \mathcal{N} \setminus \{A\}$ respectively denote the exogenous and endogenous parents of A. A causal relational model is associated with a *causal DAG*, G, whose nodes are the endogenous variables $\mathcal{N}$ and whose edges are all pairs (X, Y) (directed edges

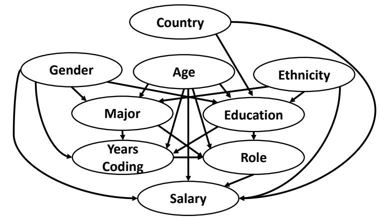


Figure 3: Example causal DAG.

from X to Y) such that $Y \in \mathcal{N}$ and $X \in Pa_{\mathcal{N}}(Y)$. The causal DAG obfuscates exogenous variables as they are unobserved. Any given set of values for the exogenous variables completely determine the values of the endogenous variables by the structural equations (we do not need any known closed-form expressions of the structural equations in this work). The probability distribution $\Pr_{\mathcal{E}}$ on exogenous variables $\mathcal{E}$ induces a probability distribution on the endogenous variables $\mathcal{N}$ by the structural equations ψ.

Figure 3 depicts a causal DAG for the SO dataset over the attributes in Table 1 as endogenous variables (we use a larger causal DAG with all 20 attributes for the SO dataset in our experiments). Given this causal DAG, we can observe that the `Role` that a coder has in their company depends on the values of their `Education`, `Age`, `Major`, and `YearsCoding` attributes.

A causal DAG can be constructed by an analyst or domain expert with background knowledge about the domain as in the above example, or using existing *causal discovery* [58] algorithms, which we study further in our experiments (Section 6).

In Pearl's model, a treatment $T = t$ (on one or more variables) is considered as an *intervention* to a causal DAG by mechanically changing the DAG such that the values of node(s) for T in G are set to the value(s) in t, which is denoted by $\mathrm{do}(T = t)$. Following this operation, the probability distribution of the nodes in the graph changes as the treatment nodes no longer depend on the values of their parents. Pearl's model gives an approach to estimate the new probability distribution by identifying the confounding factors Z described earlier using conditions such as *d-separation* and *backdoor criteria* [58], which we do not discuss in this paper.

4 FRAMEWORK FOR SUMMARIZED CAUSAL EXPLANATIONS FOR AGGREGATE QUERIES

Databases and Queries. We consider a single-relation database[2] over a schema $\mathbb{A}$. The schema is a vector of attribute names, i.e., $\mathbb{A}=(A_1, \ldots, A_s)$, where each A_i is associated with a domain $\mathrm{dom}(A_i)$, which can be categorical or continuous. A database instance D, populates the schema with a set of tuples $t=(a_1, \ldots, a_s)$ where $a_i \in \mathrm{dom}(A_i)$. We use $t[A_i]$ to denote the value of attribute A_i of tuple t. In this paper, we will only consider the *active domain* of every A_i as $\mathrm{dom}(A_i)$, i.e., the set of values of A_i in the given D.

We consider an important class of SQL queries for data analysis, with group-by and average as the aggregate function:

$$Q = \texttt{SELECT } \mathcal{A}_{gb}, \texttt{ AVG } (A_{avg}) \texttt{ FROM D WHERE } \phi \texttt{ GROUP BY } \mathcal{A}_{gb};$$

Here, $\mathcal{A}_{gb} \subseteq \mathbb{A}$ is a set of categorical group-by attributes, $A_{avg} \in \mathbb{A}$ is the average attribute, and ϕ is a predicate. The result of evaluating Q over D is denoted by $Q(D)$. We denote $|Q(D)|$, i.e., the number of groups in $Q(D)$, by m. The `WHERE` condition ϕ simply reduces the

[1]This assumption does not hold for causal inference on multiple tables and even on a single table where tuples depend on each other, which we discuss in Section 7.

[2]We discuss adjustments for supporting multi-dimensional datasets in Section 7.

table D to tuples satisfying ϕ before the techniques in this section can be applied, so we do not discuss ϕ further in this section.

As an example, consider again the query presented in Example 1.1. Here $\mathcal{A}_{gb}$= `Country`, A_{avg}= `Salary`, and ϕ is the empty predicate. The query results $Q(D)$ are shown in Figure 1, where m=$|Q(D)|$=20.

Why average queries? Our framework aims to find summarized causal explanations for the query answers in $Q(D)$ using the concept of CATE discussed in Section 3. The causal estimates by CATE as shown in Eq. (2) and (6), inherently use *expectations, i.e., weighted average*. Thus, CATEs can be used when considering the aggregate average on the outcome column Y, e.g., `AVG(Salary)` in Example 1.1. Aggregate functions like `SUM` or `COUNT` depend on the *number of units* satisfying the conditions in Eq. (2) and (6), which does not have a correspondence with the estimate of causal effect (except that larger groups might reduce variance in the estimate). While other non-causal work on explanations or summarizations for query answers [44, 46, 53, 64, 83, 84] can support other aggregate functions, methods based on causal estimates focus on average, which has been the prevailing assumption in previous work as well [50, 68, 88].

4.1 Explanation Patterns

The *patterns* in our framework of summarized explanations are *conjunctive predicates* on attribute values that are prevalent in previous work on explanation, e.g., [33, 64, 84].

Definition 4.1 (Pattern). *Given a database instance D with schema $\mathbb{A}$, a simple predicate is an expression of the form φ=A_i op a_i, where A_i∈$\mathbb{A}$, a_i∈dom(A_i), and op∈{=, <, >, ≤, ≥}. A* pattern *is a conjunction of simple predicates $\mathcal{P}$=$\varphi_1 \wedge \ldots \wedge \varphi_k$.*

For categorical attributes (e.g., `Role`), we focus solely on patterns where the operator `op` is set to =. For ordinal attributes (e.g., `Age`) we consider patterns where `op` can take any value from {=, <, >, ≤, ≥}.

Grouping and treatment patterns. Our explanations consist of pairs of patterns: (i) A **grouping pattern** $\mathcal{P}_g$ is a pattern that captures a subset of groups in $Q(D)$ and must be well-defined over $Q(D)$, i.e., a query answer $s \in Q(D)$ cannot be assigned to two grouping patterns. Therefore, $\mathcal{P}_g$ can only contain attributes W such that the Functional Dependency (FD) from the grouping attributes, $\mathcal{A}_{gb} \rightarrow W$ holds for all W in $\mathcal{P}_g$. (ii) A **treatment pattern**, $\mathcal{P}_t$, is defined over the dataset D (as opposed to $\mathcal{P}_g$ that is defined over $Q(D)$) and partitions the input tuples into treated ($T = 1$ if $\mathcal{P}_t$ evaluates to true for a tuple) and control groups ($T = 0$ if $\mathcal{P}_t$ evaluates to false). This partition is then used to assess the causal effects of the treatment pattern on the outcome $Y = A_{avg}$, the attribute for average in the query Q.

A pair of grouping and treatment pattern $(\mathcal{P}_g, \mathcal{P}_t)$ together define an **explanation pattern**. Intuitively, $\mathcal{P}_g$ specifies the subpopulation of interest (equivalent to the condition $B = b$ in Eq. (2) for CATE), while $\mathcal{P}_t$ (equivalent to treatment T) explains the observed outcome Y within that subpopulation as per the CATE value.

Definition 4.2 (Explanation Pattern). *Given a database instance D with schema $\mathbb{A}$ and a query Q with group-by attributes $\mathcal{A}_{gb}$ and attribute $A_{avg}$for average, an* explanation pattern *$(\mathcal{P}_g, \mathcal{P}_t)$ where $\mathcal{P}_g$ is a grouping pattern on $Q(D)$, i.e., the FD $\mathcal{A}_{gb} \rightarrow W$ holds in $D$for all attributes W in $\mathcal{P}_g$, and $\mathcal{P}_t$ is a pattern defined over D.*

Example 4.1. *In the first insight in Figure 2, one explanation pattern is $(\mathcal{P}_g, \mathcal{P}_t)$ with $\mathcal{P}_g$:* (`Continent = Europe`) *and $\mathcal{P}_t$:* (`Age < 35`)∧ (`Education = Master's degree`). *Note that the FD from the group-by attribute* `Country → Continent` *holds.*

Partitioning attributes for grouping and treatment patterns. We partition the attributes in $\mathbb{A}$ into two disjoint sets for grouping and treatment patterns. All attributes $W \subseteq \mathbb{A}$ such that the FD $\mathcal{A}_{gb} \rightarrow W$ holds in D are considered for grouping patterns $\mathcal{P}_g$. All other attributes U, such that the FD $\mathcal{A}_{gb} \rightarrow U$ *does not* hold in D are considered for treatment patterns $\mathcal{P}_t$. The necessity for the FD for grouping patterns is explained above, and this partitioning keeps the attributes appearing in grouping and treatment patterns disjoint simplifying the explanations. Further, note that any such attribute W individually where $\mathcal{A}_{gb} \rightarrow W$ holds, and in general, any possible grouping pattern, cannot be a valid treatment pattern. By the overlap condition given in Eq. (4), conditioned on a grouping pattern $\mathcal{P}_g$, there should be at least one unit with $T = 1$ and at least one unit with $T = 0$. If we pick an attribute-value $W = w$ with $\mathcal{A}_{gb} \rightarrow W$ or a pattern with multiple such attributes as the treatment pattern, for all tuples in D contributing to a query answer in $Q(D)$, either it evaluates to true or to false, so we do not get both $T = 1$ and $T = 0$ as required in Eq. (4) to estimate the CATE value.

Explainability. To evaluate the effectiveness of an explanation pattern $(\mathcal{P}_g, \mathcal{P}_t)$, next we define its explainability.

Definition 4.3 (Explainability). *Given a database instance D with schema $\mathbb{A}$, a query Q group-by attributes $\mathcal{A}_{gb}$and attribute A_{avg}, and a causal model $\mathcal{M}_{\mathbb{A}}$ on $\mathbb{A}$associated with a causal DAG, the explainability of an explanation pattern $(\mathcal{P}_g, \mathcal{P}_t)$ is defined as: $Explainability(\mathcal{P}_g, \mathcal{P}_t)$:=$CATE_{\mathcal{M}_{\mathbb{A}}}(\mathcal{P}_t, A_{avg} \mid \mathcal{P}_g)$, using the definition of CATE as in Eq. (2) with treatment T=$\mathcal{P}_t$, outcome Y=A_{avg}, and subpopultion defined by $\mathcal{P}_g$. The subscript $\mathcal{M}_{\mathbb{A}}$ denotes that the CATE is estimated using the causal model $\mathcal{M}_{\mathbb{A}}$ as explained in Section 3.*

In particular, CATE given by Eq. (2) in the above definition is reduced to Eq. (6) using confounding variables Z obtained from the causal DAG of $\mathcal{M}_{\mathbb{A}}$, which then can be estimated from the data D.

As mentioned, our focus is on computing CATE (Eq. 2) rather than ATE (Eq. 1) to understand the causal factors that influence outcomes for specific groups in $Q(D)$. For instance, the effect of having a `Master's degree` on `Salary` in countries in Europe can be different from that in countries in Asia. Therefore, computing the average causal effect while considering all individuals from across the globe may not provide meaningful insights. For $\mathcal{P}_g$: (`Continent = Europe`) and $\mathcal{P}_t$: (`Education = Master's degree`), we define the treatment group as individuals with a master's degree from European countries and the control group as individuals without a Masters degree from European countries. This enables us to focus on a specific group and draw relevant conclusions.

Example 4.2. *In Example 1.2 and Figure 2, there are two explanation patterns using the same grouping pattern. Here $\mathcal{P}_g$:* (`Continent = Europe`), *$\mathcal{P}_{t1}$:* (`Age < 35`) ∧ (`Education = Master's degree`), *and $\mathcal{P}_{t2}$: $(Student = yes)$. The explainability of $(\mathcal{P}_g, \mathcal{P}_{t1}) = 36K$ and that of $(\mathcal{P}_g, \mathcal{P}_{t2}) = -39K$, indicating that age below 35 and having a Master's degree has a high positive causal effect for individuals from European countries while being a student has a high negative effect.*

4.2 Problem Definition and Hardness

Our goal is to obtain a succinct yet comprehensive set of explanation patterns for the groups in $Q(D)$ from the huge search space of possible explanation patterns (e.g., the number of explanation patterns for Example 1.1 was 22350 compared to the short explanation in Figure 2). To achieve this, we frame a constrained optimization problem. We apply three constraints: (1) the number of explanation patterns should not exceed a specified threshold k, (2) the number of groups in $Q(D)$ explained by the patterns must be at least a specified θ-fraction of all the groups in $Q(D)$, and (3) an explanation pattern should not explain the same set of groups as already explained by another explanation pattern. Finally, our goal is to find the set of explanation patterns that abide by these constraints and whose overall explainability is maximized. First, we define the *coverage* of a grouping pattern $\mathcal{P}_g$.

Definition 4.4 (Coverage). *Given a database instance D and a query Q with group-by attributes $\mathcal{A}_{gb}$, a grouping pattern $\mathcal{P}_g$, and a group $s \in Q(D)$, $\mathcal{P}_g$ is said to* **cover** *s if for any tuple $t \in D$ such that $t[\mathcal{A}_{gb}] = s[\mathcal{A}_{gb}]$, it holds that $t \models \mathcal{P}_g$, i.e., t satisfies the predicate $\mathcal{P}_g$. The set of groups in $Q(D)$ covered by $\mathcal{P}_g$ is denoted by $Cov(\mathcal{P}_g)$.*

Next, we define the problem of *Summarized Causal Explanations* that we aim to solve in this work.

Definition 4.5 (Summarized Causal Explanations). *Given a database D, a causal model $\mathcal{M}_{\mathbb{A}}$, a query Q with group-by attributes $\mathcal{A}_{gb}$ and attribute $A_{avg}$for average, a collection of explanation patterns $\{\mathcal{P}_i\}_{i=1}^{l}$, an integer $k \in [1, m]$ where $m = |Q(D)|$, and a threshold $\theta \in [0, 1]$, we aim to find a set $\Phi \subseteq \{\mathcal{P}_i\}_{i=1}^{l}$ of explanation patterns such that the following conditions hold:*

(1) ***(Size constraint)*** *$|\Phi| \leq k$.*
(2) ***(Coverage constraint)*** *at least $\theta \cdot m$ groups from $Q(D)$ are covered by Φ, i.e., $\cup_{\mathcal{P}_g \in \Phi} Cov(\mathcal{P}_g) \geq \theta \times m$.*
(3) ***(Incomparability constraint)*** *There are no pairs of explanation patterns, $(\mathcal{P}_g, \mathcal{P}_t)$ and $(\mathcal{P}'_g, \mathcal{P}'_t)$ in Φ such that $Cov(\mathcal{P}_g) = Cov(\mathcal{P}'_g)$.*

(Objective) *The objective is to maximize the total explainability of Φ under the above constraints, i.e., maximize $\sum_{\mathcal{P} \in \Phi} Explainability(\mathcal{P})$.*

Example 4.3. *In Example 1.2 and Figure 2, the parameters k=3 and θ=1 are used, i.e., we aim to find a set of at most 3 explanation patterns that reveal the causes of the outcome for all groups in $Q(D)$. As Proposition 4.1 shows, even deciding whether there is a set Φ of explanation patterns for a given size constraint k and coverage constraint θ simultaneously is NP-Hard, although for this example, we find a solution covering all 20 countries in $Q(D)$ as shown in Figure 1.*

Hardness Result and enumeration of search space. Since in our optimization problem, we want to cover a certain fraction of answer tuples in $Q(D)$ ('elements') with at most k patterns ('sets'), we can show the following NP-hardness result even if we ignore the optimization objective (proof in the Appendix).

Proposition 4.1. *It is NP-hard to decide whether the Summarized Causal Explanations problem is feasible (i.e., has any solution satisfying the constraints) for a given k and θ.*

Algorithm 1: The CauSumX algorithm

input : A database relation D, an integer k, a query Q with an aggregate attribute A_{avg} and grouping attributes $\mathcal{A}_{gb}$, and a coverage threshold θ
output: A set Φ of explanation patterns.

```
1 Φ ← ∅;
2 Candidates ← GetGroupingPatterns(D, A_avg, 𝒜_gb);   // Section 5.1
3 for 𝒫_g ∈ Candidates do
4     𝒫_g.𝒫_t ← GetTopTreatment(𝒫_g, k, A_avg, D);   // Section 5.2
5 Φ ← SolveLP(Candidates, k, θ);                      // Section 5.3
6 return Φ
```

Further, Definition 4.5 assumes that the search space of explanation patterns is given, while in practice it is not efficient to enumerate all explanation patterns ahead of time. In Section 5 we give efficient algorithms that give good solutions for the optimization problem without explicitly enumerating all explanation patterns upfront using techniques from the *Apriori Algorithm* [11].

Positive and negative explanation patterns. In Figure 2, we return positive and negative patterns for each grouping pattern by slightly varying the optimization objective in Definition 4.5. For a grouping pattern $\mathcal{P}_g$, we find a treatment pattern $\mathcal{P}^+_{g,tp}$ gwith the *highest explainability value of* $(\mathcal{P}_g, \mathcal{P}^+_{g,t})$ ($\mathcal{P}^+_{g,t}$ is called a *positive treatment pattern* for $\mathcal{P}_g$). A *negative treatment pattern* $\mathcal{P}^-_{g,t}$ is defined similarly using the lowest explainability value. In our system, for each grouping pattern $\mathcal{P}_g$ we compute the sum of absolute values of two explainabilities: $|Explainability(\mathcal{P}_g, \mathcal{P}^+_{g,t})| + |Explainability(\mathcal{P}_g, \mathcal{P}^-_{g,t})|$. We treat this sum as the weight of the explanation pattern combination $(\mathcal{P}_g, \mathcal{P}^+_{g,t}, \mathcal{P}^-_{g,t})$ and return top-k explanations that satisfy the constraints in Definition 4.5. This helps understand the cause of both high and low values of outcomes for different groups without explicitly asking for explanations for high and low values as done in previous work [46, 53, 64, 84].

5 THE CAUSUMX ALGORITHM

Proposition 4.1 shows that even deciding the feasibility of the Summarized Causal Explanations problem for a given k and θ is NP-hard. Further, it is not practical to enumerate all possible explanation patterns and compute their explainability upfront from a large search space. Given three-dimensional desiderata in Definition 4.5, i.e., both size and coverage constraints and the objective of maximizing weights or explainability of patterns (unlike the standard set-cover or max-cover problems that have two), it is non-trivial to design a good approximation algorithm or heuristics for this problem. In this section, we present the CauSumX algorithm (for Causal Summarized EXplanations) that aims to address these challenges.

Overview. The pseudo-code in Algorithm 1 outlines the operation of CauSumX. It takes a database D, an integer k, a query Q with an aggregate attribute A_{avg} and grouping attributes $\mathcal{A}_{gb}$, and a coverage threshold θ as input. The output is a set of explanation patterns Φ. The algorithm consists of three steps: (1) extracting grouping patterns using a frequent itemset mining algorithm - the Apriori algorithm [11] (line 2). (2) Focusing on promising treatment patterns for each extracted grouping pattern, *materializing and evaluating only them* (lines 3–4). (3) Utilizing Linear Programming (LP) to obtain a set of explanation patterns (line 5).

5.1 Mining Grouping Patterns

Considering every possible grouping pattern (Definition 4.2) is infeasible as their number is exponential ($O(agrmax_{A_i \in \mathbb{A}} |\mathsf{dom}(A_i)|^{|\mathbb{A}|})$. Instead, our approach utilizes the Apriori algorithm [11] to generate candidate grouping patterns. The Apriori algorithm gets a threshold parameter τ, and ensures that the mined patterns are present in at least τ tuples of D. Formally, given a set of attribute $W \subseteq \mathbb{A}$ such that the FD $\mathcal{A}_{gb} \rightarrow W$ holds in D, we apply the Apriori algorithm to extract frequent itemsets (i.e., patterns) defined solely by these attributes. The algorithm guarantees that each mined pattern covers at least τ tuples from D and is well-defined over $Q(D)$, making them promising candidates for covering the necessary number of groups (see item (2) in Definition 4.5).

Post-Processing. Certain extracted grouping patterns may be superfluous, i.e., the set of groups they define from $Q(D)$ could be indistinguishable. Therefore, following the mining stage, we remove redundant grouping patterns to ensure the obtained solution will satisfy the incomparability constraint (item (3) in Definition 4.5). Additionally, we favor more succinct patterns as they are easier to comprehend. To achieve this, we utilize a hash table that records the groups from $Q(D)$ associated with each mined grouping pattern. In each group set, we retain only the most concise grouping pattern.

For example, the grouping patterns $\mathcal{P}_g^1$={`country = US`}, and $\mathcal{P}_g^2$={`continent = North America, country = US`} define the exact same set of countries. Consequently, we will discard $\mathcal{P}_g^2$.

5.2 Mining Treatment Patterns

As opposed to a standard causal analysis setting where a causal question of the form "What is the effect of treatment T on outcome Y?" is posed, in our setting, **we aim to find T that yields the highest effect on Y.** Our subsequent goal, as discussed at the end of Section 4.2, is to identify a positive treatment pattern $\mathcal{P}_{tg}^+$ and a negative treatment pattern $\mathcal{P}_{tg}^-$ for each mined grouping pattern $\mathcal{P}_g$, which respectively explain the cause for high and low outcomes of tuples in D that satisfy $\mathcal{P}_g$. However, for simplicity and without loss of generality, we present the approach for finding positive explanation patterns only denoted by $\mathcal{P}_t$ (i.e., treatments that have the highest positive CATE for a grouping pattern $\mathcal{P}_g$).

Since the number of potential treatment patterns for a $\mathcal{P}_g$ can be large (exponential in $|\mathbb{A}|$), we propose a greedy approach to materialize and assess the CATE only for promising treatment patterns. This is done by leveraging the notion of lattice traversal [15, 30]. In particular, the set of all treatment patterns can be represented as a lattice where nodes correspond to treatment patterns and there is an edge between $\mathcal{P}_t^1$ and $\mathcal{P}_t^2$ if $\mathcal{P}_t^2$ can be obtained from $\mathcal{P}_t^1$ by adding a single predicate. This lattice can be traversed in a top-down fashion while generating each node at most once. Since not all nodes correspond to treatments that have a positive CATE, we only materialize treatment nodes if all their parents have a positive CATE. Note that since the CATE is non-monotonic (meaning that adding an additional predicate to a treatment pattern can either increase or decrease its CATE value), some relevant treatment patterns may go unnoticed. For example, we observe that for the pattern $\mathcal{P}_{t1}$=(`role = QA`), by adding the predicate (`education = MA`), its CATE value increases, while the CATE decreases by adding the predicate (`education = no degree`). However, according to our

Algorithm 2: Top treatment pattern for a grouping pattern

input : A grouping pattern $\mathcal{P}_g$, the outcome attribute A_{avg}, the dataset D, a causal DAG G, and direction $\sigma \in \{+, -\}$.
output: A treatment pattern $\mathcal{P}_t$.

1 `/* Get all single-predicate patterns. */`
2 $C \leftarrow$ `GenChildren`$(D, A_{avg}, \mathcal{P}_g)$;
3 $C \leftarrow$ `ComputeCATEnFilter`$(C, D, A_{avg}, \mathcal{P}_g, G)$;
4 $\mathcal{P}_t^{max} \leftarrow$ `GetTopTreatment`(C);
5 **while** *True* **do**
6   `/* Get patterns in the next level. */`
7     $C \leftarrow$ `GenChildrenNextLevel`(C, σ);
8     $C \leftarrow$ `ComputeCATEnFilter`$(C, D, A_{avg}, \mathcal{P}_g, G)$;
9     $\mathcal{P}_t \leftarrow$ `GetTopTreatment`(C);
10   **if** $\mathcal{P}_t.CATE > \mathcal{P}_t^{max}.CATE$ **then**
11         $\mathcal{P}_t^{max} \leftarrow \mathcal{P}_t$;
12   **else**
13   Break;
14 **return** $\mathcal{P}_t^{max}$;

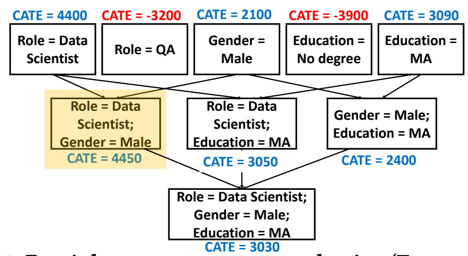


Figure 4: Partial treatment-patterns lattice (Example 5.1).

experiments with real-life data (Section 6.3), combining treatment patterns that exhibit a positive CATE is highly likely to result in a treatment with a positive CATE as well. As a result, we account for most treatments with a positive CATE.

Algorithm 2 describes the search for the treatment pattern $\mathcal{P}_t^{max}$ with the highest positive CATE value for a given grouping pattern $\mathcal{P}_g$. It traverses the pattern lattice in a top-down manner, starting from patterns with a single atomic predicate in line 2. In particular, the function `GenChildren` generates all atomic predicates of the form A_i op a_j. For each treatment pattern, it evaluates its CATE, and discards the ones with CATE values that do not have the same sign σ, either + or − (line 3). The algorithm stores the pattern with the highest positive CATE identified thus far (line 4). It then proceeds to traverse the pattern lattice only for nodes that satisfy the condition of having all their parents with a positive CATE (line 7-8), and updates $\mathcal{P}_t^{max}$ if a pattern with a higher CATE was found (lines 9-12). It terminates at the first level, which does not include the maximum value recorded (lines 13-14).

Example 5.1. *We illustrate the operation of Algorithm 2 using Figure 4. Initially, it considers all treatment patterns with a single predicate (line 2). Assuming σ=+ (searching for the treatment pattern with the highest CATE), it moves to the next level (line 5). At this level, patterns with two predicates are considered only if both of their parents have a positive CATE. Thus, the pattern* {`Role = QA, Gender = Male`} *is excluded as the CATE of* `Role = QA` *is negative (line 8). The treatment pattern with the maximum CATE ($\mathcal{P}_t^{max}$) is found at the second level (marked in yellow) (line 9). After another iteration and generating the third level of the lattice, the algorithm does not proceed to the fourth level because $\mathcal{P}_t^{max}$ is not present in the third level (line 10).*

By changing σ, Algorithm 2 can also find the treatment patterns with the lowest negative CATE. It can also be generalized to output the highest explainability (i.e., the highest absolute CATE value), or both the patterns with the highest and lowest CATE values.

Optimizations. We implement more optimizations for Algorithm 2.

(a) Pruning attributes: We eliminate attributes that do not have a causal relationship with the outcome attribute (lines 2, 7). Since these attributes have no impact on CATE values, they can be disregarded. We can detect such attributes by utilizing the input causal DAG or by removing attributes with low correlation to the outcome.

(b) Pruning treatments: In constructing the lattice (lines 2, 7), we exclude patterns with a near-zero CATE. We observe that combining patterns with a CATE close to zero value often yields a similar result. Consequently, when advancing to the next lattice level, we only consider the top 50% patterns with the highest or lowest CATE.

(c) Parallelism: The process of extracting treatment patterns for each grouping pattern (lines 2, 7) can be performed in parallel since this procedure is dependent only on the grouping pattern.

(d) Estimating CATE Values: We use fixed-size samples to estimate CATE values (lines 3, 8) and investigate the impact of sample size on runtime and accuracy (Section 6.5). Our study shows that using a sample size of 1 million tuples yields CATE estimates that closely match those obtained from the entire dataset.

5.3 Linear Program Formulation and Rounding

Proposition 4.1 shows that finding any feasible solution to the optimization problem in Definition 4.5 is NP-hard. Our study shows that intuitive combinatorial greedy algorithms targeting the size and coverage constraints and maximizing explainability are unable to find good solutions, hence we use an LP-rounding algorithm.

Given a collection of explanation patterns $\{\mathcal{P}_j\}_{j=1}^{l}$ with weights w_j corresponding to their explainability, an integer k, and a threshold θ, we construct the Integer Linear Program (ILP) shown in Figure 5 (which extends the ILP for the *max-k-cover* problem). Here g_j are the variables for patterns $\mathcal{P}_j$. t_i, $i = 1$ to m are the variables for m output groups in $Q(D)$. $s_i \in \mathcal{P}_j$ denotes that s_i satisfies $\mathcal{P}_j$.

We consider the LP-relaxation of this ILP with variables in $[0, 1]$ instead of $\{0, 1\}$, and then use an LP-solver to find solutions. If no solution is returned for this LP, we know that the original ILP did not have any feasible solution either. If any solution is returned by the LP, the original ILP may or may not have a feasible solution. We use the standard randomized rounding algorithm for max-k-cover [60] (sample k patterns with probability $\frac{g_j}{k}$), which guarantees at most k sets, an $(1 - \frac{1}{e})$-approximation to the coverage constraint $\sum_{i=1}^{m} t_i \geq \theta \cdot m$, and a $\frac{1}{k}$-approximation to the maximization objective $\sum_{j=1}^{l} g_j \cdot w_j$ in the LP as well as the ILP since *OPT LP* $\geq$ *OPT ILP* (details in the Appendix). However, the approximation to the objective holds when all explanation patterns are considered, while we use this LP-rounding algorithm in conjunction with the grouping and treatment pattern mining procedures described in Sections 5.1 and 5.2, therefore incur a trade-off between value and efficiency.

We can further improve the above algorithm by considering a fixed range of values for the coverage threshold θ, computing a solution for each value, and selecting the one that maximizes the objective value while satisfying the coverage constraint.

$$\max \sum_{j=1}^{l} g_j \cdot w_j \quad \text{s.t.} \quad (1) \sum_{j=1}^{l} g_j \leq k, \quad (2)\ t_i \leq \sum_{j: T_i \in \mathcal{P}_j} g_j \ \forall i = 1 \text{ to } m,$$
$$(3) \sum_{i=1}^{m} t_i \geq \theta \cdot m, \quad (4)\ t_i, g_j \in \{0, 1\}\ \forall i = 1 \text{ to } m,\ \forall j = 1 \text{ to } l$$

Figure 5: ILP for optimization problem (line 5 in Algorithm 1).

Table 3: Examined datasets.

Dataset	tuples	atts	max values per att	grouping patterns
German [16]	1000	20	53	10
Adult [1]	32.5K	13	94	13
SO [6]	38K	20	20	75
IMPUS-CPS [34]	1.1M	10	67	9
Accidents [55]	2.8M	40	127	15

Time complexity analysis for Algorithm 1. The maximum number of explanation patterns in a database D with attributes $\mathbb{A}$ is bounded by $|D|^{|\mathbb{A}|}$ (considering both grouping and treatment patterns and active domain of attributes), which is polynomial in *data complexity* assuming a fixed schema [81]. The number of patterns dominates the number of variables in the ILP in Figure 5. Hence even if all patterns are enumerated and evaluated explicitly, the LP relaxation of the ILP can be solved (CauSumX uses z3 [28]) and rounded to an integral solution in polynomial time in $|D|$. The additional operations in this section (e.g., computation of CATE in Algorithm 2) are polynomial in D, giving a worst-case polynomial data complexity of CauSumX. However, $|D|^{|\mathbb{A}|}$ can have a large value for large $|D|$ and $\mathbb{A}$. The suite of optimizations developed in this section reduces the running time of CauSumX (ref. Section 6).

6 EXPERIMENTAL EVALUATION

We present experiments that evaluate the effectiveness and efficiency of our proposed framework. We aim to address the following research questions. **Q1**: What is the quality of our explanations, and how does it compare to that of existing methods? **Q2** : How does each phase of CauSumX contributes to its ability to find an explanation summary that satisfies our optimization goal and constraints? **Q3** : What is the efficiency of the CauSumX algorithm? **Q4**: How sensitive are the explanations to various parameters?

Prototype implementation. CauSumX was written in Python, and is publicly available in [14]. CATE values computation was performed using the DoWhy library [72], utilizing their linear regression approach. We use the Apyori package [4] implementation of the Apriori algorithm. CauSumX generates a solution in natural language using predefined templates, as shown in Figure 2. Those templates were generated via prompt questions to ChatGPT [2], asking it to transform predicates into a human-readable text.

6.1 Experimental Setup

Datasets. We examine multiple commonly used datasets:

German: This dataset contains details of bank account holders, including demographic and financial information, along with their credit risk. The causal graph was used from [26].

Adult: This dataset comprises demographic information of individuals along with their education, occupation, annual income, etc. We used the causal graph from [26].

SO: This is discussed in Example 1.1. The causal DAG was constructed using the approach from [89].

IMPUS-CPS: This dataset is derived from the Current Population

Survey conducted by the U.S. Census Bureau, which includes demographic details for individuals, e.g., education, occupation, and annual income. We adopted the causal dag from [26].

Accidents: This dataset provides comprehensive coverage of car accidents across the USA. It includes numerous environmental stimuli features that describe the conditions surrounding the accidents, such as visibility, precipitation, and traffic signals. To construct a causal DAG, we followed the methodology outlined in [89].

Baselines. We compare CauSumX with the following baselines:
Brute-Force: The optimal solution according to Definition 4.5. This algorithm implements an exhaustive search over all possible grouping and treatment pattern combinations.
IDS: The authors of [42] have proposed a framework for generating Interpretable Decision Sets (IDS) for prediction tasks. The framework incorporates parameters restricting the percentage of uncovered data tuples and the number of rules. These parameters were assigned the same values used in our system.
FRL: The authors of [24] introduce a framework for producing Falling Rule Lists (FRL) as a probabilistic classification model. FRLs consist of a sequence of if-then rules, with the if clauses containing antecedents and the then clauses containing probabilities of the desired outcome. The order of rules in a falling rule list reflects the order of the probabilities of the outcome.
Explanation-Table: The authors of [32] introduced an efficient method to generate *explanation tables* for multi-dimensional datasets. The proposed algorithm employs an information-theoretic approach to select patterns that provide the most information gain about the distribution of the outcome attribute.

Since IDS, FRL, and Explanation-Table assume a binary outcome attribute, we binned the outcome variable in each examined scenario using the average outcome values.

We also considered ChatGPT [2] as a baseline. We provided ChatGPT with multiple prompts comprising a task description, SQL query, causal DAG, and dataset link. However, ChatGPT's responses consistently indicated its lack of direct access to external datasets. Consequently, it couldn't provide specific insights or analysis on the datasets, offering only general domain-related insights instead. Although these insights included identifying attributes with strong causal effects on the outcome (e.g., education and role in income), they did not provide specific patterns. Consequently, we excluded this baseline from presentation as it did not contribute to an in-depth analysis of the causes behind query results.

Variations of CauSumX. We consider the following variations:
Brute-Force-LP: As in Brute-Force, all grouping and treatment patterns are examined. In the final step, an approximated solution is obtained by employing the LP formulation described in Section 5.3.
Greedy-Last-Step: This baseline utilizes the approaches described in Section 5 to generate promising grouping and treatment patterns. In the final step, it employs a greedy strategy instead of solving an LP. The strategy involves iteratively selecting explanation patterns based on their explainability and the increase in coverage they offer.

Unless otherwise specified, the size constraint is set to 5, and the coverage threshold is set to 0.75. The threshold of the Apriori algorithm is set to 0.1. IDS, FRL, and Explanation-Table use their default parameters. The time cutoff is set to 3 hours. The experiments were executed on a PC with a 4.8GHz CPU, and 16GB memory.

Results summary. (1) In all examined scenarios, *our generated explanations were consistent with previous findings*. This highlights the capability of CauSumX to generate real-life causal explanations across multiple domains. The IDS, FRL, and Explanation-Table approaches *do not take into account an aggregated SQL query, focusing solely on the input dataset*. As a result, they either identify patterns that are specific to a single group or patterns that are universally applicable across all groups. (2) The approaches that utilize our grouping and treatment patterns algorithms (CauSumX, Greedy-Last-Step) exhibit significantly faster runtimes compared to the algorithms that consider all grouping and treatment patterns (Brute-Force, Brute-Force-LP). The disparity in running times between Greedy-Last-Step and CauSumX is negligible, while CauSumX is capable of achieving a better balance between satisfying the coverage constraint and maximizing the objective. (3) The runtime of CauSumX is minimally affected by the data size, thanks to the sampling optimization. It exhibits a linear growth with the number of attributes, thanks to the pruning optimizations. Additionally, CauSumX experiences a linear increase in runtime with the number of treatment patterns to consider. Furthermore, the number of grouping patterns has a negligible impact on the runtime of CauSumX due to parallelism. Finally, the summary size insignificantly affects the runtime of CauSumX. (4) The Apriori threshold affects both explainability and coverage. We show that a sample size of 1 million tuples is suitable for accurately estimating the CATE values. Finally, the explanations vary based on the used causal DAG, highlighting the significance of correct domain knowledge.

6.2 Quality Evaluation (Q_1)

We assess the quality of the explanation generated by CauSumX relative to the baselines when examining a range of aggregated queries on diverse real-world datasets. Since no ground truth is available, we examine the consistency of our findings with insights from prior research (as was done in [35, 68, 88]). For each dataset, we present the result for a representative aggregated SQL query, whose explanation can be found in the literature. Our queries are inspired by real-life sources, such as the Stack Overflow annual reports [5], media websites (e.g., The 19th Newsletter [7]), and academic papers (e.g., [12, 68]). More details and additional use cases (for the other datasets) are provided in the Appendix.

SO. As in our running example, we consider a SQL query that computes the average salary of developers across the 20 most commonly mentioned countries among respondents (accounting for more than 85% of the entire SO dataset). To define grouping patterns, we considered attributes having FDs with country: Continent, HDI, Gini, and GDP. Here, for brevity, we set the solution size to 3. As mentioned, the explanation summary generated by CauSumX(shown in Figure 2), reveals key insights regarding the factors influencing salary across different countries. It highlights that job role, age, and education level are the primary determinants of income in the examined countries. Notably, individuals in C-level positions tend to earn significantly more than students. These results align with previous research [6, 8], which emphasized the significant influence of education level and job responsibilities on

• For countries in Europe (e.g., Spain, Italy), the most substantial effect on high salaries (effect size of 34K, $p<$ 1e-3) is observed for individuals under 45. Conversely, being under 25 has the greatest adverse impact on salary (effect size of -32K, $p<$ 1e-3).
• For countries with a medium GDP level (e.g., Sweden, Spain), the most substantial effect on high salaries (effect size of 40K, $p<$ 1e-3) is observed for white individuals under 35. Conversely, being over 55 has the greatest adverse impact on salary (effect size of -34K,$p<$ 1e-4).
•For countries with a high Gini coefficient (e.g., Turkey, Brazil), the most substantial effect on high salaries (effect size of 29K, $p<$ 1e-4) is observed for white individuals under 45. Conversely, being being over 55 has the greatest adverse impact on salary (effect size of -28K, $p<$ 1e-3).

Figure 6: SO use-case example (sensitive attributes only).

• For cities in the Northeast region (e.g., Boston, Albany), the combination of overcast weather conditions and low visibility has been found to have a substantial positive effect on severity (effect size of 0.55, p < 1e-3). The presence of traffic signals has been identified as the factor with the largest adverse impact on severity (effect size of -0.42, p < 1e-4).
• For cities in the Midwest region (e.g., Chicago, Detroit), the combination of cold temperatures and snow has been found to have a substantial positive effect on severity (effect size of 0.61, p < 1e-3). Clear weather has been identified as the factor with the largest adverse impact on severity (effect size of -0.31, p < 1e-3).
• For cities in the South region (e.g., Huston, Miami), rain has been found to have a substantial positive effect on severity (effect size of 0.3, p < 1e-3). The presence of traffic calming measures has been identified as the factor with the largest adverse impact on severity (effect size of -0.44, p < 1e-3).
• For cities in the West region (e.g., Phoenix, Los Angeles), the absence of traffic signals and traffic calming measures has been found to have a substantial positive effect on severity (effect size of 0.53, p < 1e-4). City roads (as opposed to highways) has been identified as the factor with the largest adverse impact on severity (effect size of -0.25, p < 1e-4).

Figure 7: Accidents use-case example.

salaries in high-tech. Additionally, our results indicate that being below the age of 35 positively impacts salary, while being over 55 has a negative impact on income. Age discrimination toward people in the IT industry was identified in the literature [10, 47].

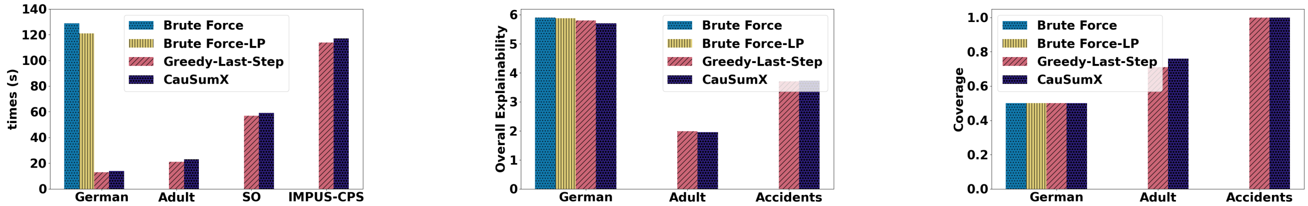
Both IDS and FRL generated rules to predict salary based on available data. Some rules were based on role and education, while others relied on correlations. For example, both baselines suggested that a decrease in computer hours corresponds to higher salaries. However, this variable is strongly associated with job roles, where managerial positions involve fewer computer hours. IDS also identified a rule indicating that individuals with children tend to earn more. However, this association is influenced by age and job roles and does not imply causality. Students typically do not have children, while experienced employees often do. Job role appears to be a more significant factor in explaining income disparity than whether someone has children. This limitation highlights the focus on associations rather than causal relationships in IDS and FRL.

The Explanation-Table baseline aimed to identify patterns associated with high or low salaries. Among the patterns identified, the attribute YearsCoding emerged as a recurring factor. According to this method, individuals who have been coding for over 30 years or have less than 2 years of coding experience tend to have lower salaries. This finding aligns with our observation that being a student or being over 55 leads to reduced salaries. However, Age and Role have stronger direct effects on salary compared to YearsCoding. This approach falls short in capturing the variations among different countries. It either identifies patterns unique to a single country (e.g., low income for individuals from India) or patterns applicable universally (e.g., observations on YearsCoding).

Focusing on Sensitive Attributes. To identify potential biases, we focused exclusively on sensitive attributes (such as ethnicity, gender, and age) when examining treatment patterns. The generated explanation is depicted in Figure 6. Our results indicate that demographic factors significantly influence salary in all countries examined. Specifically, being under 35 positively impacts income, while being over 55 has a negative effect. Similarly, being a white male correlates with higher salary. These findings align with previous research on demographic impact on income, such as the gender wage gap [7] and disparities based on ethnicity [19, 25]. This showcases CauSumX's versatility in identifying causal explanations across different attributes. It also emphasizes its potential in uncovering disparities among demographic groups, supporting efforts to combat discrimination, and promoting equality.

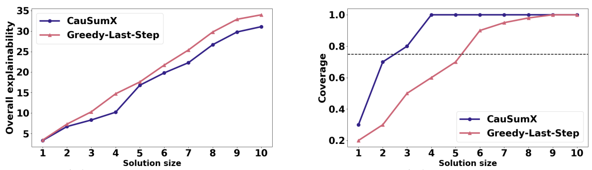
Accidents. We investigated a query that compared the average severity of car accidents across cities in the US. For grouping patterns, we considered the attributes STATE and REGION, which are functionally dependent on CITY. Our generated explanation is shown in Figure 7. Our findings indicate that adverse weather conditions, such as cold temperatures and snow, tend to escalate the severity of car accidents. Conversely, the presence of traffic signals and calming measures appears to mitigate the severity of accidents. Our results highlight variations in weather conditions across different regions. For instance, in the Midwest, cold temperatures and snow commonly contribute to severe accidents, whereas in the South, rainy weather emerges as a more significant factor for severe car accidents as snow and cold temperatures are less common. Previous studies [56, 57] have provided evidence supporting the effectiveness of traffic signals and calming measures in reducing accident severity. This analysis underscores the potential value of our system in extracting insights that go beyond common patterns and correlations. Such insights can be valuable for decision-making processes, enabling a deeper understanding of causal factors and aiding in the implementation of effective road safety measures. The IDS, FRL, and Explanation-Table baselines exceeded our time cutoff.

German. We analyze a query computing the average risk score for loan requests based on purpose. Our findings show that checking and saving accounts' status, along with credit history, greatly influence the risk score across all loan purposes. These results align with previous research [35] and the association of these attributes with the Schufa score [9], a popular credit rating score in Germany. Schufa score considers factors like credit history, existing loans, and negative incidents. However, specific details remain undisclosed. Our system identifies influential factors that affect the outcome, potentially shedding light on opaque algorithms like Schufa score.



(a) Running times (b) Overall explainability (c) Coverage

Figure 8: Performance of different variants of CauSumX. Baselines that exceed the time cutoff are excluded.



(a) Explainability (b) Coverage

Figure 9: Analysis of CauSumX and Greedy-Last-Step.

Rules generated by the baselines rely on correlations rather than causal relationships. For instance, a rule by IDS (and Explanation-Table) suggests that a loan for vacation results in a credit risk score of 1. However, not all vacation loan requests should be approved. Here, the purpose of vacation loans constituted a small number of records, and coincidentally, all of them were associated with high saving account balances. These approaches failed to appropriately prioritize this pattern, resulting in a rule that lacks sensibility in terms of causality but is accurate in terms of prediction.

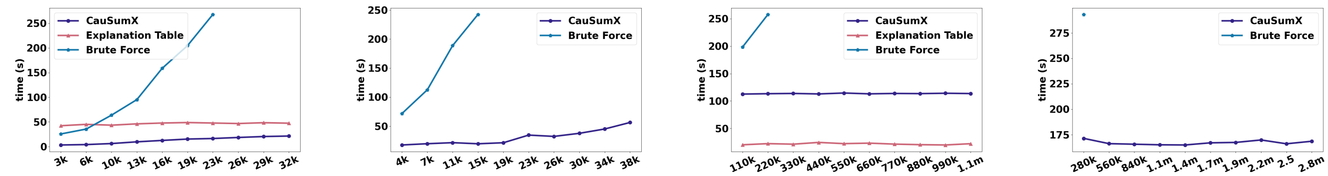
6.3 Ablation Study (Q_2)

We analyze the impact of our grouping and treatment pattern mining algorithms, as well as the LP formulation, compared to the optimal solution determined by Brute-Force based on Definition 4.5.
Runtime. Consider Figure 8(a). As expected, the algorithms that utilize our grouping and treatment patterns algorithms (CauSumX, Greedy-Last-Step) exhibit significantly faster performance compared to Brute-Forcevariants. Both Brute-Force and Brute-Force-LP were unable to process any dataset other than German, as their running times exceeded our time cutoff. *This clearly demonstrates the efficiency gained by our proposed grouping and treatment patterns mining algorithms.* The disparity in running times between Greedy-Last-Step and CauSumX is negligible. This can be attributed to the fact that the treatment pattern detection phase consumes the majority of the execution time due to the larger number of treatment patterns[3]. Additionally, the relatively low number of grouping patterns explains the minimal difference in execution times during the last phase. Despite Greedy-Last-Step being faster than solving the LP, there are not too many explanation patterns to consider.
Coverage & Explainability. Figures 8(b) and (c) display the explainability and coverage of each baseline. In German, all baselines achieve the same coverage determined by k, but the approaches considering all patterns have higher explainability. This minimal difference highlights the effectiveness of our pattern mining and pruning techniques, improving runtime without compromising explainability significantly. In Accidents, CauSumX and Greedy-Last-Step yield identical solutions covering the entire view. However, in Adult, Greedy-Last-Step achieves higher explainability but falls short in coverage compared to CauSumX. This emphasizes the better balance achieved by our LP formulation in satisfying coverage and maximizing the objective compared to the greedy approach.
In-depth Analysis. In a comprehensive analysis comparing CauSumX and Greedy-Last-Step, we explored their performance by varying the solution size k on the SO dataset. The findings are presented in Figure 9. As k increases, both algorithms exhibit improved overall explainability, with similar performance in this aspect. However, their behavior diverges in terms of coverage. CauSumX demonstrates a faster ability to satisfy the coverage constraint (shown by the dashed horizontal line). This is because CauSumX treats coverage as a constraint, while Greedy-Last-Step has no guarantees for coverage. As a result, Greedy-Last-Step only satisfies the coverage constraint for k=6. Based on our findings, CauSumX surpasses Greedy-Last-Step, as both achieve comparable objective values, but CauSumX has a higher likelihood of satisfying constraints.

6.4 Efficiency Evaluation (Q_3)

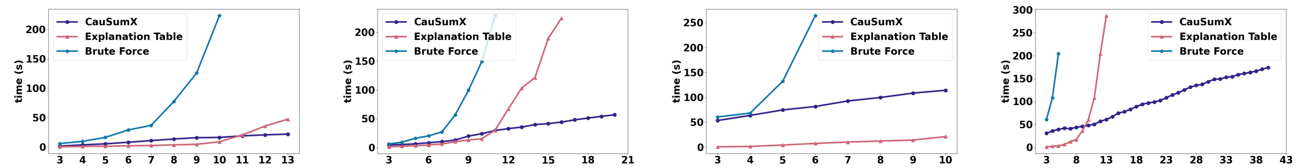
Next, we showcase the scalability of CauSumX. For brevity, we exclude the results for German. The SQL queries used are described in Section 6.2. We analyze the runtime of CauSumX while considering various parameters and compare the results to Brute-Force and Explanation-Table. We omit the results for IDS and FRL from the presentation, as their response times exceed 10 minutes.
Data Size. We analyze the impact of dataset size on runtime through random sampling of tuples. The results are shown in Figure 10. CauSumX and Brute-Force demonstrate a nearly linear increase in runtime for SO and Adult due to their full utilization of data for CATE value computation. However, CauSumX employed a sampling optimization for the larger IMPUS-CPS and Accidents datasets, resulting in a more consistent runtime. Explanation-Table's runtime is unaffected by dataset size due to sampling, but it is unable to handle datasets with more than 10 attributes (e.g., SO).
Number of Attributes. We examine the impact of attribute quantity on runtime, by randomly excluding attributes (except grouping and outcome variables) from consideration. The results are shown in Figure 11. Brute-Force and Explanation-Table show exponential runtime increases with attribute number due to the growing number of patterns to consider. In contrast, CauSumX exhibits linear growth in runtime, thanks to pruning techniques mentioned in Section 5.2 that eliminate non-promising treatment patterns.
Treatment Patterns. We analyze the impact of treatment pattern quantity on runtime. We vary the number of bins for ordinal attributes and randomly exclude values for non-ordinal attributes. The results are displayed in Figure 12. Runtime increases linearly for

[3]A breakdown analysis by step of CauSumX is given in [14]



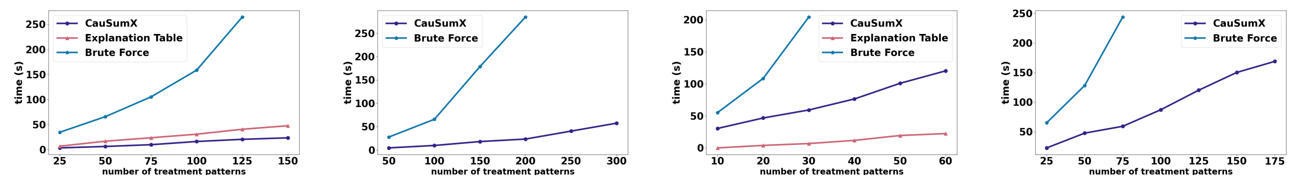
(a) Adult (b) Stack Overflow (c) IMPUS-CPS (d) Accidents

Figure 10: Time vs. dataset size.



(a) Adult (b) Stack Overflow (c) IMPUS-CPS (d) Accidents

Figure 11: Time vs. the number of attributes.



(a) Adult (b) Stack Overflow (c) IMPUS-CPS (d) Accidents

Figure 12: Time vs. the number of treatment patterns.

all algorithms as the number of patterns grows, which is expected due to the increased solution space.

Grouping Patterns. We examine the impact of grouping pattern quantity on runtime. By adjusting the threshold of the Apriori algorithm, we explore different numbers of grouping patterns. For CauSumX, the runtime remains relatively unchanged across all scenarios due to its simultaneous exploration of promising treatment patterns for each grouping pattern. However, Brute-Force's runtime increases linearly with the number of grouping patterns.

Solution Size. Lastly, we explore the impact of varying k. As this parameter affects only the final phase of CauSumX and Brute-Force (which are fast), we observe negligible changes in their runtimes.

6.5 Explanations Sensitivity (Q_4)

We evaluate the impact of various parameters on the quality of the explanation summaries. The measures we focus on are overall explainability and coverage. Full details are provided in the Appendix.

Apriori Threshold. We investigate the effect of varying the threshold parameter τ in the Apriori algorithm. Increasing τ leads to a reduction in the number of grouping patterns considered. Our findings indicate that higher τ values lead to a decrease in both explainability and coverage. Based on our findings, we recommend using a default threshold of 0.1, which provides satisfactory results in terms of runtime, explainability, and coverage. However, it can be adjusted according to specific coverage requirements.

CATE Values Estimation. We investigate the impact of sample size on CATE value estimation. As sample size increases, estimation accuracy improves. We measure estimated CATE values for the top 20 treatments with the highest CATE using different sample sizes. For a sample size of 1m tuples, CATE values have an error of less than 5% and we achieve near accurate treatment ranking based on CATE values. Thus, we conclude that a 1m tuple sample size is suitable for accurate CATE value estimation.

Causal DAG. We depart from the assumption of a pre-defined causal DAG and instead use existing causal discovery algorithms to construct DAGs. We conducted tests with multiple widely used causal discovery algorithms and evaluated the effects on overall explainability and treatment pattern ranking. Our results show that no single causal discovery algorithm consistently outperforms others. Additionally, we observe that the chosen causal DAG can impact CATE values, highlighting CauSumX's sensitivity to input causal DAGs. Thus, our key observation emphasizes the importance of leveraging available algorithms to generate candidate causal DAGs in the absence of a pre-defined DAG. By incorporating domain knowledge, one can determine the most suitable causal DAG.

7 DISCUSSIONS AND FUTURE WORK

This paper presented the CauSumX framework, whose goal is to find a summarized causal explanation for the results of an aggregated SQL query. There are several intriguing future works. First, it will be important to increase robustness of explanations against the causal DAGs, and conduct a formal study of sensitivity of the causal explanations with respect to causal DAGs. Second, our framework currently supports a single-relation database. For multi-relational databases, our definitions need to be augmented since the tuples on which the treatments are applied may not be the same as the tuples where the outcome is observed, and there is a complex dependency among tuples violating basic assumptions of causal analysis. Here the work on *Causal Relational Learning* [69] that extends Pearl's causal model may be useful, although the grouping patterns and treatment patterns across multiple tables have to be defined carefully. Third, even for a single-table database, there may be dependencies among tuples, e.g., in the popular *Flights* dataset for flight delay, the delay of one flight has an impact on subsequent flights using the same aircraft, and is also dependent on flights leaving and arriving in the same airport (i.e., the SUTVA assumption [66] no longer holds). Modeling such complex dependencies in for summarized causal explanations will be interesting future work.

REFERENCES

[1] [n. d.]. Adult Census Income Dataset. https://www.kaggle.com/datasets/uciml/adult-census-income. Accessed: 2023-05-12.

[2] [n. d.]. OpenAI Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-07-15.

[3] [n. d.]. Tableau.

[4] 2019. Apyori Python Packge. https://pypi.org/project/apyori/.

[5] 2021. 2021 Stackoverflow Developer Survey. https://insights.stackoverflow.com/survey/2021.

[6] 2021. Stack Overfow developer survey. https://insights.stackoverflow.com/survey.

[7] 2023. The 19th*. https://19thnews.org/2023/03/parenthood-stereotypes-gender-pay-gap/. Accessed: 2023-05-18.

[8] 2023. Viza jobs. https://vizajobs.com/what-do-technology-jobs-paysalary-insights-and-compensation-factors/.

[9] 2023. What is Schufa Score? https://www.settle-in-berlin.com/what-is-schufa/. Last Updated: 2023-07-15.

[10] 2029. Tech Talks. https://bdtechtalks.com/2019/03/29/ageism-in-tech-age-limit-software-developers-face/.

[11] Rakesh Agrawal, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. Santiago, Chile, 487–499.

[12] Mohamed Aljaban. 2021. Analysis of car accidents causes in the usa. (2021).

[13] Yael Amsterdamer, Daniel Deutch, and Val Tannen. 2011. Provenance for aggregate queries. In *Proceedings of the 30th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2011, June 12-16, 2011, Athens, Greece*, Maurizio Lenzerini and Thomas Schwentick (Eds.). ACM, 153–164. https://doi.org/10.1145/1989284.1989302

[14] Anonymous. 2023. Techinical Report. https://anonymous.4open.science/r/Explanation_Summarization-F736

[15] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. 2019. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 554–565.

[16] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.

[17] Abhijit V Banerjee, Abhijit Banerjee, and Esther Duflo. 2011. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs.

[18] Senjuti Basu Roy, Sihem Amer-Yahia, Ashish Chawla, Gautam Das, and Cong Yu. 2010. Constructing and exploring composite items. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 843–854.

[19] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[20] Aline Bessa, Juliana Freire, Tamraparni Dasu, and Divesh Srivastava. 2020. Effective Discovery of Meaningful Outlier Relationships. *ACM Transactions on Data Science* 1, 2 (2020), 1–33.

[21] Nicole Bidoit, Melanie Herschel, and Katerina Tzompanaki. 2014. Query-based why-not provenance with nedexplain. In *Extending database technology (EDBT)*.

[22] Shaofeng Bu, Laks VS Lakshmanan, and Raymond T Ng. 2005. Mdl summarization with holes. In *Proceedings of the 31st international conference on Very large data bases*. Citeseer, 433–444.

[23] Adriane Chapman and HV Jagadish. 2009. Why not?. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. 523–534.

[24] Chaofan Chen and Cynthia Rudin. 2018. An optimization approach to learning falling rule lists. In *International conference on artificial intelligence and statistics*. PMLR, 604–612.

[25] Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. 2020. Race and economic opportunity in the United States: An intergenerational perspective. *The Quarterly Journal of Economics* 135, 2 (2020), 711–783.

[26] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[27] D.M Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.

[28] Leonardo Mendonça de Moura and Nikolaj Bjørner. 2008. Z3: An Efficient SMT Solver. In *TACAS*. 337–340.

[29] Daniel Deutch, Nave Frost, and Amir Gilad. 2020. Explaining Natural Language query results. *VLDB J.* 29, 1 (2020), 485–508.

[30] Daniel Deutch and Amir Gilad. 2019. Reverse-Engineering Conjunctive Queries from Provenance Examples. In *Advances in Database Technology - 22nd International Conference on Extending Database Technology, EDBT 2019, Lisbon, Portugal, March 26-29, 2019*, Melanie Herschel, Helena Galhardas, Berthold Reinwald, Irini Fundulaki, Carsten Binnig, and Zoi Kaoudi (Eds.). OpenProceedings.org, 277–288. https://doi.org/10.5441/002/edbt.2019.25

[31] Daniel Deutch, Amir Gilad, Tova Milo, Amit Mualem, and Amit Somech. 2022. FEDEX: An Explainability Framework for Data Exploration Steps. *Proc. VLDB Endow.* 15, 13 (2022), 3854–3868. https://www.vldb.org/pvldb/vol15/p3854-gilad.pdf

[32] Kareem El Gebaly, Parag Agrawal, Lukasz Golab, Flip Korn, and Divesh Srivastava. 2014. Interpretable and informative explanations of outcomes. *Proceedings of the VLDB Endowment* 8, 1 (2014), 61–72.

[33] Kareem El Gebaly, Parag Agrawal, Lukasz Golab, Flip Korn, and Divesh Srivastava. 2014. Interpretable and Informative Explanations of Outcomes. *Proc. VLDB Endow.* 8, 1 (sep 2014), 61–72. https://doi.org/10.14778/2735461.2735467

[34] Sarah Flood, Miriam King, Steven Ruggles, and J Robert Warren. 2015. Integrated public use microdata series, current population survey: Version 9.0.[Machine-readable database]. *Minneapolis: University of Minnesota* 1 (2015).

[35] Sainyam Galhotra, Amir Gilad, Sudeepa Roy, and Babak Salimi. 2022. HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 1598–1611. https://doi.org/10.1145/3514221.3526149

[36] Sainyam Galhotra, Amir Gilad, Sudeepa Roy, and Babak Salimi. 2022. Hyper: Hypothetical reasoning with what-if and how-to queries using a probabilistic causal approach. In *Proceedings of the 2022 International Conference on Management of Data*. 1598–1611.

[37] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[38] Sander Greenland and James M Robins. 1999. Epidemiology, justice, and the probability of causation. *Jurimetrics* 40 (1999), 321.

[39] Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association* 81, 396 (1986), 945–960.

[40] Alexandra Kim, Laks VS Lakshmanan, and Divesh Srivastava. 2020. Summarizing hierarchical multidimensional data. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 877–888.

[41] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems* 27 (2014).

[42] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[43] Laks VS Lakshmanan, Raymond T Ng, Christine Xing Wang, Xiaodong Zhou, and Theodore J Johnson. 2002. The generalized MDL approach for summarization. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 766–777.

[44] Laks VS Lakshmanan, Jian Pei, and Jiawei Han. 2002. Quotient cube: How to summarize the semantics of a data cube. In *VLDB'02: Proceedings of the 28th International Conference on Very Large Databases*. Elsevier, 778–789.

[45] Seokki Lee, Bertram Ludäscher, and Boris Glavic. 2020. Approximate summaries for why and why-not provenance (extended version). *arXiv preprint arXiv:2002.00084* (2020).

[46] Chenjie Li, Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2021. Putting Things into Context: Rich Explanations for Query Answers using Join Graphs. In *Proceedings of the 2021 International Conference on Management of Data*. 1051–1063.

[47] Yin Lin, Brit Youngmann, Yuval Moskovitch, HV Jagadish, and Tova Milo. 2021. On detecting cherry-picked generalizations. *Proceedings of the VLDB Endowment* 15, 1 (2021), 59–71.

[48] Ester Livshits, Leopoldo E. Bertossi, Benny Kimelfeld, and Moshe Sebag. 2020. The Shapley Value of Tuples in Query Answering. In *ICDT*, Vol. 155. 20:1–20:19.

[49] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 623–631.

[50] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. XInsight: EXplainable Data Analysis Through The Lens of Causality. *Proc. ACM Manag. Data*, Article 156 (jun 2023), 27 pages.

[51] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2009. Why so? or why no? functional causality for explaining query answers. *arXiv preprint arXiv:0912.5340* (2009).

[52] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The complexity of causality and responsibility for query answers and non-answers. *arXiv preprint arXiv:1009.2021* (2010).

[53] Zhengjie Miao, Qitian Zeng, Boris Glavic, and Sudeepa Roy. 2019. Going beyond provenance: Explaining query answers with pattern-based counterbalances. In *Proceedings of the 2019 International Conference on Management of Data*. 485–502.

[54] Tova Milo, Yuval Moskovitch, and Brit Youngmann. 2020. Contribution Maximization in Probabilistic Datalog. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 817–828.

[55] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. 2019. A countrywide traffic accident dataset. *arXiv preprint arXiv:1906.05409* (2019).

[56] Göran Nilsson. 1982. Effects of speed limits on traffic accidents in Sweden.

[57] Ndhlovu Pardon and Chigwenya Average. 2013. The effectiveness of traffic calming measures in reducing road carnage in masvingo urban. *International Journal* 3, 2 (2013), 2305–1493.

[58] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).
[59] J. Pearl and D. Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
[60] Prabhakar Raghavan and Clark D Tompson. 1987. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica* 7, 4 (1987), 365–374.
[61] Alon Reshef, Benny Kimelfeld, and Ester Livshits. 2020. The Impact of Negation on the Complexity of the Shapley Value in Conjunctive Queries. In *PODS*, Dan Suciu, Yufei Tao, and Zhewei Wei (Eds.). 285–297.
[62] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology.
[63] Sudeepa Roy, Laurel Orr, and Dan Suciu. 2015. Explaining query answers with explanation-ready databases. *Proceedings of the VLDB Endowment* 9, 4 (2015), 348–359.
[64] Sudeepa Roy and Dan Suciu. 2014. A formal approach to finding explanations for database queries. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 1579–1590.
[65] Donald Bruce Rubin. 1971. *The use of matched sampling and regression adjustment in observational studies*. Ph. D. Dissertation. Harvard University.
[66] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[67] Omer Sagi and Lior Rokach. 2021. Approximating XGBoost with an interpretable decision tree. *Information Sciences* 572 (2021), 522–542.
[68] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*. 1021–1035.
[69] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. 2020. Causal Relational Learning. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). ACM, 241–256. https://doi.org/10.1145/3318464.3389759
[70] Gayatri Sathe and Sunita Sarawagi. 2001. Intelligent rollups in multidimensional OLAP data. In *VLDB*. 307–316.
[71] Holger Schielzeth. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* 1, 2 (2010), 103–113.
[72] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).
[73] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 10 (2006).
[74] Bill Shipley. 2016. *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge university press.
[75] P. Spirtes et al. 2000. *Causation, prediction, and search*. MIT press.
[76] David Steurer. 2014. Max Coverage—Randomized LP Rounding. http://www.cs.cornell.edu/courses/cs4820/2014sp/notes/maxcoverage.pdf.
[77] Yuchao Tao, Amir Gilad, Ashwin Machanavajjhala, and Sudeepa Roy. 2022. DPXPlain: Privately Explaining Aggregate Query Answers. *Proc. VLDB Endow.* 16, 1 (2022), 113–126. https://www.vldb.org/pvldb/vol16/p113-tao.pdf
[78] Balder ten Cate, Cristina Civili, Evgeny Sherkhonov, and Wang-Chiew Tan. 2015. High-level why-not explanations using ontologies. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 31–43.
[79] Jin Tian and Judea Pearl. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence* 28, 1-4 (2000), 287–313.
[80] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.
[81] Moshe Y. Vardi. 1982. The Complexity of Relational Query Languages (Extended Abstract). In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing* (San Francisco, California, USA) *(STOC '82)*. ACM, New York, NY, USA, 137–146. https://doi.org/10.1145/800070.802186
[82] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[83] Yuhao Wen, Xiaodan Zhu, Sudeepa Roy, and Jun Yang. 2018. Interactive summarization and exploration of top aggregate query answers. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 2196.
[84] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining away outliers in aggregate queries. (2013).
[85] Yu Xie, Jennie E Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological methodology* 42, 1 (2012), 314–347.
[86] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017. Scalable Bayesian rule lists. In *International conference on machine learning*. PMLR, 3921–3930.
[87] Brit Youngmann, Sihem Amer-Yahia, and Aurélien Personnaz. 2022. Guided Exploration of Data Summaries. *Proc. VLDB Endow.* 15, 9 (2022).
[88] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2023. On Explaining Confounding Bias. *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (2023).
[89] Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. 2023. Causal Data Integration. *arXiv preprint arXiv:2305.08741* (2023).
[90] Cong Yu, Laks Lakshmanan, and Sihem Amer-Yahia. 2009. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th international conference on extending database technology: Advances in database technology*. 368–378.

In this part, we provide missing proofs and additional experiments.

A PROOFS

id	A_1	A_2	A_3	O
t_1	1	0	0	0
t_2	1	0	0	0
t_3	1	1	0	0
t_4	0	0	1	0
t_5	0	1	1	0
t_{S_1}	1	-35	7	0
t_{S_2}	12	1	-4	0
t_{S_3}	55	97	1	0

Figure 13: Example query view for the reduction from set cover in the proof of Proposition 4.1. The sets are $S_1 = \{1, 2, 3\}$, $S_2 = \{3, 5\}$, and $S_3 = \{4, 5\}$. If $k = 2$, the patterns that cover $\frac{5+2}{5+3} = \frac{7}{8}$ of the tuples are $A_1 = 1$ and $A_3 = 1$, indicating that S_1, S_3 is a cover.

Proof of Proposition ??. In the decision version of the Set Cover problem we are given with a universe of elements $U = \{x_1, \ldots, x_{n'}\}$, a collection of m subsets $S_1, \ldots, S_{m'} \subseteq U$ and a number k. The question is whether there exists a cover of U of at most k' subsets.

Given an instance of the set cover problem, we build an instance of the ExPO problem as follows. We build a relation R with $m' + 1$ attributes, $\mathbb{A} = (A_1, \ldots, A_{m'}, O)$, and containing $n' + m'$ tuples. For each element $x_i \in U$, we create a tuple t_i, such that $t_i[A_j] = 1$ iff $x_i \in S_j$. We further add m' tuples t_{S_j} such that $t_{S_j}[A_j] = 1$, $t_{S_j}[O] = 0$, and $t_{S_j}[A_p] = l \neq 0$ for all $p \neq j$ where l is a unique number not used anywhere else in an attribute of R. The query Q is the query that returns all the tuples in the relation R, i.e., Q = SELECT $A_1 \ldots A_{m'}$, COUNT(*) FROM R GROUP BY $A_1, \ldots, A_{m'}$. Here, $\mathcal{P}_g$ can be any predicate, since the FD that needs to hold is $id \rightarrow \mathcal{P}_g$. Note that for each set of tuples defined by a pattern can only have an outcome of 0, as the outcome of all tuples is 0. Therefore, the explainability of all explanation patterns (Definition 4.3) is 0. For ExPO, we further define $\theta = \frac{n'+k'}{n'+m'}$, $k = k'$, and $\tau = 0$. The underlying causal DAG, G, only contains the edges of the form $A_j \rightarrow O$ for all $1 \leq j \leq m'$. We claim that there exists a cover of U with at most k sets iff there exists a solution Φ to ExPO such that $|\Phi| \leq k$, all tuples are covered, and $\sum_{\varphi \in \Phi} explainability(\varphi) \geq 0$.

($\Rightarrow$) Assume that we have a collection $S_{j_1}, \ldots, S_{j_k}$ such that $\cup_{j=j_1}^{j_k} S_j = U$. We show that there is a solution for ExPO as follows. For each S_{j_1}, we choose for the solution the pattern $\mathcal{P}_g^{j_i} : A_{j_i} = 1$. We show that $\Phi = \{(\mathcal{P}_g^{j_1}, \emptyset), \ldots, (\mathcal{P}_g^{j_k}, \emptyset)\}$ is a solution to ExPO. First, we note that all tuples of the form t_i are covered by at least one explanation pattern by their definition. For the m remaining tuples, we have coverage of at most k tuples. These are the tuples $t_{S_{j_i}}$ that have $A_{j_i} = 1$. Thus, the number of covered tuples is exactly $n' + k'$ out of $n' + m'$ tuples in R. If there are fewer than k tuples we can augment the original cover with arbitrary sets to obtain a cover of size k.

($\Leftarrow$) Assume that we have a solution to ExPO with the aforementioned parameters. We show that we can find a solution to the set cover problem. Suppose the cover is $\Phi = \{(\mathcal{P}_g^{j_1}, \emptyset), \ldots, (\mathcal{P}_g^{j_k}, \emptyset)\}$. We first claim that no grouping pattern that includes $A_i = 0$ in a conjunction can be included in Φ as such a pattern will not cover any tuple t_{S_j} since these tuples do not have an attribute with value 0 by definition (and any other number other than 1 will only cover a single tuple). Thus, the number of covered tuples will be $< \frac{n'+k'}{n'+m'} = \theta$, which would contradict the assumption that this is a valid solution to ExPO. Hence, all patterns are of conjunctions of $A_i = 1$. For each treatment pattern of the form $\mathcal{P}_g = \wedge_{j=i_a}^{i_b} (A_j = 1) \wedge (A_p = l)$, we choose an arbitrary attribute in the conjunction A_j if $(A_j = 1) \in \mathcal{P}_g$ and choose S_j for the cover. Finally, if there is an uncovered element x in U and Φ includes a pattern in of the form $\mathcal{P}_g = (A_j = l)$ where $l \neq 1$, we choose for the cover a set S that covers x arbitrarily. We claim that the chosen collection of k sets is a cover of U. To see this, recall that we claimed that the coverage of Φ is at least $n + k$. If the coverage includes tuples of the form t_{S_j}, then each pattern covers a single tuple. Suppose these patterns are $\mathcal{P}_a, \ldots, \mathcal{P}_b$. When building the coverage, instead of these patterns, we add a set that covers elements that are not yet covered by existing patterns. Thus, there are at least $b - a$ covered elements from U in addition to the $n + k - (b - a)$ tuples covered by the patterns. Thus, the set cover we have assembled contains $n - (b - a) + (b - a) = n$ elements and covers all elements in U. ◻

Randomized rounding algorithm. The solution of the ILP formulation (Section 5.3) is determined by the values of the variables g_i, indicating the selected explanation patterns. We compute a solution by using any LP solver, then apply the following randomized rounding procedure [60]:
(1) If no solution is returned (LP is infeasible), return *"no solution"*.
(2) Let $g_1, \ldots, g_l$ and $t_1, \ldots, t_m$ be a solution to the LP.
(3) Interpret the numbers for $g_1/k, \ldots, g_l/k$ as probabilities for the explanation patterns $\mathcal{P}_j$, $j = 1$ to l.
(4) Choose k explanation patterns independently at random according to these probabilities.
(5) Return the collection Φ with the k chosen explanation patterns.

We can show that if all grouping and treatment patterns are considered, this procedure yields a solution that covers at least $(1 - \frac{1}{e}) \times \theta m$ groups from $Q(D)$ in expectation, and its overall explainability, in expectation, is at least a $(\frac{1}{k})$ fraction of the corresponding optimal solution.

Proposition A.1. *The following holds for the LP-rounding algorithm of the ILP in Figure 5:*

(1) If the LP-rounding Algorithm returns "no solution", then it is correct, i.e., no solutions exist for the ILP.
(2) Otherwise, the algorithm returns a collection Φ of k explanation patterns that covers at least $(1 - \frac{1}{e}) \times \theta \cdot m$ groups in expectation, and has an expected overall explainability of $\geq \frac{1}{k} \times OPT\ ILP$.

The proof adapts the proof from [76] for the randomized rounding algorithm for the Maximum Coverage problem to our setting.

Proof of A.1. (1) A feasible solution of ILP($\mathcal{I}$) is also a feasible solution of LP($\mathcal{I}$). Hence if there are no fractional solution to LP($\mathcal{I}$), there are no integral solutions as well.

(2) In this case the LP($\mathcal{I}$) returns some solution.

(a) **Claim: for every group T_i, the probability that G covers T_i is at least $(1-\frac{1}{e}).t_i$.** If we choose a random pattern according to the probabilities $g_1/k, \cdots, g_l/k$, it covers group T_i with probability $\sum_{j: T_i \in \mathcal{P}_j} g_j/k \geq t_i/k$ by (??). Therefore, the probability that none of the k patterns chosen by the rounding algorithm in case (2) covers T_i is at most $(1 - t_i/k)^k$ and the probability that T_i is covered is $\geq 1 - (1 - t_i/k)^k$ which is $\geq (1 - \frac{1}{e}).t_i$, since the left is concave and the right is linear, and the inequality holds at the end points $t_i = 0, 1$ in the interval $[0, 1]$.

(b) **Claim: at least $(1 - \frac{1}{e}) \times \theta m$ groups are covered in expectation.** Let M_i be a random variable such that $M_i = 1$ if T_i is covered by some pattern in G and $= 0$ otherwise. The number of groups covered by G is $\sum_{i=1}^{m} M_i$. Expected number of groups covered by G is $E[\sum_{i=1}^{m} M_i] = \sum_i E[M_i]$ (by linearity of expectation) $= \sum_i Pr[M_i = 1] \geq \sum_i (1 - \frac{1}{e}).t_i$ (by claim (a)) $= (1 - \frac{1}{e}) \sum_{i=1}^{m} t_i \geq (1 - \frac{1}{e}).\theta m$ by (??) in the LP.

(c) **Claim: The total weight of patterns in G is $\geq \frac{1}{k} \times OPT\ LP(\mathcal{I})$.** Let C_j be a random variable such that $C_j = 1$ if $\mathcal{P}_j \in G$ and $= 0$ otherwise. The total weight of patterns in G is $\sum_{j=1}^{l} C_j.w_j$. Expected weights

• To buy a new car, having a caching account with at least 200 DM and paying back all credits at this bank duly has the most significant positive effect on the risk score (effect size of 0.56, $p<$ 1e-3). Conversely, requesting a loan with a duration exceeding 48 months has the largest adverse impact on credit risk (effect size of -0.49, $p<$ 1e-5).
• To buy domestic appliances, requesting a loan with a duration not exceeding 12 months and paying back all credits at this bank duly has the most significant positive effect on the risk score (effect size of 0.34,$p<$ 1e-3). Conversely, requesting a loan with a duration exceeding 48 months has the largest adverse impact on credit risk (effect size of -0.69, $p<$ 1e-5).
• To buy furniture or equipment, having a caching account with at least 200 DM has the most significant positive effect on the risk score (effect size of 0.3, $p<$ 1e-5). Conversely, requesting a loan with a duration exceeding 45 months has the largest adverse impact on credit risk (effect size of -0.78, $p<$ 1e-3).
• To get a loan for repairs, have a caching account with at least 200 DM and a saving account with at least 1000 DM has the most significant positive effect on the risk score (effect size of 0.5, $p<$ 1e-3). Conversely, not having a checking account and renting a house has the largest adverse impact on credit risk (effect size of -0.66, $p<$ 1e-4).
• To get a loan for retraining, having a owning a house has the most significant positive effect on the risk score (effect size of 0.4, $p<$ 1e-2). Conversely, requesting a loan with a duration exceeding 60 months has the largest adverse impact on credit risk (effect size of -0.66, $p<$ 1e-3).

Figure 14: German use-case example.

• For blue-collar occupations (e.g., Machine-op-inspct, Craft-repair, Transport-moving), being an adult who is married has the most significant positive effect on the death rate (0.25,$p<$ 1e-3). Conversely, being unmarried has the largest adverse impact on income (effect size of -0.2, $p<$ 1e-4).
• For white-collar occupations (e.g., Exec-managerial, Prof-specialty, Adm-clerical), being a male with a bachelor's degree or higher has the most significant positive effect on income (effect size of 0.38, $p<$ 1e-4). Conversely, being unmarried has the largest adverse impact on income (effect size of -0.23, $p<$ 1e-3).
• For service occupations (e.g., Sales, Other-service), being married has the most significant positive effect on income (effect size of 0.53, $p<$ 1e-3). Conversely, being unmarried female has the largest adverse impact on income (effect size of -0.39, $p<$ 1e-4).

Figure 15: Adult use-case example.

of patterns in G is $E[\sum_{j=1}^{l} C_j] = \sum_j E[C_j]$ (by linearity of expectation) = $\sum_j Pr[C_j = 1] = \sum_j g_j/k = \frac{1}{k}.OPT\ LP(\mathcal{I})$.

(d) **Claim: The total weight of groups in G is $\geq \frac{1}{k} \times OPT\ ILP(\mathcal{I})$.** Since an optimal solution of ILP($\mathcal{I}$) is a feasible solution of LP($\mathcal{I}$), we have

$$OPT\ ILP(\mathcal{I}) \leq OPT\ LP(\mathcal{I}) \tag{7}$$

Combining (7) with claim (c), (d) follows. (a) and (d) prove the proposition. □

B ADDITIONAL EXPERIMENTS

Here we provide missing details for the experiments as well as additional experiments.

German.: We analyze a query computing the average risk score for loan requests based on purpose. Due to the absence of functional dependencies in the dataset, each group in the aggregated view necessitates a distinct explanation, representing the individual groups in the result view. The explanation summary generated by our CauSumX is presented in Figure **??**. Among the ten purposes examined, four purposes could not be explained, as none of the considered treatments were found to be statistically significant. This outcome can be attributed to the low number of tuples associated with those particular purposes in the dataset.

Our findings reveal the significant impact of checking and saving accounts' status and credit history on the risk score for all loan purposes. These results align with previous research [35] and the association of these attributes with the Schufa score [9], a widely used credit rating score in Germany. The Schufa score considers factors like credit history, existing loans, and negative incidents. However, specific details remain undisclosed. Our system identifies influential attributes that affect desired outcomes, potentially shedding light on opaque algorithms like Schufa score. Rules generated by IDS, FRL, and Explanation-Table rely on correlations rather than causal relationships. For instance, a rule generated by IDS (and Explanation-Table) suggests that a loan for a vacation corresponds to a credit risk score of 1. However, not all vacation loan requests should be approved. Here, the purpose of vacation loans constituted a small number of records, and coincidentally, all of them were associated with high saving account balances. These approaches failed to appropriately prioritize this pattern, resulting in a rule that lacks sensibility in terms of causality but is accurate in terms of prediction.

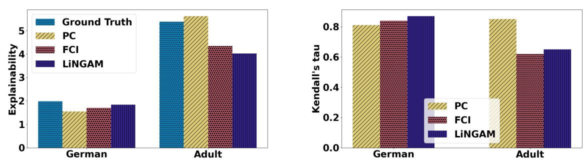
Adult.: We analyzed average income across occupations using an aggregated query. The income variable is a binary attribute, where 1 represents an annual income greater than \$50k and 0 indicates otherwise. To establish the grouping patterns, we utilized the attribute OCCUPATION CATEGORY, which exhibited a functional dependency with the OCCUPATION variable. All other variables were utilized to define the treatment patterns. The generated explanation summary can be found in Figure **??**. Marital status, education, and gender were identified as significant factors affecting income across all occupations, aligning with prior research [68, 80]. Salimi et al.[68] found a high representation of married males and a strong positive association between marriage and high income in this dataset. Although there is no direct causal link between marital status and income, dataset inconsistencies resulted in marital status having the strongest impact in our analysis due to adjusted gross income based on filing status, which reflects household income. Additionally, [68] demonstrated that males tend to have higher education levels and higher education is associated with higher incomes. Our results support these findings, revealing variations across occupation categories, particularly in white-collar occupations where higher education predominantly influences income. This highlights the benefit of our approach in providing detailed causal explanations. IDS, FRL, and Explanation-Table yielded comparable outcomes, with marital status being the best predictor of income, followed by gender, and age. However, they fail to explain variations among occupations.

Breakdown Analysis of the CauSumXAlgorithm. We analyze the operation of CauSumX by step. The runtime analysis is depicted in Figure 16. We observe that in all cases, mining the treatment pattern phase (Algorithm 2) consumes most of the time. Since the number of grouping patterns is relatively small (as there are not too many FDs in the examined cases), the first and last steps are relatively fast. This aligns with our time complexity analysis and demonstrates the need for an efficient approach for avoiding iterating over all possible treatment patterns.

Apriori Threshold. We investigate the effect of varying the threshold parameter τ in the Apriori algorithm. Increasing τ leads to a reduction in the number of grouping patterns considered. Recall that Brute Force examines all possible grouping patterns, equivalent to setting τ=0. However, even when τ=0, CauSumX and Brute Force yield different explainability scores. This is because CauSumX does not explore all treatment patterns and therefore is not guaranteed to find the optimal ones. Our findings for

Dataset	Graph Name	Number of Edges	Density
German	Used causal DAG	20	0.05
	PC	43	0.11
	FCI	12	0.03
	LiNGAM	8	0.02
Adult	Used causal DAG	36	0.23
	PC	38	0.24
	FCI	10	0.06
	LiNGAM	18	0.11
SO	Used causal DAG	28	0.07
	PC	75	0.19
	FCI	41	0.1
	LiNGAM	7	0.01

Table 4: Causal DAG statistics.



(a) Explainability **(b) Kendall's τ**

Figure 19: Modifying the causal DAG.

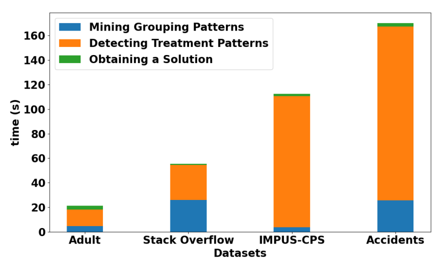
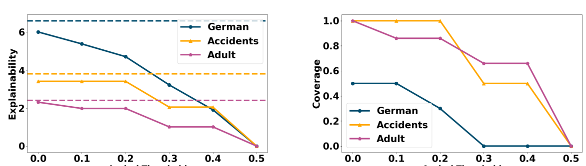
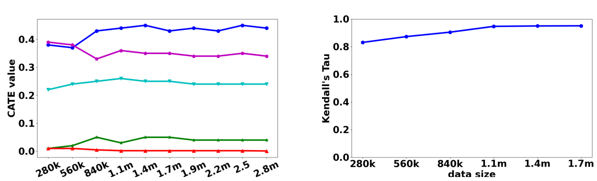


Figure 16: Runtime by-step of the CauSumX algorithm



(a) Explainability vs. Threshold **(b) Coverage vs. Threshold**

Figure 17: The Effect of Apriori Threshold.



(a) CATE value vs. sample size **(b) Kendall's τ vs. sample size**

Figure 18: CATE Values Estimation (Accidents dataset).

the German, Adult, and Accident datasets are shown in Figure 17 (similar trends were observed for the other datasets). Figure 17(a) presents the impact of the Apriori threshold on explainability. The dashed lines represent the explainability achieved by the Brute Force Algorithm. In Figure 17(b), we illustrate the effect of the threshold on coverage. Note that even when setting τ=0, it is still not possible to cover all groups in the German dataset. This limitation arises because each group requires a separate explanation (due to the absence of FDs in this dataset), and the fact that the coverage is further restricted by the size constraint (which is 5 in our setting). As expected, higher threshold values lead to a decrease in both explainability and coverage. Based on our findings, we recommend using a default threshold of 0.1, which provides satisfactory results in terms of runtime, explainability, and coverage. However, the analyst can adjust this threshold according to specific coverage requirements.

CATE Values Estimation. We conducted an investigation to assess how the sample size affects the estimation of CATE values. As the sample size increases, the running time also increases, but the accuracy of the estimations improves. Figure 18 illustrates the results for the Accident dataset. In Figure 18(a), we present the estimated CATE values for 5 random treatments using various sample sizes. In Figure 18(b), we evaluate the agreement between rankings using Kendall's τ correlation coefficient. We randomly selected 20 treatments and ranked them based on their CATE values, comparing this ranking with rankings obtained using different sample sizes. Notably, for a sample size of $1m$ tuples, the CATE values exhibit an error of no more than 5%, and the Kendall's τ reaches a high and stable value of 0.95. Similar trends were observed for the IMPUS-CPS dataset. Consequently, we conclude that a sample size of $1m$ tuples is suitable for accurate estimation of the CATE values.

Causal DAG. In this experiment, we deviate from the assumption of having a pre-defined causal DAG. Instead, we leverage existing causal discovery algorithms to construct the DAGs and examine their impact on the results. Modifying the DAG can lead to changes in the CATE values, consequently affecting their ranking based on CATE values. Hence, we present the effects on both overall explainability and the ranking of treatment patterns when using different causal DAGs. We conducted tests using three widely used causal discovery algorithms: PC [75], FCI [75], and LiNGAM [73]. Statistics on the obtained causal DAGs are given in Table 4. Our results for the German and Adult and SO datasets are shown in Figure 19 (similar patterns were observed for the other datasets). Figure 19(a) illustrates the obtained explainability scores for the German and Adult datasets (for SO, the explainability scores are on a different range and thus omitted from the presentation) using different causal DAGs. In Figure 19, we present the Kendall tau values, comparing the ranking of top-20 treatments based on their CATE values with the ranking obtained using the ground truth causal DAG. Notably, no single causal discovery algorithm outperforms all others. While for the German dataset, LiNGAM yielded results that closely aligned with our ground truth, for the Adult dataset, the PC algorithm demonstrates superiority. This is because each algorithm relies on specific assumptions to construct a causal DAG, leading to varying performance across datasets depending on the validity of those assumptions. We observe that the generated causal DAGs tend to be sparser than the ground truth DAGs. Hence, our key finding is that when a causal DAG is unavailable, one can utilize available algorithms to generate candidate causal DAGs, and then, by leveraging domain knowledge, determine the most suitable causal DAG that aligns with the domain expertise.