Causal Explanations for Disparate Trends: Where and Why?

TAL BLAU, Ben-Gurion University of the Negev, Israel
BRIT YOUNGMANN, Technion, Israel
ANNA FARIHA, University of Utah, USA
YUVAL MOSKOVITCH, Ben-Gurion University of the Negev, Israel

During data analysis, we are often perplexed by certain *disparities* observed between two groups of interest within a dataset. To better understand an observed disparity, we need *explanations* that can pinpoint the data regions where the disparity is most pronounced, along with its causes, i.e., factors that alleviate or exacerbate the disparity. This task is complex and tedious, particularly for large and high-dimensional datasets, demanding an automatic system for discovering *explanations* (data regions and causes) of an observed disparity. It is critical that explanations for disparities are not only interpretable but also actionable—enabling users to make informed, data-driven decisions. This requires explanations to go beyond surface-level correlations and instead capture *causal* relationships. We introduce ExDis, a framework for discovering causal Explanations for Disparities between two groups of interest. ExDis identifies data regions (subpopulations) where disparities are most pronounced (or reversed), and associates specific factors that causally contribute to the disparity within each identified data region. We formally define the ExDis framework and the associated optimization problem, analyze its complexity, and develop an efficient algorithm to solve the problem. Through extensive experiments over three real-world datasets, we demonstrate that ExDis generates meaningful causal explanations, outperforms prior methods, and scales effectively to handle large, high-dimensional datasets.

CCS Concepts: • **Computing methodologies → Causal reasoning and diagnostics**; • **Information systems → Data analytics**; *Data mining*; • **Mathematics of computing → Exploratory data analysis**.

Additional Key Words and Phrases: Explanation, Disparity.

ACM Reference Format:
Tal Blau, Brit Youngmann, Anna Fariha, and Yuval Moskovitch. 2026. Causal Explanations for Disparate Trends: Where and Why?. *Proc. ACM Manag. Data* 4, 1 (SIGMOD), Article XXX (February 2026), 26 pages. https://doi.org/10.1145/XXXXXXX

1 Introduction

Data is the main building block of modern, data-driven decision-making. People rely on the trends observed in the data to gain insights, and in turn, use those insights to draw conclusions or even make important decisions. For large, high-dimensional datasets, certain observed trends often require further drilling down or *explanations*. For example, after observing a *disparate* trend that females are more likely to experience nervous breakdowns and anxiety attacks than non-females, one might wonder: "For which subpopulation is this disparity more pronounced?", "Which factors contribute to this disparity?", "Are there particular countries/races where this trend is reversed?", and so on. Manually searching for these answers is like finding a needle in a haystack, which demands automated ways to pinpoint the subpopulation or data region where a trend is amplified

Authors' Contact Information: Tal Blau, Ben-Gurion University of the Negev, Israel, tbl@post.bgu.ac.il; Brit Youngmann, Technion, Israel, brity@technion.ac.il; Anna Fariha, University of Utah, Salt Lake City, UT, USA, afariha@cs.utah.edu; Yuval Moskovitch, Ben-Gurion University of the Negev, Israel, yuvalmos@bgu.ac.il.

ACM 2836-6573/2026/2-ARTXXX
https://doi.org/10.1145/XXXXXXX

Table 1. A sample dataset over a partial schema of the Stack Overflow Annual Developer Survey dataset [1].

Gender	**Ethnicity**	**Education**	**Role**	**YrsProfCoding**	**TC**
Non-binary	White	BS	Business analyst	6-8	83K
Male	South Asian	PhD	Data analyst	4-6	124K
Female	South Asian	MS	Back-end developer	2-4	75K
Male	East Asian	BS	Back-end developer	6-8	59K

or substantially reversed from the global trend, and identify factors that are connected to these disparate trends.

Understanding *causal reasons* behind disparities in outcomes between two groups is essential for making informed, data-driven decisions to address inequities. For instance, if a policymaker identifies the factors that causally lead to lower average salaries in a specific subpopulation compared to the rest of the population, they can design targeted interventions to mitigate the gap.

In this paper, we propose ExDis for automatically Explaining an observed Disparate trend. We proceed to provide three examples, highlighting unique use cases, to motivate the need to discover *explanations* (data regions and associated causes) of an observed disparate trend between two groups of interest within a dataset.

Example 1.1 (Investigating a disparate trend). *An analyst Miro is examining tech workers' total-compensation data (Table 1), which contains information about individuals' demographics (gender, ethnicity, age, etc.), role, professional coding experience, their own and parents' education, total compensation (TC), etc. Contrary to the common knowledge, Miro observes that the average TC of data or business analysts ($106K) is about 10% higher than the average TC of back-end developers ($96K)—a surprising trend! Miro wants to identify large subpopulations that contribute significantly to this trend, and uncover causes behind it. He discovers that one of such subpopulation is White individuals aged between 25–34, which constitutes 35% of the population, within which, on average, analysts ($115K) earn 9.5% more than developers ($105K). Miro further discovers that having worked as a professional coder is a major contributing factor to this disparity, particularly for this subpopulation. Specifically, among White individuals aged between 25–34, having 6–8 years of professional coding experience causes a TC increase of* $\underrightarrow{\$44K}$ *for* `analysts`*, and only* $\underrightarrow{\$10K}$ *for* `back-end developers`*, further exacerbating the TC-gap.*

Miro wonders if there are other significant data regions with similar properties. What are the major causes of the disparity in those regions? Unfortunately, the dataset contains a number of multi-valued attributes, making manual exploration of all possible data regions infeasible. Furthermore, exploring all possible factors that can cause a significant disparity in the target attribute (TC) requires a more involved search, making it impossible for Miro to do it manually.

Example 1.2 (Debugging bias). *While analyzing a health insurance coverage dataset [71], Soha observes a disparate trend that individuals with an occupation that involves manual labor have a 13% lower chance (78%) of being covered by health insurance than the overall population across all occupations (91%). This indicates a "blue-collar bias" [25] and Soha wishes to discover data regions where this bias is significant and uncover why. Turns out that among non-natives, which makes up 16% of the population, the disparity is even more severe. Within this subpopulation, manual-labor workers have a 65% chance of being insured, which is 19% lower than the 84% coverage rate for all occupations. Furthermore, among not-natives, earning between 25K–55K boosts the chance of having health insurance for* `manual-labor workers` *by* $\underrightarrow{2\%}$ *, where it hurts the chance for* `people with any occupation` *by* $\underleftarrow{1\%}$ *.*

Example 1.3 (Discovering reverse trends). *Generally, males have a lower likelihood (37%) of feeling nervous frequently than non-males (45%). Madison is investigating a Medical Expenditure Panel Survey dataset [2] and they want to find subpopulations where a reverse trend exists, i.e., males have a higher likelihood of feeling nervous than non-males (this phenomenon is known as Simpson's Paradox [75]). One such subpopulation is divorced people aged between 51–63 who have a doctor's recommendation to exercise, where males have a higher likelihood (47%) of feeling nervous than non-males (43%). Interestingly, within this subpopulation, not currently smoking exacerbates the situation for* `males` *(increases the likelihood of feeling nervous by* $\underrightarrow{21\%}$ *) but improves the situation for* `non-males` *(decreases the likelihood of feeling nervous by* $\underleftarrow{14\%}$ *). However, discovering such a reverse trend, they must manually examine all possible subpopulations and try out all possible treatments within each subpopulation!*

The above examples motivate investigating disparities in an aggregated *outcome variable* (e.g., TC) between *two (possibly overlapping) groups* of interest (e.g., analysts vs back-end developers), aiming to identify (1) *where* the disparities are most pronounced (or reversed), such as a specific data region or subpopulation and (2) *why*, i.e., what factors further alleviate/exacerbate the disparity. An example of a real-world use case is the famous UC Berkeley gender bias case [8], where aggregate admissions data suggested bias against women, but deeper analysis revealed Simpson's paradox driven by department-level differences. Another real-world example is the kidney stone treatment study [29], where an inferior treatment appeared to be more effective overall, even though the superior treatment performs better for both small and large kidney stones. Our goal is to detect and explain such disparities.

1.1 Desiderata

There are three key desiderata for the aforementioned problem. **First,** a single causal explanation rarely accounts for the disparity observed across different subpopulations, which may exhibit disparity for different underlying reasons. Thus, the first goal is to identify *high-utility* subpopulations—groups where a strong and meaningful causal explanation for the disparity exists. **Second,** while small subpopulations may exhibit strong causal explanations, insights derived from such groups are often not generalizable. To ensure broader applicability and avoid misleading insights, chosen subpopulations must have sufficient *support*, i.e., they should cover a reasonable portion of the overall dataset. **Third,** selecting the top-k subpopulations purely based on utility and support may lead to redundancy. E.g., "principal engineers" and "people aged between 35–45" may comprise the same individuals, as most principal engineers are 35–45 years old. Thus, beyond finding high-utility and high-support subpopulations, we must minimize the overlap among the reported k subpopulations, ensuring *diversity* [47, 77, 85].

1.2 Problem

Given a database D, an outcome variable O, causal background knowledge in the form of a causal DAG $\mathcal{G}_D$ by Pearl's graphical causal model [50], two groups of interest g_1 and g_2, and a parameter k, our goal is to generate a set of k *disparity explanations* that, collectively, best explain the disparities between g_1 and g_2 w.r.t an aggregation over O, according to $\mathcal{G}_D$. In this work, we consider `AVERAGE` as the aggregation function since it satisfies the requirements for causal analysis (details are in Section 4.1). Each explanation consists of two components: (1) a *subpopulation* where the disparity is pronounced (or reversed), and (2) a *treatment pattern* that causally affects g_1 and g_2 disparately, within that subpopulation. The quality of a set of disparity explanations is primarily determined by the causal strength of the treatment patterns they reveal, measured by the *Average Treatment Effect (ATE)* [50]. This forms our main optimization objective: to identify a set of subpopulations for which the associated treatments strongly explain the observed disparity. In addition to maximizing causal explainability, two important constraints guide the selection process: (1) Each explanation must have sufficient *support*—i.e., the subpopulation it describes should cover a sizable fraction of the data, ensuring representativeness and generalizability. (2) The selected subpopulations must exhibit low *overlap*, encouraging *diversity* in the explanation set.

1.3 Challenges and Limitations of Prior Work

The key challenge lies in identifying subpopulations that are associated with strong causal explanations for observed disparities, while simultaneously satisfying constraints on support and diversity. Prior work mostly focused on selecting the top-k subpopulations based on observed disparities or coverage [3, 49], often neglecting causal explainability or redundancy. Moreover, unlike approaches that rely solely on observed outcome disparities [68], we emphasize the importance of discovering subpopulations where the disparity is causally explained by specific treatments. We extend prior work in two key directions: (1) We address a more difficult variant of the problem by optimizing over a set of k subpopulations under support and diversity constraints, and (2) We focus on identifying high-quality *causal* explanations, not only data regions with high disparities.

Recent works [41, 82] have used causal inference to explain aggregate query results and disparities between two groups within it. However, they do not support overlapping groups, which is essential for bias debugging as demonstrated in Example 1.2. While our form of causal explanation (the same treatment resulting in different outcomes for different groups) is already established in prior works, such as XInsight [41], they focus only on a global explanation. In contrast, we expand the granularity of explanation to subpopulation level, since trends in different subpopulations can diverge significantly from global trends. As such, we aim to identify causal explanations within different subpopulations rather than providing a single explanation for the entire data or query outcome. Finally, operating on the entire data, rather than the aggregate view, poses another challenge as the search space becomes significantly larger (we explain this in Section 5.2).

1.4 Contributions

We present a novel framework named ExDis to explain the disparity between the average outcome variable between two groups of interest. We make the following contributions:

(1) We **formalize the problem** of generating a set of causal explanations to account for disparities in outcomes between two (possibly overlapping) groups of interest. Specifically, we define an optimization problem that seeks to *maximize the causal utility* of the selected explanations—i.e., their ability to causally explain the observed disparity, subject to three constraints: (1) a bound on the number of explanations (size k), (2) minimum support for each explanation to ensure representativeness, and (3) limited overlap between subpopulations to reduce redundancy. We show that this problem is NP-hard (Section 4).

(2) We **develop the ExDis framework**, which operates in three steps. First, ExDis finds candidate subpopulations. Then it identifies local explanations for each candidate subpopulation by adapting a previous work on finding treatments with substantial causal effect [82]. Finally, it uses an effective greedy strategy to find a k-sized explanation set (Section 5).

(3) We present a thorough **empirical analysis** over 3 real-world datasets and present **3 case studies** that include 5 baselines, and 4 variants. We show that ExDis generates higher quality explanations than the existing approaches and can find alternative explanations that existing approaches miss. We also find ExDis scalable and efficient in practice, with its runtime being linear w.r.t the number of data tuples (Section 6).

We review related work in Section 2, provide background on causal inference in Section 3, and discuss our limitations and directions for future work in Section 7.

2 Related work

2.1 Identifying Interesting Subgroups in High-dimensional Data

Prior work identifies the most intriguing data subsets for exploration [5, 13, 23, 28, 38, 46, 49, 60–63, 81]. Other studies focus on uncovering compelling data visualizations [73, 87], or pinpointing data subsets where models underperform [16, 49, 68]. One of our key objectives is to identify subpopulations with a significant disparity in the average outcomes between two groups of interest.

DivExplorer [49] is designed to analyze the behavior of classification models, with the primary goal of identifying data regions where a performance metric (e.g., false positive rates) deviates significantly compared to the entire dataset. FairDebugger [68] aims to identify data subsets responsible for fairness violations in the outcomes of a random forest classifier. It pinpoints the most impactful data samples that significantly influence the model's predictions. However, unlike ExDis, both of these systems focus solely on detecting data regions with unexpected behavior, without uncovering the underlying *causal* factors driving the observed trends. We empirically compare ExDis against these baselines in Section 6.

2.2 Query Results Explanation

Extensive research has been devoted to explain the results of aggregate SQL queries. Multiple works leverage *data provenance* to generate explanations for query results [9, 14, 17, 35, 37, 42, 43, 70], while some rely on non-causal interventions [12, 18, 19, 54, 55, 69, 79], entropy-based techniques [20], and counterbalancing patterns [44]. This line of work differs from ours, as our goal is to explain the disparity among two, possibly overlapping, groups of interest via a small set of causal explanations.

Recent works [41, 59, 82, 83] use causal inference to explain aggregate query results. Prior work [59, 83] proposed methods to find confounding variables that explain the correlation between the grouping attribute and the outcome in group-by-average queries. Both provide the same explanation for all groups in the query results. CauSumX's [82] goal is to provide explanations for an entire aggregate view by combining similar explanations for brevity. To adapt it to our setting, it could be used to explain the aggregate results of two groups. However, our objective differs: rather than identifying treatments that influence each group *individually*, we focus on finding what *differentiates* the groups. Specifically, we search for treatments that benefit one group but have the opposite effect on the other. In contrast, CauSumX simply aims to identify what influences the outcome within each group. Consequently, the treatments identified by CauSumX cannot serve as explanations for the observed disparity between the two groups. Furthermore, CauSumX operates only on aggregate views with a relatively small number of grouping patterns, making it unlikely to scale in our setting.

XInsight [41] identifies both causal and non-causal patterns to explain disparities between two groups in aggregate queries. In contrast, we support overlapping groups and identify specific causal explanations within different subpopulations, rather than seeking a single treatment for the entire dataset. As our experiments confirm (Section 6.2), in many cases, no universal explanation suffices, and disparities are better understood through localized causal insights.

2.3 Rule Mining

Association rule mining is a widely studied problem [26, 33], which aims to identify frequent relationships in datasets. Rule-based interpretable models utilize these techniques to derive predictive rules, aiming to balance accuracy and interpretability [32, 40, 58, 64]. Recent works have explored rule generation from causal relationships [51, 52, 67] and heterogeneous treatment effect estimation [76, 80]. However, they do not address our problem of identifying where disparities are significant and their causes.

CURLS [88] aims at finding a set of causal rules that delineate subgroups with a significant treatment effect and small outcome variances. Like how we limit #explanations and #predicates within each explanation, CURLS limits #rules and their size to ensure interpretability. While both works aim to minimize the overlap between explanations (rules in CURLS), our goal is to expose and explain disparities between groups, while CURLS focuses on identifying subpopulations with significant effects. Notably, subpopulations with strong treatment effects do not necessarily show high disparities across groups. In fact, subpopulations with weaker overall effects can reveal stronger disparities, as opposing effects between the groups within a subpopulation may cancel each other out.

3 Background on Causal Inference

We use Pearl's model for *observational causal analysis* [50] and present below a few concepts according to it. The broad goal of *causal inference* is to estimate the effect of a *treatment variable* T on an *outcome variable* O (e.g., the effect of `YrsProfCoding` on `TC`).

The Average Treatment Effect (ATE) quantifying the difference in expected outcomes between treated and untreated groups [50, 57]. ATE conceptually assumes a scenario where treatment is assigned randomly. However, in observational data, treatment is not assigned randomly, and *confounding variables* that influence both treatment and outcome must be accounted for. To estimate ATE for a binary treatment T on an outcome O, the following definition is used:

$$ATE(T, O) = \mathbb{E}_Z\left[\mathbb{E}[O \mid T = 1, \mathbf{Z}] - \mathbb{E}[O \mid T = 0, \mathbf{Z}]\right]$$

Here, $\mathbf{Z}$ represents the set of confounding variables, ensuring that the causal effect is isolated from confounding influences.

Example 3.1. *Suppose that we want to estimate the causal effect of* `YrsProfCoding` *on* `TC` *from the Stack Overflow (SO) dataset. Since the values of* `YrsProfCoding` *were not assigned at random, and having more or fewer years of professional coding experience and obtaining a high total compensation may depend on other attributes like* `Ethnicity`*,* `Education`*, and* `Role`*, we must control for these confounding variables when estimating* ATE(`YrsProfCoding`, `TC`).

Pearl's model provides ways to account for these confounders $\mathbf{Z}$ to get an unbiased causal estimate under additional assumptions: (1) The unconfoundedness assumption states that if we condition on $\mathbf{Z}$, then T is independent of O, given $\mathbf{Z}$. Intuitively, it means that after conditioning on $\mathbf{Z}$, T is as good as randomly assigned. (2) The Overlap assumption ensures that for every combination of confounders, there is a nonzero probability of receiving the treatment, allowing for valid comparisons across groups.

Pearl's model gives a systematic way (e.g., the backdoor criterion [50]) to find a sufficient set of confounding variables $\mathbf{Z}$ when a *causal DAG* is available. A causal DAG is a specific type of Bayesian network, where nodes represent random variables (i.e., data attributes) and edges signify potential direct causal influence, that provides a simple way to represent causal relationships among variables. Causal DAGs can be constructed by a domain expert or using *causal discovery* algorithms [24]. In line with prior work [21, 36, 82], we assume the causal DAG is given as part of the input.

Example 3.2. *Figure 2 depicts a causal DAG for the SO dataset over a subset of attributes in Table 1, which indicates that* `YrsProfCoding` *depends on an individual's* `Role`*,* `Ethnicity`*, and* `Education`.

For this work, we focus on estimating the causal effect of a treatment T on an outcome O within a specific subpopulation, characterized by a *pattern* ψ. (We define the set of patterns in Section 4.) Consequently, our goal is to compute the Conditional Average Treatment Effect (CATE) [27, 56] rather than the ATE, as CATE captures the treatment effect within a targeted subpopulation. To

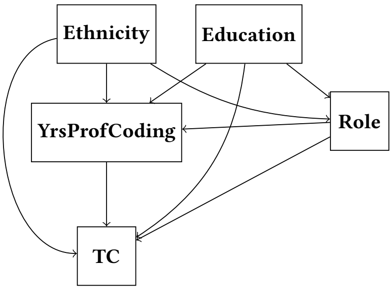

Fig. 2. Partial causal DAG for the Stack Overflow dataset.

estimate the CATE for a binary treatment T on an outcome O for a subpopulation ψ, we use the following definition:

$$CATE(T, O|\psi) = \mathbb{E}_Z\left[\mathbb{E}[O \mid T = 1, \mathbf{Z}, \psi] - \mathbb{E}[O \mid T = 0, \mathbf{Z}, \psi]\right]$$

where $\mathbf{Z}$ represents a sufficient set of confounding variables.

4 Discovering Disparity Explanations

We consider a single-relation database instance D over a schema $\mathbb{A}$=$(A_1, \ldots, A_m)$, where each attribute A_i is associated with a domain $\mathtt{dom}(A_i)$. The outcome attribute O∈$\mathbb{A}$ can be either categorical or continuous. We also assume all other attributes in $\mathbb{A}$ are categorical (if not, we can discretize them to make them categorical). A tuple in D is denoted as t=$(a_1, \ldots, a_m)$ where a_i∈$\mathtt{dom}(A_i)$. We use bold letters to represent a subset of attributes $\mathbf{A}$⊆$\mathbb{A}$. To specify a *subpopulation* (subset of tuples) from D, we use *patterns* [20, 37, 39, 55, 79, 82] that comprise conjunctive *predicates* on attribute values.

Definition 4.1 (Pattern). *Given a database instance D over schema $\mathbb{A}$, a simple predicate φ is an expression of the form A_i op a_i, where A_i∈$\mathbb{A}$, a_i∈$\mathtt{dom}(A_i)$, and op ∈ $\{=, \neq\}$. A* pattern *ψ is a conjunction of simple predicates, i.e., $\psi = \varphi_1 \wedge \ldots \wedge \varphi_k$. We use $\psi(D) \subseteq D$ to denote the subpopulation within D defined by ψ.*

Example 4.1. *Examples of two simple predicates in the SO dataset (Table 1) are* `Ethnicity = Asian` *and* `Role = Data analyst`. *An example of a pattern is* `Ethnicity = Asian ∧ Role = Data analyst`.

The two groups of interest, g_1 and g_2, are defined by the patterns ψ_{g_1} and ψ_{g_2} respectively. In this work, we only consider equality or inequality predicates, in line with previous work on explanations that deem such predicates intuitive and understandable [3, 20, 54].

4.1 Disparity Explanations

For a database D, we aim to discover explanations for an observed disparity in average outcome AVG(O) between g_1 and g_2. Our building blocks are *disparity explanations* that identify *where* the average outcomes for g_1 and g_2 differ significantly and *why*.

Example 4.2. *An analyst over the SO dataset (Table 1) is interested in finding explanations of a surprising disparate observation: the average* TC *of data or business analysts (\$106k) is \$10k higher than the average* TC *of back-end developers (\$96k).*

We focus on comparing the average outcomes of two groups. Our framework is designed to uncover *causal explanations* for differences in these averages, leveraging the concept of CATE as discussed in Section 3. Since CATE inherently relies on *expectations* (weighted averages), it is particularly suited for analyzing aggregate averages of outcomes. Aggregate functions such as SUM or COUNT, on the other hand, depend on the *number of tuples* in the data, which does not directly align with causal effect estimates. While non-causal approaches to explanations [34, 37, 44, 55, 78, 79] can support a variety of aggregate functions, methods that are based on causal estimates typically focus on averages [36, 41, 59, 82, 83].

4.1.1 Mutable and immutable attributes. We assume the attributes set $\mathbb{A} \setminus \{O\}$ is partitioned into two disjoint sets: *mutable* attributes that can be used to define what affects the outcome (e.g., years of coding professionally, education) and *immutable* attributes, which are inherent and cannot be changed (e.g., ethnicity, gender). We use immutable attributes to define the subpopulations. Formally, $\mathbf{I}=\{I_1, I_2, \dots\} \subset \mathbb{A}$ denotes the set of immutable attributes and $\mathbf{M}=\{M_1, M_2, \dots\} \subset \mathbb{A}$ denotes the set of mutable attributes, where $\mathbf{M} \cap \mathbf{I} = \emptyset$ and the outcome $O \notin \mathbf{M} \cup \mathbf{I}$. This makes sure that explanations consist solely of mutable attributes that can imply corrective measures to reduce the disparity. We assume that a domain expert provides this categorization of attributes. This is similar to prior work on counterfactual explanations, where certain attributes are excluded in the explanation as they are non-actionable [22, 30, 31].

Definition 4.2 (disparity explanation). *Given a database instance D over a schema $\mathbb{A}$, two disjoint sets of mutable and immutable attributes $\mathbf{M} \subset \mathbb{A}$ and $\mathbf{I} \subset \mathbb{A}$, an outcome variable $O \in \mathbb{A}$, and two groups of interest g_1 and g_2, a disparity explanation ϕ is defined as a pair of patterns (ψ_g, ψ_e) where:*

(1) ψ_g is defined by attributes in $\mathbf{I}\subset\mathbb{A}$, highlighting a subpopulation with significant disparity between g_1 and g_2 in terms of AVG*(O).*

(2) ψ_e is defined by attributes in $\mathbf{M}\subset\mathbb{A}$, indicating a treatment that can explain the disparity between g_1 and g_2 within $\psi_g(D)$.

To assess the impact of the treatment ψ_e on the outcome O within the subpopulation $\psi_g(D)$, we compare the causal effect of ψ_e on O within the two subpopulations: $(\psi_g \wedge \psi_{g_1})(D)$ and $(\psi_g \wedge \psi_{g_2})(D)$.

Example 4.3. *Continuing with our running example, where g_1 is* analysts *and g_2 is* back-end developers*, an example disparity explanation is: Among white individuals aged between 25–34, having 6-8 years of professional coding experience boosts the* TC *for* analysts *more than* back-end developers*. Here, the subpopulation pattern ψ_g is defined by* Ethnicity = White ∧ Age = 25 − 34 *and the treatment pattern ψ_e is* YrsProfCoding *= 6–8. Within this subpopulation, the average* TC *for* analysts *and* back-end developers *are \$115K and \$105K, respectively (gap is \$10K).*

4.2 Problem Formulation

We formally define the problem of finding disparity explanations. We assume that we are given a database instance D with schema $\mathbb{A}$, a causal model $\mathcal{G}_D$ on $\mathbb{A}$, outcome $O \in \mathbb{A}$, and two groups g_1 and g_2 defined by the patterns ψ_{g_1} and ψ_{g_2}. Let $\{\phi_1, \phi_2, \dots, \phi_l\}$ be a set of possible disparity explanations. Our goal is to find a bounded-sized set of disparity explanations Φ to identify subsets of the data that (1) provide insights into the disparity between g_1 and g_2, and (2) avoid redundancy across subsets to cover different data regions.

Usefulness of an explanation. A disparity explanation $\phi = (\psi_g, \psi_e)$ is considered useful if (i) it reveals a significantly different effect of the treatment ψ_e on the outcome O for the subpopulation $\psi_g(D)$ between g_1 and g_2, and (ii) it constitutes a significant portion of the data. To this end, we define the *disparity score*, which measures the magnitude of the disparity, and the *support* of a disparity explanation that allows us to eliminate disparity explanations that constitute only minor portions of the data.

The disparity score Δ of a disparity explanation $\phi = (\psi_g, \psi_e)$ measures the absolute difference between the two CATE values: one computed over the subpopulation $(\psi_g \wedge \psi_{g_1})(D)$ and the other over $(\psi_g \wedge \psi_{g_2})(D)$. The difference is normalized by the maximal outcome value. Formally,

$$\Delta(\phi) = \frac{\left|CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_1}) - CATE_{\mathcal{G}_D}(\psi_e, O|\psi_g \wedge \psi_{g_2})\right|}{\max\{|o| \mid o \in O\}}$$

In order to prioritize disparity explanations that cover a large portion of the given database, we use the notion of *support*. The support of a disparity explanation $\phi = (\psi_g, \psi_e)$ is defined by the fraction of tuples $\in D$ that take part in the explanation, namely, tuples that satisfy the patterns in the disparity explanation. Formally,

$$support(\phi) = \frac{|\psi_{g\wedge g_1}(D) \cup \psi_{g\wedge g_2}(D)|}{|D|}$$

Intuitively, the higher the support of a disparity explanation, the more interesting it is, as it applies to a larger portion of the population. We prefer disparity explanations with high support.

Example 4.4. *Continuing from Example 4.3, the support for the disparity explanation, over the subpopulation White individuals aged between 25 to 34—working as either analysts or back-end developers—is* $\frac{16{,}508}{47{,}702} = 34.6\%$*. The disparity score for the corresponding explanation—with the identified cause of having 6–8 years of coding experience—is* $\frac{|44{,}058-10{,}552|}{2{,}000{,}000} = 0.016$*. Note that this is a real explanation found during our empirical analysis (Table 5, row 5).*

Diversity among the disparity explanations. We are interested in a diverse set of disparity explanations to reveal and explain the difference in outcome for the two groups of interest. Given two groups of interest g_1 and g_2, we use $D_{g_1 \cup g_2}$ to denote the subset of D containing tuples that belong to at least one of the groups. More formally, $D_{g_1 \cup g_2} = \psi_{g_1}(D) \cup \psi_{g_2}(D)$. Given two disparity explanations $\phi = (\psi_g, \psi_e)$ and $\phi' = (\psi_{g'}, \psi_{e'})$, defined over subpopulations g and g', respectively, and the same outcome variable O, we use the Jaccard similarity between $\psi_g(D_{g_1 \cup g_2})$ and $\psi_{g'}(D_{g_1 \cup g_2})$ to measure the similarity between ϕ and ϕ'. Formally:

$$\textsc{sim}(\phi, \phi') = \frac{|\psi_g(D_{g_1 \cup g_2}) \cap \psi_{g'}(D_{g_1 \cup g_2})|}{|\psi_g(D_{g_1 \cup g_2}) \cup \psi_{g'}(D_{g_1 \cup g_2})|}$$

We are now ready to formally define the problem of selecting disparity explanations. At a high level, our goal is to select a bounded-sized diverse set of disparity explanations with support above a given threshold, such that their combined disparity score is maximized, with bounded pairwise similarity to reduce redundancy.

Problem 1 (Disparity Explanation Selection). *Given a database instance D with schema $\mathbb{A}$, a causal model $\mathcal{G}_D$ on $\mathbb{A}$, outcome $O \in \mathbb{A}$, two groups of interest g_1 and g_2, a set of possible disparity explanations Φ_c, a budget $k \in \mathbb{N}^+$, a support threshold σ, and a similarity threshold τ, select a disparity explanation set $\Phi \subseteq \Phi_c$, such that:*

*(1) **(size constraint)** $|\Phi| \leq k$,*
*(2) **(support constraint)** $\forall \phi_i \in \Phi$, $support(\phi) \geq \sigma$,*
*(3) **(diversity constraints)** $\forall \phi_i, \phi_j \in \Phi$, $\textsc{sim}(\phi_i, \phi_j) \leq \tau$, and*
*(4) **(objective)** $\Delta(\Phi) = \sum_{\phi \in \Phi} \Delta(\phi)$ is maximized.*

Complexity Analysis. A naïve solution to Problem 1 requires (1) materializing all subsets $\Phi \subseteq \Phi_c$ s.t $|\Phi| \leq k$, (2) validating each subset w.r.t the support and diversity constraints, and (3) finding the valid subset that maximizes the objective function. Note that the number of possible patterns grows exponentially with the number of attributes and their domain sizes, leading to an exponential explosion of the number of candidate disparity explanations Φ_c. Concretely, $|\Phi_c| = \Pi_{A_i \in \mathbf{M}}(|dom(A_i)| + 1) \cdot \Pi_{A_j \in \mathbf{I}}(|dom(A_j)| + 1)$, rendering enumeration of all possible explanations infeasible. Next, we show that even if the full search space could be materialized, finding the optimal solution remains intractable.

Proposition 4.1. *Given a set of candidate disparity explanations Φ_c, a budget k, a support threshold σ, a similarity threshold τ, and a bound B, determining whether $\exists\Phi \subseteq \Phi_c$ s.t $|\Phi| \leq k, \forall\phi_i \in \Phi$, $support(\phi) \geq \sigma, \forall\phi_i, \phi_j \in \Phi$, $\textsc{sim}(\phi_i, \phi_j) \leq \tau$ and $\sum_{\phi\in\Phi} \Delta(\phi) \geq B$ is NP-hard.*

The proof is given in our technical report [11]. It is based on a reduction from the Independent Set problem, indicating that the problem is hard w.r.t the number of disparity explanations, which itself is exponential in the number of attributes, as explained above.

4.3 System Parameters

Data-Specific and Scenario-Specific Parameters. We draw distinction between two types of users: an "admin" or domain expert (equivalent to DB administrator in traditional RDBMS), who will set up ExDis for the "end-users" (equivalent to SQL programmers). The admin is responsible for setting up *data-specific parameters*—such as mutable and immutable attributes $\mathbf{M}$ and $\mathbf{I}$, the causal DAG $\mathcal{G}_D$, and the support threshold σ. These parameters are scenario-agnostic and are set according to the data properties such as whether it is realistic to treat certain data attributes and the known causal relationships among data attributes. The distinction between immutable and mutable attributes ensures explanations are actionable for policymakers (e.g., race cannot be changed) but typically requires external knowledge to identify. However, the recent rise of powerful LLMs with general-domain knowledge can help bypass the need for domain expertise in identifying immutable attributes.

In contrast, the *scenario-specific parameters* are specified by the end-users. These include (1) the outcome attribute O and two groups of interest g_1 and g_2 based on the desired scenario, (2) the desired number of explanations k, and (3) the diversity threshold τ.

Parameter Tuning. Several established ways exist for setting the support threshold σ. E.g., domain experts often set σ based on what counts as "frequent enough" to be meaningful for their applications. For instance, in retail data mining, 1% of transactions are often considered actionable, while in medical data, even rare patterns can be important. Alternatively, in an exploratory tuning, a common practice is to start with a relatively high σ (to keep the search space manageable), then gradually lower it until the number of candidate explanation patterns becomes too high (resulting in performance regression). Finally, an automated method is the elbow method [7], where one can plot the number of candidate patterns vs. σ to observe a sharp increase to choose σ near that point. Tuning the diversity threshold (τ) can be guided by the budget parameter k: a high value for τ may result in fewer than k explanations, which indicates infeasibility (k explanations do not exist that satisfy the diversity requirement). Thus, in an exploratory setting, a practical way to tune τ is to set it such that k explanations are retrieved.

5 The ExDis Algorithm

Since the number of possible disparity explanations can grow exponentially with the number of attributes and their domain values, enumerating all possible explanations is infeasible. Moreover, as shown in Proposition 4.1, even if the full search space could be materialized, finding the optimal solution remains NP-hard. We therefore propose a highly scalable heuristic approach. Our method builds on prior work in subpopulation mining [4] (Section 5.1) and explanation generation [82] (Section 5.2), but our problem formulation is novel and introduces new challenges. A key contribution of our work is the integration of pattern mining with causal analysis in a practically efficient

Algorithm 1: Apriori-Based Subpopulation Miner

```
Input: Dataset $D$, immutable attributes $\mathbf{I}$, min support threshold $\sigma$
Output: Set of candidate subpopulations $C$
$C_1 \leftarrow \{\{(A = v)\} \mid A \in \mathbf{I}, v \in \mathrm{dom}(A), \mathrm{support}(A = v) \geq \sigma\}$
$C \leftarrow C_1$
$k \leftarrow 2$
while $C_{k-1} \neq \emptyset$ do
    $\mathcal{L}_k \leftarrow$ GenerateCandidates($C_{k-1}$)
    $C_k \leftarrow \emptyset$
    foreach $c \in \mathcal{L}_k$ do
        $\mathrm{support}(c) \leftarrow \frac{|\{r \in D | r \text{ satisfies } c\}|}{|D|}$
        if $support(c) \geq \sigma$ then
            $C_k \leftarrow C_k \cup \{c\}$
    $C \leftarrow C \cup C_k$
    $k \leftarrow k + 1$
return $C$
Function GenerateCandidates($C_{k-1}$):
    return all $k$-itemsets formed by joining pairs of $(k-1)$-itemsets in $C_{k-1}$ that share the first $k-2$
      items, and whose all $(k-1)$ subsets are in $C_{k-1}$
```

way. Specifically, our framework must explore a vast search space to identify subpopulations that exhibit both significant disparities and high-quality causal explanations. In addition, it must satisfy a diversity constraint that requires computing similarity between *every* pair of explanations. To improve efficiency, we employ clustering. As we demonstrate in Section 6.2, no existing method can be directly applied to solve this problem. Although ExDis does not provide theoretical guarantees, our experiments show that it achieves performance comparable to that of an exhaustive search that computes the optimal solution.

The ExDis framework comprises three steps: (1) the *subpopulation miner*, which identifies subpopulations with sufficient support; (2) the *explanation miner*, which uncovers causal explanations for each candidate subpopulation; and (3) the *greedy search*, which efficiently selects k explanations adhering the diversity constraint.

5.1 Subpopulation Miner

Our first objective is to identify candidate subpopulations (data regions) where the disparity between groups g_1 and g_2 is significant and supported by sufficient data. To this end, we adapt the classical *Apriori* algorithm [4] to efficiently discover all subpopulations whose support exceeds a given threshold σ. This procedure is restricted to the set of immutable attributes $\mathbf{I}$, ensuring that the resulting candidate subpopulations are defined exclusively by immutable attributes. The full pseudocode is given in Algorithm 1, which modifies the standard Apriori algorithm to restrict candidate generation and frequency counting to the immutable attributes $\mathbf{I}$.

ExDis supports various use cases, such as investigating surprising observations, debugging fairness issues, or identifying reverse trends. Each scenario may require search for specific subpopulations where the average outcome for g_1 is either higher or lower than that for g_2. To accommodate this, we introduce a filtering step, which retains subpopulations that meet the relevant condition.

5.2 Explanation Miner

For each subpopulation pattern ψ_g found in the previous step, given an outcome variable O, and two groups g_1 and g_2, the next step is to explain the disparity within the subpopulation, i.e., to identify the treatment pattern ψ_e with the highest disparity score. To this end, we adapt the treatment mining step of CauSumX [82] to our setting. CauSumX provides causal explanations for the results of aggregate queries. An explanation pattern consists of a set of tuples from the aggregate view (i.e., the query results), and a treatment pattern is used to quantify the causal effect (in terms of CATE) of the treatment on the outcome within the relevant subview. CauSumX employs a heuristic lattice traversal approach to identify promising treatment patterns with high CATE values. However, unlike CauSumX, which estimates the CATE value, we estimate the disparity score of the treatment pattern under consideration. Furthermore, we adjust the search to the context, e.g., when exploring parts of the data where the average outcome for g_1 exceeds that of g_2, the explanation should elucidate this phenomenon—specifically, we aim to identify treatments that favor g_1 over g_2, and filter out the others.

We note that CauSumX operates only on aggregate views (i.e., the results of aggregate queries), where considering a relatively small number of grouping patterns is sufficient (the average was 24 [82]). In contrast, identifying a treatment for each subpopulation requires searching for subpopulations with significant disparity across the entire dataset. This requires consideration of a much larger number of potential grouping patterns (in our experiments, the average was 184). As a result, this module is the primary bottleneck of ExDis. To improve runtime, we introduce the following optimizations:

Limiting patterns. To ensure conciseness (and thus interpretability) of explanations while reducing runtime, we restrict the search for treatment patterns to at most two predicates (similar to [3]). However, this is *not* a fundamental limitation of our technique. If the user wishes to get longer explanations, ExDis can explore patterns with any number of predicates. This, of course, comes with a cost of a longer runtime. In our evaluation, even when allowing 3 predicates, most explanations had at most two predicates. Thus, to optimize the runtime, we limit the number of predicates to 2.

Parallelization. The independence of subpopulations allows treatment pattern discovery to be performed in parallel. We exploit this property to parallelize the computation, thereby improving scalability and reducing runtime.

Caching. Computing the disparity score for a given treatment pattern requires adding it as a node to the underlying causal DAG [82]. Often, this may lead to a DAG that has been previously encountered. To avoid redundant computations, we cache results related to previously encountered causal DAGs.

Sampling. As was done in [82], instead of focusing on obtaining precise CATE values, we estimate CATE from a random sample of the data. We use a fixed sample size of 50,000 tuples, guided by our empirical findings, which indicates that this sampling size achieves highly accurate CATE estimations while maintaining a relatively low runtime. However, the sampling ratio is a customizable system parameter and the user is free to tune it for more accurate results.

5.3 Greedy Search

Given the set of candidate disparity explanations $\{\phi_i\}_{i=1}^{l}$ obtained in the previous two steps, our goal is to identify a set of k explanations with the highest disparity scores, adhering to the constraints of Problem 1. A key challenge is balancing scalability and diversity in selecting candidate disparity explanations. To address this, we introduce a clustering step: rather than evaluating all candidate explanations individually, we group similar subpopulations and assign a representative explanation to each cluster, thereby reducing redundancy and improving efficiency while maintaining coverage

Table 3. Details of the datasets for experiments and case studies.

Dataset	#Tuples	|**I**|	|**M**|	g_1	g_2	$\|g_1\|$	$\|g_2\|$	O	AVG$_{g_1}$	AVG$_{g_2}$
Stack Overflow (SO) [1]	47,702	4	6	Data/business analysts	Back-end developers	4,088	28,987	Total Compensation (TC)	$106K	$96K
American Community (ACS) [71]	1,420,652	15	10	Manual labor	Overall data	99,790	1,420,652	Likelihood of having a health insurance	78.6%	91.5%
Medical Expenditure Panel (MEPS) [2]	13,528	7	5	Males	Non-males	5,731	7,797	Likelihood of feeling nervous frequently	41.6%	46.9%

of distinct subpopulations. The rationale is twofold: (1) it enables us to handle a large number of candidate subpopulations without exhaustive enumeration, and (2) selecting a representative from each cluster ensures that the final set of explanations is diverse, avoiding multiple explanations that represent the same subpopulation.

Specifically, we proceed as follows. We first cluster the candidate explanations using a hierarchical clustering algorithm (using Scipy [74] implementation) based on the symmetric difference among the subpopulations. Then, from each cluster, we select a random representative explanation and assign its disparity score to the entire cluster. We then iteratively select k disparity explanations. At the first iteration, we pick a random explanation from the cluster with the highest disparity score. At the $j-th$ iteration (for $1 < j \leq k$), we select the explanation ϕ^* such that:

$$\phi^* = \underset{\phi \in \Phi \wedge \textsc{sim}(\phi, \phi') < \tau}{\arg\max} \Delta(\phi), \quad \text{for } \phi' \in \Phi_{j-1}.$$

where Φ is the set of candidate explanations that consist of a random explanation from each cluster, and Φ_{j-1} is the set of explanations selected up to iteration j.

5.4 Complexity Analysis

The maximum number of disparity explanations in a database D with attributes $\mathbb{A}$ is bounded by $|D|^{|\mathbb{A}|}$ (considering both subpopulation and treatment patterns), which is polynomial in terms of data complexity, assuming a fixed schema [72]. Our greedy search is also polynomial in the number of explanations considered. Additional operations, such as calculating CATE values, are polynomial in D, leading to worst-case polynomial data complexity. As we demonstrate in Section 6.4, ExDis is capable of efficiently handling large, high-dimensional datasets in practice.

6 Experimental Evaluation

We present an experimental evaluation of the effectiveness of ExDis in practical settings. We aim to address the following questions:

- **Q1**: How does the quality of ExDis-generated disparity explanations compare to that of existing methods? (Section 6.2)
- **Q2**: How is the quality of the explanations affected by various parameters, and how to tune the system parameters? (Section 6.3)
- **Q3**: How efficient and scalable is ExDis? (Section 6.4)
- **Q4**: How do our proposed optimizations affect ExDis' runtime performance? (Section 6.5)

6.1 Experimental Setup

All experiments were performed on a Windows computer, Intel CPU, with 16 GB memory. We implemented ExDis in Python 3 and used DoWhy library [65] to compute the CATE values. CATE values are estimated using a linear regression model. For simplicity of visualization, we omit p-values, but note that all reported causal effects are statistically significant (i.e., $p < 0.05$). Our source code is publicly available [10].

6.1.1 Datasets, causal DAGs, and preprocessing. We used three popular datasets (Table 3) and obtained the corresponding causal DAGs given in prior work [84]. To process continuous numerical attributes, we applied equal-width binning across 10 bins.

SO: The Stack Overflow Developer Survey [1] dataset contains responses from developers worldwide, covering topics such as professional experience, education, technologies used, and employment-related information, such as annual total compensation (TC).

ACS: The American Community Survey (ACS) [71] is a nationwide survey conducted by the U.S. Census Bureau, with demographic, social, economic, and housing data. We focused on 7 states: California, Texas, Florida, New York, Pennsylvania, Illinois, and Ohio.

MEPS: The Medical Expenditure Panel Survey (MEPS) [2] dataset provides information on health-care utilization, expenditures, insurance coverage, and demographics of individuals in the U.S.

6.1.2 System parameters. Unless otherwise specified, we used the following default parameters: the explanation set size $k = 5$; minimum support threshold $\sigma = 0.05$ (considering only groups covering at least 5% of the data); maximum similarity threshold for the diversity constraint $\tau = 0.55$ based on Jaccard similarity; and 10 clusters in the greedy search phase.

6.1.3 Use cases. Throughout our experimental evaluation, we examine three scenarios that represent different use cases of ExDis (as mentioned in Section 1). The description of the groups, the outcome variables, and relevant statistics are given in Table 3.

(1) Investigating a surprising fact. In tech, developers generally earn more than analysts. However, an analysis of the SO dataset revealed a surprising trend: data analysts (g_1) earn more on average than backend developers (g_2). To analyze this, we fix the outcome O as total compensation (TC). We aim to identify which subpopulations significantly contribute to this disparity and the underlying explanations in terms of treatments that favor g_1 over g_2.

(2) Fairness debugging. In the ACS dataset, we observe people in certain type of occupation to have lower than average rate of health-insurance coverage. Specifically, we fix the outcome O to indicate whether an individual holds a health insurance; g_1 represents individuals employed in manual-labor occupations (cleaning, maintenance, farming, fishing, construction, etc.), where the health insurance coverage rate is only 78.6%; and g_2 represents the entire dataset across all occupations, with a coverage rate of 91.5%. Our objective is to investigate the underlying factors contributing to this discrepancy by identifying subpopulations for which the average insurance coverage rate of g_1 is significantly lower than that of g_2, along with a causal explanation specific to each subpopulation.

(3) Finding reverse trends. We investigate an intriguing observation in the MEPS dataset: while typically males (g_1) have a lower likelihood of feeling nervous frequently than non-males (g_2), our analysis reveals subpopulations where a reverse trend holds. To dig deeper, we fix the outcome O to indicate whether an individual feels nervous frequently. Our goal is to pinpoint subpopulations where the relative trend involving g_1 and g_2 is reverse compared to the global trend and explore the underlying causes.

6.1.4 Baselines. We consider the following baselines:

Brute Force. We employ an exhaustive Brute Force algorithm, which considers all possible k-sized explanation sets. Note that in the absence of an absolute ground-truth, the results obtained from this technique can be treated as the optimal solution for Problem 1.

Top-k. This baseline ignores the diversity constraint and simply returns the top-k explanations ranked by their disparity scores.

XInsight. XInsight [41] is designed to identify both causal and non-causal patterns to explain disparities between two groups in aggregate SQL queries. Unlike ExDis, which provides local (possibly different) explanations for each subpopulation, XInsight provides a single global explanation for the entire data. Since XInsight includes a causal discovery phase, we ensure a fair comparison as

Table 4. Overall disparity scores (Δ), runtimes, and diversity visualizations for explanations generated by various baselines and ExDɪs across three use cases. We report the disparity scores w.r.t the Brute Force baseline since it gives the ground truths. In the heat-map, the diagonal is black, denoting 0 self distance from an explanation to itself. Lighter colors denote less similar pair of explanations, offering diversity.

Dataset	SO				ACS				MEPS			
Approach	Δ (%)	Runtime (s)	#Explanations	Diversity	Δ (%)	Runtime (s)	#Explanations	Diversity	Δ (%)	Runtime (s)	#Explanations	Diversity
Brute Force	100	181	5		100	3130	5		100	19	5	
Top-K	118	180	5		121	1514	5		166	14	5	
DivExplorer	N/A	33	0	N/A	86	1055	5		N/A	4	0	N/A
FairDebugger	N/A	2258	0	N/A	26	717	2		N/A	103	0	N/A
ExDɪs (this paper)	55	62	5		84	1170	5		100	17	5	

follows: for each treatment pattern (i.e., explanation) identified by ExDɪs, we report its causal effect over the entire dataset (in the "Global" column in Tables 5–7). We aim to empirically demonstrate that subpopulation-specific explanations may not be valid globally.

DivExplorer. DivExplorer [49] analyzes the behavior of classification models to identify data regions where a performance metric (e.g., false positive rate) deviates significantly from its value over the entire dataset. Given a divergence metric, it identifies data subsets, defined by patterns, where the metric shows a significant disparity compared to the overall population. We use DivExplorer as a baseline, setting the divergence function to the disparity score. To adapt it to our setting, we use a fixed treatment. This treatment is selected as the one that yields the highest disparity score between the two groups of interest in the overall data. We then used the last step of ExDis to find a k-size solution.

FairDebugger. FairDebugger [68] identifies training data subsets that contribute to fairness violations in random forest models by evaluating how their removal affects model outcomes. To adapt FairDebugger to our setting, we fix the treatment pattern to the one exhibiting most disparity across the entire dataset. We then assess the influence of each subpopulation by measuring the change in disparity after removing it. The difference in disparity scores quantifies the subpopulation's contribution to the overall disparity. We then used the last step of ExDis to find a k-size solution.

6.2 Explanation Quality

Table 4 shows a quantitative comparison contrasting the explanations generated by ExDɪs with those of the baselines over three use cases. Details of the explanations generated by each of these approaches are in our technical report [11].

6.2.1 Investigating a surprising fact (SO). The explanations generated by ExDɪs are shown in Table 5. Notably, ExDɪs identifies subpopulations where the average salary of analysts is higher than that of back-end developers, with the salary gap often exceeding that observed in the overall population (analysts: \$106,542; back-end developers: \$96,609). For almost all cases, ExDɪs provides a different causal explanation to highlight the factor contributing to the disparity within the subpopulations. In all five explanations, the causal effect of the chosen treatment on the entire population is not statistically significant as shown in the "Global" column, emphasizing the importance of providing "local" explanations for each subpopulation. *This observation highlights the distinction between our approach and XInsight (that provides a global explanation, as shown in the Global column), demonstrating that subpopulation-level explanations differ from those provided at the entire population level.*

In this scenario, ExDɪs shares only 1 out of 5 explanations with the optimal solution by Brute Force. However, ExDɪs selects a highly diverse set of explanations, achieving 55% of the optimal disparity score produced by Brute Force. In contrast, although the Top-k baseline achieves high

Table 5. Disparity explanations discovered by ExDis for the SO dataset. Subpopulation patterns are highlighted in Orange, while treatment patterns are highlighted in Cyan. We highlight the two groups of interest using Yellow and Pink. The first explanation highlights that for White heterosexual individuals whose parents attended secondary school, the average TC for `analysts` observes an increase of $154,024 when the treatment hoping to become a manager in the next 5 years is applied. In contrast, the same treatment yields only an increase of $31,354 for `back-end developers`. We observe that this treatment is not a good explanation globally due to not being statistically significant.

Disparity Explanation	**Support**	**Total Compensation (TC)**				**Δ**
		Subpopulation (ExDis)		**Global (XInsight)**		
		Average	**CATE**	**Average**	**CATE**	
For White heterosexual individuals whose parents attended secondary school, TC growth is more influenced by hoping to become a manager in the next 5 years for analysts compared to back-end developers.	11.14%	$129,749 $110,026	$154,024 ↑ $31,354 ↑	not statistically significant		0.061
For White males aged between 18-24 years old, TC growth is more influenced by have 3-5 years in professional coding for analysts compared to back-end developers.	10.64%	$90,909 $67,754	$75,692 ↑ $24,164 ↑	not statistically significant		0.025
For males aged between 25-34 years old whose parents hold a bachelor degree, TC growth is more influenced by working in a company size between 100 - 499 workers for analysts compared to back-end developers.	13.21%	$105,694 $96,085	$70,069 ↑ $19,807 ↑	not statistically significant		0.025
For White heterosexual males aged between 25-34 years old whose parents hold a Master degree, TC growth is more influenced by not having coding as a hobby for analysts compared to back-end developers.	7.56%	$117,879 $103,957	$84,648 ↑ $43,292 ↑	not statistically significant		0.020
For White individuals aged between 25–34, TC growth is more influenced by working 6-8 years in professional coding for analysts compared to back-end developers.	34.69%	$115,777 $105,988	$44,058 ↑ $10,552 ↑	not statistically significant		0.016

Table 6. Disparity explanations discovered by ExDis for the ACS dataset.

Disparity Explanation	**Support**	**Likelihood of having a health insurance**				**Δ**
		Subpopulation (ExDis)		**Global (XInsight)**		
		Average	**CATE**	**Average**	**CATE**	
For White individuals from the Southern region who speak Spanish, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	8.18%	52.42% 75.17%	7.74% ↓ 4.87% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.126
For individuals from the Southern region who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	25.09%	72.77% 81.24%	6.38% ↓ 1.83% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.082
For White natives individuals from Texas who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	12.21%	71.86% 80.52%	5.20% ↓ 2.83% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.080
For White males from the Southern region, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	18.00%	67.56% 74.20%	4.15% ↓ 2.69% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.068
For individuals who were born in USA, the likelihood of having a health insurance decreases when they have no personal earnings for manual labor occupations, whereas it increases for all occupations.	75.67%	84.11% 88.97%	3.46% ↓ 1.36% ↑	78.61% 91.58%	8.02% ↓ 3.63% ↓	0.048

disparity scores, the selected explanations are highly similar to one another (as indicated by the heatmap in Table 4), highlighting the importance of incorporating similarity-awareness when selecting explanations.

In this scenario, both DivExplorer and FairDebugger failed to identify any valid solution. This is because they relied on a fixed treatment (as discussed in Section 6.1.4). In every subpopulation they identified, either the average income of backend developers exceeded that of analysts, or the treatment disproportionately benefited backend developers. According to our problem definition, such cases do not qualify as valid solutions.

6.2.2 Fairness debugging (ACS). The disparity explanations generated by ExDis for the second use case are presented in Table 6. ExDis identified subpopulations in which the percentage of manual-labor workers with health insurance was lower than in the overall population. Notably, all explanations (i.e., treatments) involved personal earnings, underscoring the strong causal relationship between income level and the likelihood of having health insurance in the U.S. Interestingly, in this case, a single explanation was sufficient to account for the disparity across all identified subpopulations. However, this explanation behaves differently when applied to the entire dataset. Specifically, for the overall population, not having personal earnings decreases the likelihood of having health insurance across all occupations. In contrast, for each of the reported subpopulations, not having personal earnings actually increases the probability of having health insurance when considering all individuals within those subpopulations. We note that even in the optimal solution found by Brute Force (as well as in the Top-K solution), only two distinct treatments were selected. This suggests that, in this case, the space of relevant treatments is limited, indicating few plausible explanations for the observed disparity.

Compared to the Brute-Force baseline, ExDis recovered 2 out of the 5 disparity explanations while exploring a significantly smaller search space. Top-K overlapped with Brute-Force in just one explanation and exhibited higher similarity among its selected explanations compared to ExDis. Both DivExplorer and FairDebugger identified different subpopulations than ExDis; FairDebugger detected only two (highly overlapping) subpopulations, while DivExplorer found five, though with high similarity among them.

6.2.3 Finding reverse trends (MEPS). Table 7 shows the disparity explanations for MEPS. ExDis identifies subpopulations where the average likelihood of experiencing nervous attacks very frequently for males (g_1) is higher than that of non-males (g_2), in contrast to the opposite trend observed in the overall population. ExDis generated a solution identical to that of the Brute-Force approach, while Top-k identified explanations with substantial overlap among themselves. In this scenario, both DivExplorer and FairDebugger failed to produce meaningful results. The subpopulations they identified either did not align with the use case. Specifically, they did not exhibit a trend that reversed the one observed in the overall dataset, or, in the few relevant subpopulations they did find, the fixed treatment yielded a disparity score of zero, rendering the explanation ineffective for explaining the observed disparity.

Results Summary

- The top-k baseline results in overlapping disparity explanations, demonstrating the need to consider the similarity among selected explanations.
- The output of ExDis closely matches that of Brute Force, demonstrating that our solution prioritizes efficiency without compromising quality.
- Explanations for disparity at the entire population level (as generated by XInsight) do not necessarily account for disparities within subpopulations, highlighting the need to find a specific local explanation for each subpopulation.
- DivExplorer and FairDebugger were less successful in identifying subpopulations with high disparity between g_1 and g_2, resulting in low-score explanations.

Remark. Note that DivExplorer and FairDebugger serve as baselines only for identifying subpopulations that exhibit significant disparities between the two groups of interest. However, unlike ExDis, they do not offer any causal explanations. To enable comparison, we use a fixed treatment while using these two baselines, which is not necessarily the optimal treatment that maximizes subpopulation-level disparity scores. Therefore, these approaches result in a low total disparity

Table 7. Disparity explanations discovered by ExDis for the MEPS dataset.

Disparity Explanation	Support	Likelihood of feeling nervous frequently				Δ
		Subpopulation (ExDis)		Global (XInsight)		
		Average	CATE	Average	CATE	
For individuals who were never married, are from the Southern region, don't have a doctor's recommendation to exercise, and aren't diagnosed with Diabetes, the likelihood of feeling nervous frequently decreases less for males who do not currently smoke compared to non-males.	5.78%	46.08% 41.71%	13.06% ↓ 20.73% ↓	37.58% 45.10%	3.39% ↓ 3.94% ↓	0.076
For individuals who are White, were never married, are under 29, and don't have Asthma or Diabetes, the likelihood of feeling nervous frequently decreases less for males than non-males when they are uninsured for the whole year.	8.41%	49.20% 48.95%	14.16% ↓ 18.48% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.043
For individuals who were never married, don't have a doctor's recommendation to exercise, and were born in USA, the likelihood of feeling nervous frequently increases more for males who have private insurance compared to non-males.	14.25%	45.90% 45.74%	10.84% ↑ 7.51% ↑	37.58% 45.10%	4.63% ↑ 5.76% ↑	0.033
For individuals who are White, were never married, don't have a doctor's recommendation to exercise, and don't have Asthma, the likelihood of feeling nervous frequently decreases less for males who were uninsured the whole year compared to non-males.	9.74%	47.25% 46.64%	12.52% ↓ 14.09% ↓	37.58% 45.10%	7.02% ↓ 4.86% ↓	0.015
For individuals aged between 30-42 years who don't have Diabetes, the likelihood of feeling nervous frequently increases more for males who have health insurance compared to non-males.	19.64%	43.93% 43.57%	7.47% ↑ 6.47% ↑	37.58% 45.10%	2.34% ↑ 4.54% ↑	0.010

score (which is expected) when the subpopulation-level disparity scores are added to obtain the score for the objective function (Problem 1). Nonetheless, the comparison allows for a qualitative contrast between the subpopulations identified by the baselines and those discovered by ExDis.

6.3 Parameters Sensitivity

In this section, we investigate the impact of various parameters on our objective function, namely disparity score Δ. Our goal is to gain insights into effective default parameter settings.

6.3.1 Robustness to similarity threshold τ. Figure 8 (a) shows how our objective function, i.e., the disparity score Δ changes with varying similarity threshold τ across three datasets: SO, ACS, and MEPS. As τ increases, all datasets exhibit a rising trend in disparity. This is expected because low similarity threshold significantly restricts the feasible solution space. For a fixed budget k (we used $k = 11$ for this experiment, to observe the impact of τ without any other restriction), relaxing τ allows for the inclusion of a larger number of disparity explanations, which can cumulatively increase the overall disparity score. This is because more similar subpopulations are permitted to coexist, potentially capturing less diverse and more overlapping causes of disparity. However, setting τ too high undermines the goal of maintaining diversity among the explanations, as excessive overlap can dilute the quality and reduce the distinctiveness of the explanation set. Thus, τ must be carefully chosen to balance comprehensiveness and diversity.

6.3.2 Robustness to number of clusters. Figure 8 (b) shows how the disparity score Δ varies with the number of clusters across three datasets: SO, ACS, and MEPS. As the number of clusters increases, all datasets exhibit an initial rise in disparity, which eventually plateaus. Notably, the ACS dataset shows the most significant increase, with the disparity growing rapidly up to 5 clusters before stabilizing around 0.4. In contrast, SO and MEPS show more modest increases, leveling off at lower disparity scores. These results suggest that finer-grained clustering help improve the disparity scores, but has diminishing return.

6.3.3 Robustness to the Causal DAG. The quality of the solution may depend on the quality of the underlying causal DAG. To evaluate this, we assess the impact of using different causal DAGs, generated by commonly used causal discovery methods. We consider the following DAGs: (1) **DEFAULT**,

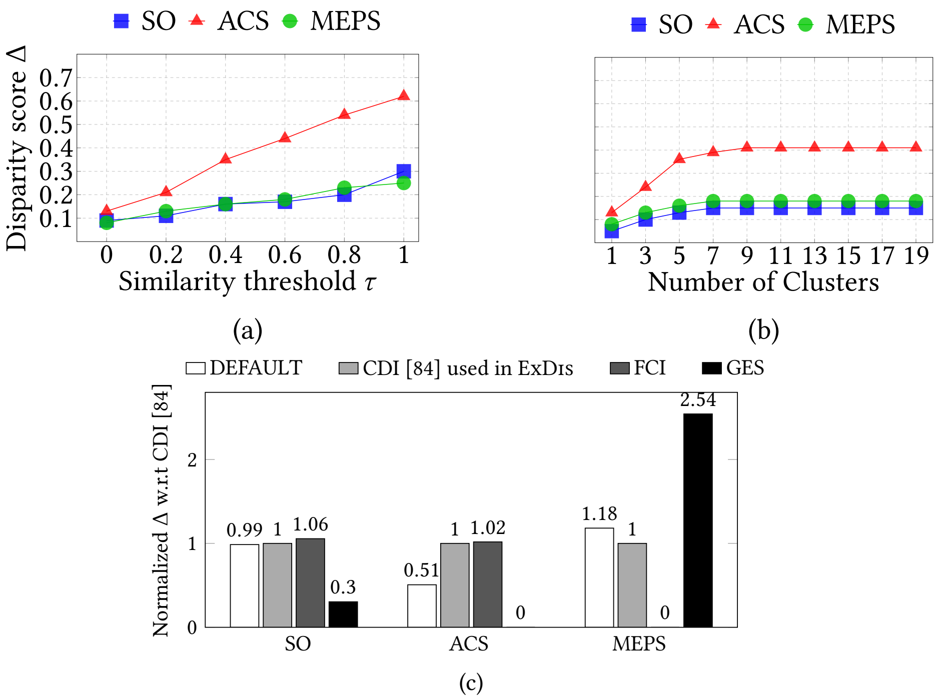

Fig. 8. Effect of various system parameters on the disparity score. (a) & (b) The absolute disparity scores are reported here to show direct impact of the similarity threshold τ and the number of clusters on the disparity score Δ. (c) Effect of casual DAG modification. Disparity scores here are shown as a relative value w.r.t the disparity score of CDI [84] (which ExDis uses).

a default two-layer causal DAG where immutable attributes affect both mutable attributes and the outcome, and mutable attributes also affect the outcome (2) Causal Data Integration (**CDI**) [84], which ExDis uses, (3) **FCI** [66], and (4) **GES** [15].

The results are depicted in Figure 8 (c). We report the relative disparity scores (Δ) computed using the different DAGs, w.r.t. the disparity score of CDI. Observe that for the default DAG, the results closely resemble those obtained using the CDI DAG employed by ExDis. In contrast, the FCI and GES DAGs produce more varied results across different use cases. This variability is due to the challenges that causal discovery algorithms face when applied to real-world data [24, 48], often leading to noisier and less reliable DAGs. Nonetheless, the disparity explanations generated using FCI and GES DAGs were largely consistent with those derived from CDI, with similar explanations selected. This suggests that ExDis remains robust and capable of producing meaningful explanations even when the underlying causal DAG is noisy or imperfect.

Results Summary

• A higher τ can increase the disparity score by allowing overlapping explanations at the cost of compromising diversity.
• Finer-grained clustering helps improve the disparity scores, but has diminishing return.
• Even with imperfect causal DAGs, ExDis is able to produce meaningful and robust disparity explanations.

Table 9. Results using different ATE estimators across SO and MEPS.

Dataset / Estimator	SO			MEPS		
	Δ (%)	Runtime (s)	#Explanations	Δ (%)	Runtime (s)	#Explanations
ExDis + Linear Regression	55	62	5	100	17	5
ExDis + Propensity Score Stratification	100	4870	5	0	331	0
ExDis + Propensity Score Weighting	94	4655	5	100	1070	5
ExDis + Propensity Score Matching	92	67379	5	100	7265	2

6.3.4 Robustness to ATE estimator. By default, we use DoWhy [65] to estimate ATE values via a linear regression model. Next, we evaluate the robustness of the generated explanations under alternative ATE estimators. Specifically, we examine three widely used methods: propensity score stratification, propensity score weighting, and propensity score matching, all implemented in DoWhy [65], following standard causal inference practices [6, 53]. The results are shown in Table 9. Our experiments reveal three main findings:

(1) As expected, the linear regression estimator is the fastest and is therefore used by default in ExDis.
(2) Explanations generated using different estimators are largely consistent, typically selecting the same treatments with only minor differences in the estimated CATE values, demonstrating that ExDis is robust to the choice of ATE estimator.
(3) The lack of results for MEPS when using the prosperity-score stratification estimator stems from the high variance of stratified ATE estimates which causes the CATE values to be not statistically significant. This is caused by sample fragmentation and limited overlap across prosperity-score bins. This behavior is consistent with standard findings in causal inference: while stratification enhances interpretability, it can reduce statistical power when strata are small or poorly balanced [45].

6.3.5 Robustness to seed selection. As discussed in Section 5.3, we select a random representative from each cluster and assign its disparity score to the entire cluster, which introduces some randomization. By default, we used a fixed seed. Next, we examine how different seeds affect the results. In particular, we run each experiment on the SO and ACS datasets using seven different random seeds, and report the average and variance of the objective value, diversity, and runtime. The results are shown in Table 10. Our results show that the results were largely consistent across different seeds. Our results suggest that ExDis is highly robust to randomization: across all seeds, the variance in both the objective value and diversity is very small (e.g., ≤ 0.09 for SO and ≤ 0.04 for MEPS), indicating that the clustering-based optimization introduces minimal noise. Similarly, runtime remains stable, with negligible variation across runs (< 0.1 s on average). These findings confirm that the random seed used for representative selection has a minor impact on both the quality and efficiency of the results.

6.4 Efficiency & Scalability

In this section, we present results demonstrating the efficiency of various components of ExDis, analyze the impact of different parameters on its runtime, and evaluate its scalability as the dataset grows both vertically (in tuples) and horizontally (in attributes).

Step-by-step breakdown of runtime: We present a step-wise breakdown of runtime in Table 11. Not surprisingly, The step "explanation miner", which focuses on identifying the causal explanation for each subpopulation, is the most computationally expensive one, accounting for over 80% of

Table 10. Average and variance (over 7 random seeds) of objective value (Δ), diversity, and runtime for ExDis.

Dataset	Δ (%) (avg ± var)	Diversity (avg ± var)	Runtime (avg ± var)
SO	51.79 (7.86)	0.2996 (0.0136)	60.283 (0.038)
MEPS	92.63 (3.25)	0.2437 (0.0091)	10.25 (0.065)

Table 11. Breakdown of runtime by steps (seconds).

Dataset	Subpopulation Miner	Explanation Miner	Fast Greedy Search
SO	0.4	70.1	0.5
ACS	3.0	1294.6	13.0
MEPS	0.3	11.3	0.1

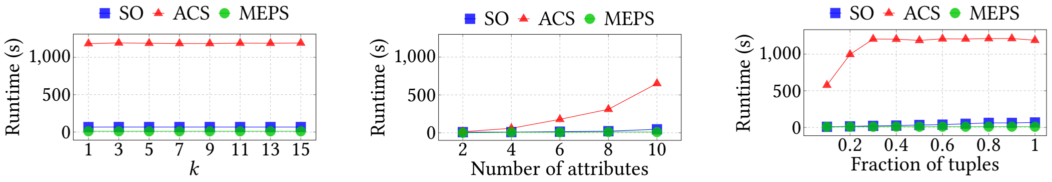

Fig. 12. Effects of various parameters on runtime: (left) the budget parameter k, (center) number of attributes, and (right) fraction of data.

the total runtime in all examined scenarios. Nevertheless, ExDis generates the solution within a reasonable time, even for large, high-dimensional datasets like ACS.

Effect of various parameters on runtime: Next, we analyze how various parameters impact runtime. Since parameter variations involve sampling, we repeat each experiment 5 times and report the average runtime across all runs.

Solution size k. Figure 12 (left) shows the impact of the solution size k on the runtime. Note that k only affects the last step (fast greedy search), which selects the explanations from the candidates mined in the previous steps. Recall that this step evaluates the pairwise intersection between the subpopulations corresponding to the disparity explanations to account for the diversity constraint. As k increases, the number of pairwise comparisons grows and so does the runtime. The effect of k on runtime is negligible for the smaller datasets (MEPS and SO) compared to the larger dataset ACS, where computing intersections among subpopulations takes longer due to the dataset's size.

Number of attributes. Figure 12 (center) shows the impact of the number of attributes on runtime. In this experiment, we randomly sampled subsets of attributes to retain and removed the rest from the dataset. The number of attributes influences the search-space size, as more attributes result in a larger set of subpopulations and treatment patterns to consider. Theoretically, runtime should grow exponentially with the number of attributes. However, this worst-case behavior was not always observed in practice, as several factors influence computation—such as the structure of the underlying causal DAG, the choice of mutable and immutable attributes, and other system parameters. Nevertheless, we observe that, as expected, the number of attributes in the dataset significantly affects runtime. For the largest dataset, ACS, we do in fact observe exponential growth in runtime as the number of attributes increases.

Dataset cardinality. Figure 12 (right) illustrates the impact of the dataset cardinality (in number of tuples) on runtime. In this experiment, we varied the dataset size using specific-sized horizontal

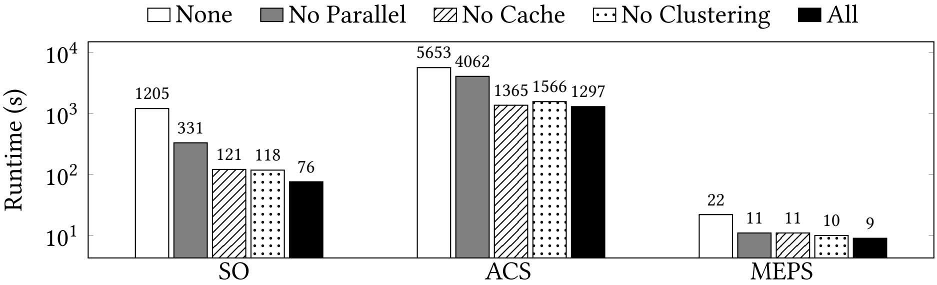

Fig. 13. Effect of various settings of using optimization techniques on runtime across three datasets. Note that the y-axis is in log scale.

dataset slices. We find that the runtime is linearly influenced by the number of tuples. For the ACS dataset, we applied the sampling optimization (during the Explanation Miner phase, as explained in Section 5.2). We found that the growth in runtime is more moderate up to around 30% of the ACS data (which is 500,000 tuples). Beyond this point, the Explanation Miner module operates on a random fixed-sized sample of 500,000 tuples.

Results Summary

- The explanation miner step of ExDɪs, which focuses on identifying the causal explanation for each subpopulation, takes the longest time, accounting for over 80% of the total runtime.
- The runtime grows (almost linearly) with the solution size k because of the pair-wise similarity computation.
- The runtime is greatly influenced by the number of attributes in the dataset, as it affects the size of the search space.
- The runtime grows linearly with the number of dataset tuples.

6.5 Ablation Study

To assess the impact of our proposed optimizations on the runtime of ExDɪs, we compare 5 variants: (1) None, implying no optimization was applied, (2) No Parallel, denoting the setting where parallelization was removed, (3) No Cache, denoting the setting where caching was removed, (4) No Clustering, denoting the setting where clustering was removed, (5) All, denoting the ExDɪs setting where all optimizations were applied.

Figure 13 illustrates how the removal of individual optimization techniques affects runtime performance across three datasets: SO, ACS, and MEPS. The y-axis is plotted on a logarithmic scale to clearly illustrate differences in runtimes. As expected, without any optimization, we observe the highest runtimes across all three datasets. We also observe that ExDɪs, with all optimizations, yield the best runtime performance. Removing clustering increases runtimes moderately, indicating clustering provides a noticeable but relatively modest speedup. Eliminating caching also causes a moderate performance degradation. While caching was introduced to avoid redundant computations, it offers relatively smaller gains. This is because when causal DAGs are small, modifying them may require less overhead than reading and writing them from cache. However, we observe substantial reduction in runtime using caching for the larger datasets SO and ACS. The absence of parallelization significantly worsens performance, especially on SO and ACS datasets, indicating its significant contribution in boosting the runtime performance across the board.

Overall, while all optimizations contribute to performance gains, parallelization yields the most substantial runtime improvements. Clustering also helps, though to a lesser extent. This experiment underscores the critical value of each optimization, highlighting how removal of any of them can degrade runtime efficiency.

7 Conclusions and Future Work

We have presented ExDɪs, a framework for discovering causal explanations for disparities between two groups of interest. ExDɪs identifies data regions where disparities are most pronounced (or reversed), and associates specific factors that causally contribute to the disparity. We acknowledge that several factors can influence the quality of the disparity explanations, including data quality, the quality of the underlying causal model, and system parameters. In Section 6.3.3, we showed that meaningful results can still be obtained even with imperfect causal DAGs. We also provided insights on tuning system parameters to achieve satisfactory results across different use cases and datasets.

ExDɪs currently operates on single-relation databases, assuming no dependencies among tuples. This design ensures compliance with SUTVA [57], a standard assumption in causal inference [21, 36, 82, 86]. While this simplifies analysis, it limits applicability to multi-relation databases where tuple dependencies naturally arise. As discussed in [82], extending treatment and grouping patterns to multi-table settings introduces substantial complexity and remains an open research direction. Nevertheless, ExDɪs can operate on normalized data by performing joins at additional computational cost, assuming no tuple-dependencies. We also assume that a causal DAG is provided by domain experts, a common practice in causal analysis [21, 36, 82, 86]. In practice, this requirement can be alleviated by leveraging existing causal discovery algorithms [24]. As shown in Section. 6.3 (Fig. 2c), ExDɪs remains robust even when the DAG is noisy or imperfect. Finally, the current implementation focuses on two-group comparisons. While extending to multiple groups is conceptually straightforward, it may introduce cognitive complexity for users, reducing interpretability. Supporting multi-group and multi-relation analyses thus presents a promising direction for future work.

Acknowledgments

This work was supported in part by the United States-Israel Binational Science Foundation (Grant No. 2024101), the Israel Science Foundation (Grant No. 934/25 & 2121/22), the NSF (Grant No. CNS-2346555), a Stena Center for Financial Technology Seed Grant, a One Utah Data Science Hub Seed Grant, and the Frankel Center for Computer Science at BGU.

References

[1] 2021. 2021 Stackoverflow Developer Survey. https://insights.stackoverflow.com/survey/2021.

[2] Agency for Healthcare Research and Quality (AHRQ). 2024. Medical Expenditure Panel Survey (MEPS) - Data Overview. Accessed: 2024-01-30.

[3] Shunit Agmon, Amir Gilad, Brit Youngmann, Shahar Zoarets, and Benny Kimelfeld. 2024. Finding Convincing Views to Endorse a Claim. *Proceedings of the VLDB Endowment* 18, 2 (2024), 439–452.

[4] 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.

[5] 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.

[6] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[7] Purnima Bholowalia and Arvind Kumar. 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications* 105, 9 (2014).

[8] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

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

[10] Tal Blau, Brit Youngmann, Anna Fariha, and Yuval Moskovitch. 2025. Causal Explanation for Disparity. GitHub repository.

[11] Tal Blau, Brit Youngmann, Anna Fariha, and Yuval Moskovitch. 2025. Causal Explanations for Disparate Trends: Where and Why? *CoRR* abs/2512.08679 (2025). arXiv:2512.08679

[12] Pierre Bourhis, Daniel Deutch, and Yuval Moskovitch. 2020. Equivalence-Invariant Algebraic Provenance for Hyperplane Update Queries. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*. ACM, 415–429.

[13] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *14th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2019, Vancouver, BC, Canada, October 20-25, 2019*. IEEE, 46–56.

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

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

[16] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Ki Hyun Tae, and Steven Euijong Whang. 2019. Automated data slicing for model validation: A big data-ai integration approach. *IEEE Transactions on Knowledge and Data Engineering* 32, 12 (2019), 2284–2296.

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

[18] 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.

[19] Daniel Deutch, Amir Gilad, and Yuval Moskovitch. 2015. Selective Provenance for Datalog Programs Using Top-K Queries. *Proc. VLDB Endow.* 8, 12 (2015), 1394–1405.

[20] 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.

[21] 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.

[22] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *SIGMOD*. ACM, 577–590.

[23] Floris Geerts, Bart Goethals, and Taneli Mielikäinen. 2004. Tiling databases. In *Discovery Science: 7th International Conference, DS 2004, Padova, Italy, October 2-5, 2004. Proceedings 7*. Springer, 278–289.

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

[25] Ricky Charles Godbolt. 2011. *Black and Blue: African Americans, Blue-Collar Bias, and the Construction Industry in Prince George's County, Maryland*. Ph. D. Dissertation. University of Phoenix.

[26] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. 2000. Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter* 2, 1 (2000), 58–64.

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

[28] Manas Joglekar, Hector Garcia-Molina, and Aditya G. Parameswaran. 2019. Interactive Data Exploration with Smart Drill-Down. *IEEE Trans. Knowl. Data Eng.* 31, 1 (2019), 46–60.

[29] Steven A Julious and Mark A Mullee. 1994. Confounding and Simpson's paradox. *BMJ* 309, 6967 (1994), 1480–1481. arXiv:https://www.bmj.com/content doi:10.1136/bmj.309.6967.1480

[30] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2023. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *Comput. Surveys* 55, 5 (2023), 95:1–95:29.

[31] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In *FAccT*. ACM, 353–362.

[32] 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).

[33] Trupti A Kumbhare and Santosh V Chobe. 2014. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies* 5, 1 (2014), 927–930.

[34] 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.

[35] Seokki Lee, Bertram Ludäscher, and Boris Glavic. 2020. Approximate Summaries for Why and Why-not Provenance. *Proceedings of the VLDB Endowment* 13, 6 (2020).

[36] Benton Li, Nativ Levy, Brit Youngmann, Sainyam Galhotra, and Sudeepa Roy. 2025. Fair and Actionable Causal Prescription Ruleset. *Proceedings of the ACM on Management of Data* 3, 3 (2025), 1–28.

[37] 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.

[38] Jinyang Li, Yuval Moskovitch, and H. V. Jagadish. 2023. Detection of Groups with Biased Representation in Ranking. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2167–2179.

[39] 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.

[40] 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.

[41] 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.

[42] 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).

[43] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. 2010. The Complexity of Causality and Responsibility for Query Answers and non-Answers. *Proceedings of the VLDB Endowment* 4, 1 (2010).

[44] 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.

[45] Luke W Miratrix, Jasjeet S Sekhon, and Bin Yu. 2013. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75, 2 (2013), 369–396.

[46] Yuval Moskovitch, Jinyang Li, and H. V. Jagadish. 2023. Dexer: Detecting and Explaining Biased Representation in Ranking. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*. ACM, 159–162.

[47] Zafeiria Moumoulidou, Andrew McGregor, and Alexandra Meliou. 2021. Diverse Data Selection under Fairness Constraints. In *24th International Conference on Database Theory, ICDT 2021, March 23-26, 2021, Nicosia, Cyprus (LIPIcs, Vol. 186)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 13:1–13:25.

[48] RT O'donnell, Ann E Nicholson, B Han, Kevin B Korb, MJ Alam, and LR Hope. 2006. Incorporating expert elicited structural information in the CaMML causal discovery program. In *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*. 1–16.

[49] Eliana Pastor, Luca De Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*. 1400–1412.

[50] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).

[51] Drago Plecko and Elias Bareinboim. 2023. Causal fairness for outcome control. *Advances in Neural Information Processing Systems* 36 (2023), 47575–47597.

[52] Drago Plecko and Elias Bareinboim. 2024. Causal fairness analysis. *ECAI* (2024).

[53] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[54] 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.

[55] 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.

[56] Donald Bruce Rubin. 1971. *The use of matched sampling and regression adjustment in observational studies*. Ph. D. Dissertation. Harvard University.

[57] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[58] Omer Sagi and Lior Rokach. 2021. Approximating XGBoost with an interpretable decision tree. *Information Sciences* 572 (2021), 522–542.

[59] 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.

[60] Sunita Sarawagi. 2000. User-adaptive exploration of multidimensional data. In *VLDB*. ResearchGate GmbH, 307–316.

[61] Sunita Sarawagi. 2001. User-cognizant multidimensional analysis. *The VLDB Journal* 10 (2001), 224–239.

[62] Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. 1998. Discovery-driven exploration of OLAP data cubes. In *Advances in Database Technology—EDBT'98: 6th International Conference on Extending Database Technology Valencia,*

Spain, March 23–27, 1998 Proceedings 6. Springer, 168–182.

[63] Gayatri Sathe and Sunita Sarawagi. 2001. Intelligent rollups in multidimensional OLAP data. In *VLDB*. 307–316.

[64] Holger Schielzeth. 2010. Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution* 1, 2 (2010), 103–113.

[65] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020).

[66] P. Spirtes et al. 2000. *Causation, prediction, and search.* MIT press.

[67] Hao Sun, Evan Munro, Georgy Kalashnov, Shuyang Du, and Stefan Wager. 2021. Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066* (2021).

[68] Tanmay Surve and Romila Pradhan. 2024. Example-based Explanations for Random Forests using Machine Unlearning. *CoRR* abs/2402.05007 (2024).

[69] Yuchao Tao, Amir Gilad, Ashwin Machanavajjhala, and Sudeepa Roy. 2022. DPXPlain: Privately Explaining Aggregate Query Answers. *Proc. VLDB Endow.* 16, 1 (2022), 113–126.

[70] 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.

[71] U.S. Census Bureau. 2024. American Community Survey (ACS) - Data. Accessed: 2024-01-30.

[72] 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. doi:10.1145/800070.802186

[73] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. Seedb: Efficient data-driven visualization recommendations to support visual analytics. In *VLDB*, Vol. 8. NIH Public Access, 2182.

[74] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.

[75] Clifford H Wagner. 1982. Simpson's paradox in real life. *The American Statistician* 36, 1 (1982), 46–48.

[76] Tong Wang and Cynthia Rudin. 2022. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS Journal on Computing* 34, 3 (2022), 1626–1643.

[77] Yue Wang, Alexandra Meliou, and Gerome Miklau. 2018. RC-Index: Diversifying Answers to Range Queries. *Proc. VLDB Endow.* 11, 7 (2018), 773–786.

[78] 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.

[79] Eugene Wu and Samuel Madden. 2013. Scorpion: Explaining away outliers in aggregate queries. (2013).

[80] Yu Xie, Jennie E Brand, and Ben Jann. 2012. Estimating heterogeneous treatment effects with observational data. *Sociological methodology* 42, 1 (2012), 314–347.

[81] Brit Youngmann, Sihem Amer-Yahia, and Aurélien Personnaz. 2022. Guided Exploration of Data Summaries. *Proc. VLDB Endow.* 15, 9 (2022).

[82] Brit Youngmann, Michael Cafarella, Amir Gilad, and Sudeepa Roy. 2024. Summarized Causal Explanations For Aggregate Views. *Proceedings of the ACM on Management of Data* 2, 1 (2024), 1–27.

[83] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2023. On Explaining Confounding Bias. *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (2023).

[84] Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. 2023. Causal Data Integration. *Proceedings of the VLDB Endowment* 16, 10 (2023), 2659–2665.

[85] 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.

[86] Anna Zeng, Michael Cafarella, Batya Kenig, Markos Markakis, Brit Youngmann, and Babak Salimi. 2025. Causal DAG Summarization. *Proceedings of the VLDB Endowment* 18, 6 (2025), 1933–1947.

[87] Xiaozhong Zhang, Xiaoyu Ge, Panos K Chrysanthis, and Mohamed A Sharaf. 2021. Viewseeker: An interactive view recommendation framework. *Big Data Research* 25 (2021), 100238.

[88] Jiehui Zhou, Linxiao Yang, Xingyu Liu, Xinyue Gu, Liang Sun, and Wei Chen. 2024. CURLS: Causal Rule Learning for Subgroups with Significant Treatment Effect. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 4619–4630. doi:10.1145/3637528.3671951

Received July 2025; revised October 2025; accepted November 2025