ExDis: Causal Explanations for Disparate Trends

Tal Blau
Ben-Gurion University of the Negev
tbl@post.bgu.ac.il

Brit Youngmann
Technion
Haifa
brity@technion.ac.il

Anna Fariha
University of Utah
Salt Lake City
afariha@cs.utah.edu

Yuval Moskovitch
Ben-Gurion University of the Negev
yuvalmos@bgu.ac.il

ABSTRACT

In today's data-driven world, insights collected from the data and trends observed in the data significantly contribute to decision making. However, users are often perplexed by certain surprising data trends, especially the *disparate* ones. For example, upon observing a disparate trend that "men are more likely to have a heart-attack than women", a health-care professional wonders, "is there a certain demographic where the trend is more pronounced or even reversed?", "what factors further exacerbate or alleviate such disparity?". To this end, we introduce ExDis, a system for automatically identifying data regions where an observed Disparity is pronounced (or reversed) and Explaining the associated causes that exacerbate (or alleviate) the disparity. ExDis equips policy-makers to recognize the factors that causally contribute to certain disparities and implement targeted corrective measures.

Link to demo video: http://users.cs.utah.edu/~afariha/exdis.mp4

1 INTRODUCTION

Drawing conclusions from data based on observed trends is common practice. However, when analyzing large, high-dimensional datasets, these trends often require deeper exploration or *explanations* to help the analyst gain a better understanding of the data. For instance, after identifying a notable disparity between two groups in the data, an analyst might be interested in identifying subpopulations in the data where the disparity is more pronounced or even reversed. Uncovering the causal reasons behind the observed disparity can further enhance data understanding.

Example 1. *The Medical Expenditure Panel Survey (MEPS) dataset provides detailed information on healthcare utilization, expenditures, insurance coverage, and demographic characteristics of individuals in the United States. Based on this dataset, in general, males have a lower likelihood (37%) of feeling nervous frequently than non-males (45%). Soha, an analyst analyzing the data, is interested in finding subpopulations where a reverse trend exists, i.e., males have a higher likelihood of feeling nervous than non-males. Indeed, a closer look at the data can reveal such cases. For instance, one such subpopulation is "divorced people with age between 51–63 who have a recommendation to exercise from doctor", 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 21%) but improves the situation for non-males (decreases the likelihood by 14%).*

We aim to *explain* disparities in the average of an outcome for two groups in the data that may overlap, e.g., one group can be the entire data. An explanation should pinpoint *data regions* showing interesting facets of the disparity, either emphasizing or contradicting the observed trend and provide *causal reasons* behind disparities, explaining the causes that exacerbate (or alleviate) the disparity.

A single causal explanation is often insufficient to explain the observed disparity for the entire population. In fact, in different subpopulations, the reasons behind the disparity between the two groups may vary, with some contributing more to the disparity than others. Therefore, we aim to discover *high-utility* subpopulations, for which a strong causal explanation exists for the observed disparity. Preferring explanations consisting of subpopulations with high utility in terms of having a strong causal factor may result in small subpopulations with low support with respect to the entire data, which is undesirable as insights drawn from such a small subpopulation are not statistically significant. To avoid reporting small subpopulations, we consider only subpopulations with *high support* (data coverage) as explanations. Finally, reporting multiple high-utility and high-support subpopulations where the disparity is most pronounced may result in redundancy. E.g., "never married" and "people under the age of 18" may comprise the same individuals, as most people under the age of 18 never married. Therefore, beyond finding high-utility and high-support subpopulations, we aim to minimize the overlap among the reported subpopulations, and, thus, ensure *high diversity* among the explanations.

Manual exploration of the data to discover explanations (data regions and causes) of an observed disparity in a dataset can be complex and tedious, particularly when the dataset is large and high-dimensional. To this end, we introduce ExDis (Explaining Disparate trends), a system that automatically identifies data regions where an observed disparity is pronounced (or reversed) and associates specific factors that causally contribute to the disparity. ExDis, utilizes the solution presented [4], enables users to upload a dataset, specify two groups of interest, set a budget parameter k, and define thresholds for the minimum support of explanations and for pairwise explanation similarity. It then generates a set of k high-utility explanations, each with support above the designated threshold, while ensuring that the pairwise similarity between any two explanations does not exceed the specified threshold.

Demonstration. We will demonstrate the usefulness of ExDis in different scenarios using three real-life datasets where participants will be able to interact with the system to explore the datasets and ask ExDis to explain disparities among selected groups.

Related work. Previous work introduced methods to identify intriguing data subsets for exploration [7, 10] and detect subsets responsible for fairness violations in classifier outcomes [11]. We focus on identifying subpopulations with substantial disparities between two, possibly overlapping, groups and providing causal explanations for these disparities. To achieve this, we build upon the DivExplorer algorithm [7], modifying it to suit our setting, enabling an effective identification of subpopulations.

Recent works [6, 12] used causal inference to explain aggregate query results. CauSumX [12] focuses on causal explanations for group-by-average queries, identifying influential factors (patterns) that drive outcomes. While our goals differ, we adapt CauSumX's treatment mining algorithm to find localized causal explanations for the subpopulations with significant disparity. XInsight [6] explains group disparities in aggregate query results by identifying both causal and non-causal patterns. However, unlike XInsight, which does not support overlapping groups, we provide localized causal insights rather than a single explanation for the entire dataset. We argue that disparities are best understood through precise, subpopulation-specific causal reasoning.

2 TECHNICAL BACKGROUND

We next provide an overview of the model underlying ExDis and refer the readers to [4] for more details.

Background on Causal Inference. We use Pearl's model for *observational causal analysis* [8]. The broad goal of causal inference is to estimate the effect of a *treatment variable* T on an outcome O. One common measure of causal estimate is *Average Treatment Effect* (ATE), defined as the difference in the average outcomes of the treated and control groups:

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

Since ATE is computed over observational data, the treatment and control groups may not be assigned randomly. Therefore, to mitigate the effect of *confounding factors* (i.e., attributes that can affect both the treatment and outcome), we must control for *confounding variables* [8] (Z in Eq. (1)). A sufficient set of confounders can be determined by applying graphical criteria [8], which can be evaluated against a *causal Directed Acyclic Graph (DAG)*. A causal DAG represents potential direct causal relationships between variables in a given dataset [8]. It can be constructed by a domain expert, or by using existing causal discovery algorithms [5].

In ExDis, where the explanation of the disparity between groups may vary among different subpopulations, we are interested in computing the *Conditional Average Treatment Effect* (CATE), which measures the effect of a treatment on an outcome within *a subpopulation of interest.* Given a subpopulation defined by a predicate $B = b$, we compute $CATE(T, O \mid B = b)$ by adding this predicate to the conditioning sets in Eq. 1.

Disparity Explanations. We consider a database D associated with a causal DAG $\mathcal{G}$. A *pattern* [12] is a conjunction of predicates (attribute-value assignments), e.g., {`Gender = Female` $\wedge$ `Race = Asian`}. In this work, we only consider equality or inequality predicates for enhanced interpretability. The two groups of interest, g_1 and g_2, are defined by the patterns ψ_{g_1} and ψ_{g_2}, respectively. Given an outcome attribute O, we aim to discover explanations for an observed disparity in the average value of 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*. Restricting to average is typical for causal explanations [9], as causal effects estimate the expected difference between groups.

We assume the dataset attributes are partitioned into two disjoint sets: *actionable* (mutable) attributes that can be used to define what affects the outcome (e.g., currently smokes, exercises) and *immutable* attributes, which are inherent and cannot be changed (e.g., race, age), that can be used to identify where the disparity is significant. This categorization ensures that treatments consist solely of mutable attributes that can imply corrective measures to reduce the disparity. Given a database D with an outcome variable O and two groups g_1 and g_2, a disparity explanation ϕ is defined as a pair of patterns (ψ_g, ψ_e) where: ψ_g is defined by immutable attributes, describing a subpopulation with significant disparity between g_1 and g_2 in terms of AVG(O), and ψ_e is defined by mutable attributes, indicating a treatment that explains the disparity between g_1 and g_2 within the subpopulation defined by $\psi_g(D)$. 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 2. *Continuing with our example, where g_1 is* `males` *and g_2 is* `non-males`*, an example disparity explanation is: Among "divorced people with age between 51—63 who have a recommendation to exercise from doctor", the treatment "not currently smoking" increases the* `Likelihood of Feeling Nervous` *for* `males`*, while it decreases for* `non-males`*. Here, the subpopulation pattern ψ_g is defined by* `MaritalStatus=Divorced` $\wedge$ `Age=[51 − 63]` $\wedge$ `DoctorRecommendsExercise=True` *and the treatment pattern ψ_e is* `SmokesCurrently=False`.

Problem Formulation. Our goal is to find a bounded-sized set of disparity explanations Φ to identify subpopulations of the data that (1) provide insights into the disparity between g_1 and g_2, and (2) avoid redundancy across different subpopulations to cover different data regions. To this end, we consider the *usefulness* of an explanation and the *diversity* among a set of disparity explanations.

Usefulness. The *usefulness* of a disparity explanation encompasses 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.

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. 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:

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

Disparity explanation selection problem: 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. More formally, given a bound over the number of explanations k, along with thresholds on their support σ and similarity τ the goal is to select a set a disparity explanation set Φ 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.

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, even if the full search space could be materialized, finding the optimal solution remains NP-hard [4]. To this end, ExDis employ a highly scalable heuristic approach based on the algorithm presented in [4].

3 EXDIS OVERVIEW

Figure 1 provides an overview of ExDis, consisting of two components: (1) the Disparity Explanation Configuration Wizard, which assists users in formulating the input, and (2) the Explanations Generator, which implements the explanation mining algorithm [4]. ExDis converts the generated explanations to human-understandable, natural language format utilizing an LLM (GPT 4o Mini).

Disparity Explanation Configuration Wizard. ExDis includes a step-by-step wizard to facilitate the formulation of the problem as defined above. First, the user uploads a dataset (or selects one from the existing datasets in the system). Then, the user specifies the target attribute and the two groups of interest. Note that the groups may overlap. If the user doesn't specify a condition, then the entire dataset is considered as a group. ExDis is designed to explain disparities between two groups. It can be used to investigate the difference between the two groups (or a single group and the entire data) where the difference matters the most, regardless of their direction, or, as demonstrated in Example 1 to find opposite trends (one increasing and the other decreasing) between the two groups. The user can specify whether they are only interested in finding reverse (opposite) trends. Once the user sets the groups, the system displays the dataset color-coded by the two groups of interest. The user also marks the *mutable* and *immutable* attributes (see Figure 2). Next, the user defines the causal DAG. This may be done manually or by uploading a predefined graph from a DOT file or use a default DAG. ExDis also features a causal discovery method [5] that can be used to obtain one. Finally, the user sets the number of desired explanations, k, as well as the diversity and support thresholds. Once the input to the Disparity Explanation is set, the information is transferred to the Explanations Generator.

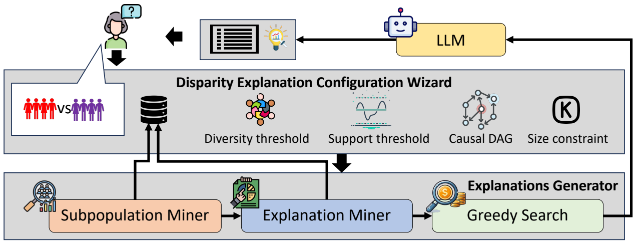


Figure 1: The ExDis architecture. The user provides a database and two groups of interest, ExDis returns causal explanations highlighting the locations and causes of the disparity between the groups.

Explanations Generator. The Explanations Generator operates in 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 to the diversity constraint.

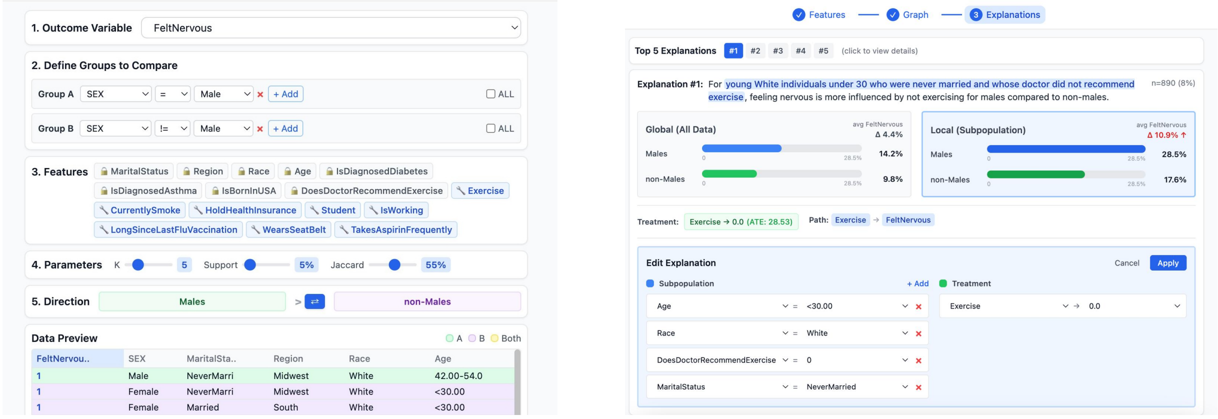
Subpopulation Miner: The number of possible disparity explanations may be exponential in the number of attributes in D. ExDis avoids generating all possible disparity explanations and instead, generates only the promising ones. It first mines subpopulations with significant disparity using the subpopulation-miner module. We adapt DivExplorer [7], which analyzes the divergence of learning models, to measure divergence as the difference in average outcome values. To generate candidate subpopulation patterns, we restrict DivExplorer to immutable attributes.

Explanation Miner: Next, ExDis searches for an explanation pattern for each subpopulation. We adapt the treatment-mining step of the CauSumX [12], which provides causal explanations for aggregate queries, to maximize the difference between the two CATE values. Unlike CauSumX, which targets treatment patterns with high CATE values, our approach estimates disparity scores for candidate patterns. It incorporates parallelization, caching, and sampling-based optimizations to improve interactivity and efficiency.

Greedy search: Given the set of candidate disparity explanations obtained in the previous two steps, our goal is to select a set of k explanations that maximize disparity scores, while satisfying the diversity constraint. Comparing the similarity between every pair of explanations is computationally expensive. To address this, we introduce a clustering step that groups similar subpopulations and assigns a representative explanation to each cluster, thereby reducing redundancy and improving efficiency while maintaining coverage of distinct subpopulations. Specifically, we cluster the candidate explanations using a hierarchical clustering algorithm based on the symmetric difference among the subpopulations. 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.



1. Outcome Variable: FeltNervous

2. Define Groups to Compare

Group A: SEX = Male ☐ ALL

Group B: SEX != Male ☐ ALL

3. Features: MaritalStatus, Region, Race, Age, IsDiagnosedDiabetes, IsDiagnosedAsthma, IsBornInUSA, DoesDoctorRecommendExercise, Exercise, CurrentlySmoke, HoldHealthInsurance, Student, IsWorking, LongSinceLastFluVaccination, WearsSeatBelt, TakesAspirinFrequently

4. Parameters: K 5, Support 5%, Jaccard 55%

5. Direction: Males > non-Males

Data Preview (○ A ○ B ○ Both)

FeltNervou..	SEX	MaritalSta..	Region	Race	Age
1	Male	NeverMarri	Midwest	White	42.00-54.0
1	Female	NeverMarri	Midwest	White	<30.00
1	Female	Married	South	White	<30.00

Top 5 Explanations #1 #2 #3 #4 #5 (click to view details)

Explanation #1: For young White individuals under 30 who were never married and whose doctor did not recommend exercise, feeling nervous is more influenced by not exercising for males compared to non-males. n=890 (8%)

Global (All Data)	avg FeltNervous Δ 4.4%
Males	14.2%
non-Males	9.8%

Local (Subpopulation)	avg FeltNervous Δ 10.9% ↑
Males	28.5%
non-Males	17.6%

Treatment: Exercise → 0.0 (ATE: 28.53) Path: Exercise → FeltNervous

Edit Explanation

Subpopulation:
- Age = <30.00
- Race = White
- DoesDoctorRecommendExercise = 0
- MaritalStatus = NeverMarried

Treatment: Exercise → 0.0

Figure 2: (Left) The ExDis Disparity Explanation Configuration Wizard. In step 2, the user specifies the target attribute, groups of interest, and the *immutable* and *mutable* attributes. (Right) The ExDis Disparity Explanations Screen.

4 DEMONSTRATION

We will demonstrate the usability and expressiveness of ExDis through three interactive demo scenarios, each reflecting a distinct and practically motivated use case. Together, these scenarios illustrate how ExDis enables analysts to efficiently identify subpopulations in which a disparate trend is pronounced and to uncover the factors driving these disparities. Across all scenarios, participants will first be guided through preloaded examples (described below) and will then be able to interactively adjust system parameters to observe their effects on the results. For the demonstration, we will use three real-life datasets: (1) the **MEPS dataset** [3], which contains information on healthcare utilization, expenditures, insurance coverage, and demographic characteristics of individuals in the United States; (2) **Stack Overflow Developer Survey** [1] containing responses from developers worldwide, covering topics such as professional experience, education, technologies used, and employment-related information; and (3) **ACS** [2] a nationwide survey conducted by the U.S. Census Bureau, with demographic, social, economic, and housing data.

1. Investigating a Disparate Trend. The first scenario is demonstrated using the Stack Overflow dataset. We present participants with a surprising global trend: on average, data and business analysts earn more than back-end developers. Using ExDis, participants will interactively identify large subpopulations that contribute most to this disparity and examine how specific factors, such as years of coding experience, affect total compensation differently across the two groups (analysts versus back-end developers). For example, participants would learn that among White individuals aged 25-34, who constitute 35% of the dataset, having 6-8 years of professional coding experience leads to an average total compensation increase of $44K for analysts, compared to only $10K for back-end developers, thereby further widening the compensation gap.

2. Debugging Bias. The second scenario focuses on bias analysis using the ACS dataset. We begin by presenting participants with a “blue-collar bias", showing that individuals in manual-labor occupations have a 13% lower chance of being covered by health insurance than the general population. Using ExDis, participants will identify subpopulations where this bias is amplified and examine factors that affect insurance coverage in opposite ways for the two groups (manual-labor workers versus the overall population). For example, we will show that using ExDis we can learn that among non-native individuals, who comprise 16% of the dataset, the disparity is substantially more pronounced: manual-labor workers have a 65% probability of being insured, compared to an 84% coverage rate across all occupations. Moreover, within this subpopulation, earning between $25K–$55K increases the likelihood of insurance coverage for manual-labor workers by 2%, while decreasing it by 1% for the overall population.

3. Discovering Reverse Trends. The third scenario demonstrates ExDis’s ability to uncover reverse trends (Simpson’s Paradox). Using the MEPS dataset, we will show participants that generally, males have a lower likelihood (37%) of feeling nervous frequently than non-males (45%). Using ExDis, we will find subpopulations where a reverse trend exists, i.e., males have a higher likelihood of feeling nervous. For example, the participants would learn that among divorced individuals aged 51-63 with a doctor’s recommendation to exercise, males exhibit a higher likelihood of feeling nervous than non-males. We will further show how ExDis identifies treatments (e.g., smoking status) that exacerbate the outcome for one group while mitigating it for another.

REFERENCES

[1] Stackoverflow developer survey, 2021. insights.stackoverflow.com/survey/2021.
[2] American community survey. www.census.gov/programs-surveys/acs, 2024.
[3] Medical expenditure panel survey. meps.ahrq.gov/mepsweb, 2024.
[4] T. Blau, B. Youngmann, A. Fariha, and Y. Moskovitch. Causal explanations for disparate trends: Where and why? *Proc. ACM Manag. Data*, 4(1), 2026.
[5] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
[6] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. Xinsight: explainable data analysis through the lens of causality. *Proc. ACM Manag. Data*, 1(2), 2023.
[7] E. Pastor, L. De Alfaro, and E. Baralis. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *SIGMOD*, 2021.
[8] J. Pearl. *Causality*. Cambridge university press, 2009.
[9] B. Salimi, J. Gehrke, and D. Suciu. Bias in olap queries: Detection, explanation, and removal. In *SIGMOD*, 2018.
[10] G. Sathe and S. Sarawagi. Intelligent rollups in multidimensional olap data. In *VLDB*, pages 307–316, 2001.
[11] T. Surve and R. Pradhan. Example-based explanations for random forests using machine unlearning. *CoRR*, abs/2402.05007, 2024.
[12] B. Youngmann, M. Cafarella, A. Gilad, and S. Roy. Summarized causal explanations for aggregate views. *SIGMOD*, 2(1), 2024.