Causal Explanations for Disparate Trends: Where and Why?

Tal Blau[1], Brit Youngmann[2], Anna Fariha[3], Yuval Moskovitch[1]

[1] Ben-Gurion University of the Negev
[2] Technion - Israel Institute of Technology
[3] University of Utah

① Motivation

- **Data Trends Drive High-Stakes Decisions**
  - Directly impacts policy design and societal equity.
- **The Core Challenges: Where and Why?**
  - **Where:** Disparities are often hidden within specific subpopulations.
  - **Why:** Surface correlations fail to reveal the true underlying causes.
- **Manual Search is Infeasible**
  - Finding localized gaps in high-dimensional data is like finding a needle in a haystack.
  - Crucial trends can easily be missed or completely reversed.

The Goal

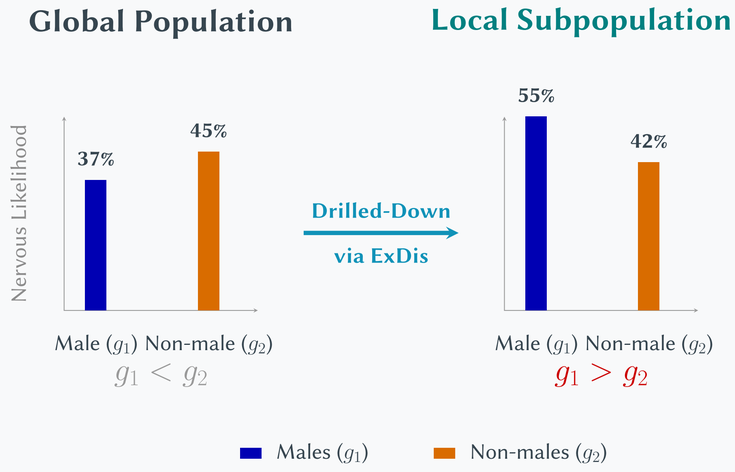
- Automate the discovery of **causal, actionable, and diverse** explanations across data regions.

② Example

- **MEPS Mental Health Study:** Looking at severe nervousness across different subpopulations.
- **Overall Averages:** Non-males (g_2) report higher rates (45%) than Males (g_1) (37%).
  - Basic policies might just target g_2 based on this broad trend. *But does this match the actual risk for specific people?*

Key insight

- **The Problem with Averages:** Looking only at the big picture hides real differences in people's lives by treating everyone the same.



Simpson's Paradox Uncovered: Non-males show higher nervousness globally ($g_1 < g_2$), but the trend completely reverses ($g_1 > g_2$) within the low-income, uninsured subpopulation isolated by ExDis.

- `ExDis` finds hidden groups where the overall trend flips completely.
- In this specific group, the numbers reverse: Males (g_1) jump to 55%, while Non-males (g_2) drop to 42%.
- To truly understand differences between groups, we must **find these hidden pockets** where the big trends break down.

③ Core Definitions & Problem Formulation

1. Base Elements :

- **Outcome Variable ($O \in A$):** The target aggregate metric monitored for disparities. Example: Total Compensation (TC).
- **Pattern (ψ):** A conjunction of equality predicates that isolates a specific subpopulation $\psi(D) \subseteq D$. Example: $\psi = (Ethnicity = Asian \wedge Role = Dataanalyst)$.
- **Groups of Interest (g_1, g_2):** disjoint patterns (ψ_{g_1}, ψ_{g_2}) whose aggregate outcomes are being contrasted. Example: $g_1 = DataAnalysts$ vs. $g_2 = Back-EndDevelopers$.

2. Attribute Partitioning ($A \setminus \{O\} = I \cup M$) two disjoint sets:

- **Immutable Attributes (I):** define **where** ψ_g. Example: Age.
- **Mutable Attributes (M):** define **why** ψ_e. Example: $EducationalLevel$.

3. Disparity Explanation (ϕ) A pair of patterns (ψ_g, ψ_e) where ψ_g over I — the subpopulation where g1 and g2 diverge. ψ_e over M — the treatment that explains the gap inside.

4. Disparity Score (Δ) :

$$\Delta(\phi) = \frac{|CATE_{G_D}(\psi_e, O \mid \psi_g \wedge \psi_{g_1}) - CATE_{G_D}(\psi_e, O \mid \psi_g \wedge \psi_{g_2})|}{\max |O|}$$

5. Support :

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

6. Diversity (Jaccard Similarity) :

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

Problem: Disparity Explanation Selection

Given a database instance D, a causal model G_D, an outcome O, two groups of interest g_1 and g_2, a fixed budget $k \in N^+$, a support threshold σ, and a similarity threshold τ, select an optimal explanation set $\Phi \subseteq \{\phi_1, \ldots, \phi_l\}$ such that:

1. **Size Constraint:** The size of the explanation set is bounded by the budget ($|\Phi| \leq k$).
2. **Support Constraint:** Each explanation covers a representative fraction of data ($\forall \phi_i \in \Phi : support(\phi_i) \geq \sigma$).
3. **Diversity Constraint:** Pairwise subpopulation overlap is strictly bounded to eliminate redundancy ($\forall \phi_i, \phi_j \in \Phi : SIM(\phi_i, \phi_j) \leq \tau$).
4. **Objective Function:** Maximize the collective causal explanation strength:

$$Maximize \Delta(\Phi) = \sum_{\phi \in \Phi} \Delta(\phi)$$

④ Algorithm

Step 1: Subpopulation Miner (Data Slicing)

- Restricts the search space strictly to immutable attributes (I).
- Performs a single-pass from the *Apriori* which automatically prunes and discards any region failing the support threshold ($support < \sigma$).

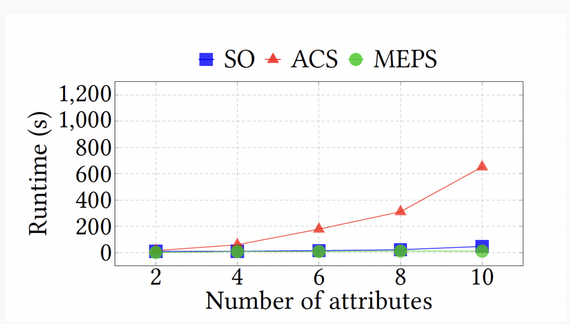
Step 2: Explanation Miner (Causal Resolution)

- Heuristically traverses a treatment attribute lattice over mutable attributes (M).
- Overcomes the primary system computational bottleneck through target optimizations: (Parallelization, Caching, Sampling).

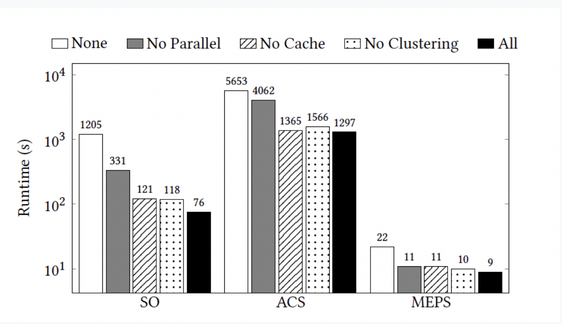
Step 3: Fast Greedy Search Selector (Diverse Maximization)

- Hierarchical Clusters overlapping subpopulations.
- Iteratively extracts the top-k explanations from each cluster that maximize combined disparity (Δ) while strictly respecting the pairwise similarity threshold (τ).

⑤ Experimental Results



Scalability vs. Dataset Size: ExDis achieves **linear** scaling, bypassing standard search bottlenecks.



Ablation Study: every optimization is absolutely critical for extracting non-redundant, high-quality explanations.

Feel free to reach out with any questions!
tbl@post.bgu.ac.il, brity@technion.ac.il,
afariha@cs.utah.edu, yuvalmos@bgu.ac.il

See the paper link here: