CSE 5249 Paper Presentation

# Summarized Causal Explanations For Aggregate Views

By: Youngchun Kim, Michael Cafarella, Amir Gilad, and Sudeepa Roy

Aditya Vadlamani & Ram Sai Ganesh

# Motivation

- Data analysts from different domains interact with databases via SQL queries to gain insights. The results of the queries can get very large a method of getting **insightful** and **automatic** explanations is desired
    - If ***causal explanations*** can be derived, then sounds data-driven decision-making is possible

- A simple, but important, class of SQL queries are the ***group-by-and-average*** queries. This class of queries provide an aggregate view of the database by partitioning the population and returning how the average varies across the sub-populations.

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

# Example


- Consider this group-by SQL Query
- Huge variation in average TC across countries Why?
  What is causing this discrepancy?
- Cannot find a pertinent explanation manually,
  - Dataset size
  - # of attributes
- Other sources of aggregated explanations –
  - Viz tools like Tableau & PowerBI
  - *Provenance* of answers
  - Query result explanations
  - **no differentiation** between **causal & non-causal** reasons

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


Table 1. A subset of the Stack Overflow dataset.

| ID | Country | Continent | Gender | Age | Role | Education | Major | Salary |
|---|---|---|---|---|---|---|---|---|
| 1 | US | N. America | Male | 26 | Data Scientist | PhD | C.S | 180k |
| 2 | US | N. America | Non-binary | 32 | QA developer | B.Sc. | Mech. Eng. | 83k |
| 3 | India | Asia | Male | 29 | C-suite executive | B.Sc. | C.S | 24k |
| 4 | India | Asia | Female | 25 | Back-end developer | M.S. | Math. | 7.5k |
| 5 | China | Asia | Male | 21 | Back-end developer | B.Sc. | C.S | 19k |

# Related Work

- **Query Result Explanation:**
  - Using data provenance
  - Non-causal interventions
  - Detection of confounders to explain Simpson's Paradox
- **Causal Inference:**
  - Heterogeneous treatment effects
  - **Assumption** about known treatment/outcome variables.
  - CauSumX **mines treatment pairs** to **find high causal effects** for sets of groups.
- **Interpretable Prediction Models:**
  - Rule mining for predictive rules
  - *IDS, FRL rule gen* – comparison used later.
- **Data Summarization:**
  - Condensing input into interpretable subsets
  - *Explanation tables* – baseline used later.

| Related Work | Causal | Entire View | Supports Groups |
|---|---|---|---|
| Query Result Explanation | ✗ | ✗ | ✗ |
| | ✗ | ✗ | ✓ |
| | ✓ | ✗ | ✓ |
| | ✓ | ✗ | ✗ |
| Interpretable Prediction Models | ✗ | ✓ | ✗ |
| Data Summarization | ✗ | ✓ | ✗ |
| | ✗ | ✓ | ✗ |
| CauSumX | ✓ | ✓ | ✓ |

**CauSumX:** Summarized & causal explanation for the **entire aggregated view**, accounting for inter-group variation*

* *more on this later: we condition on sensitive groups for this reason.*

# Problem Statement and Contributions

- We want to have a method which can provide ***causal explanations*** for the ***entire aggregate view*** while ***accounting for variations between groups***

- This paper presents a novel framework, *CauSumX*, which generates causal explanations for an entire aggregate view from a query. This is done in a three-step algorithm which:
  - Mines for frequent grouping patterns using the Apriori algorithm (from the 1990s, not 1900s 😅)
  - Mines for promising treatment patterns for each grouping pattern
  - Solves an ILP optimization problem to maximize causal explainability subject to constraints

# Background: Causal Inference

**Goal:** Estimate the effect of a *treatment*, **T**, on an *outcome*, **Y.**

The gold standard for causal inference are *randomized control experiments.* In such settings we can use the following metrics:

- *Average Treatment Effect: $$ATE(T, Y) = \mathbb{E}(Y | do(T = 1)) - \mathbb{E}(Y | do(T = 0))$$*
- *Conditional ATE: $$CATE(T, Y | B=b) = \mathbb{E}(Y | do(T = 1), B = b) - \mathbb{E}(Y | do(T = 0), B = b)$$*

Randomization can help mitigate confounding variables. However, randomized control experiments can be impractical. An alternative method would be *observational causal analysis.*

# Background: Causal Inference

Observational Causal Analysis allows for causal inference under some additional assumptions:

1. Unconfoundedness: $Y \perp\!\!\!\perp T | Z = z$
2. Overlap: $0 < \mathbb{P}(T = 1 | Z = z) < 1$

The metrics from before can be rewritten as

- *Average Treatment Effect: $$ATE(T, Y) = \mathbb{E}_Z[\mathbb{E}(Y | T = 1, Z=z) - \mathbb{E}(Y | T = 0, Z=z)]$$*
- *Conditional ATE: $$CATE(T, Y | B = b) = \mathbb{E}_Z[\mathbb{E}(Y | T = 1, Z=z, B=b) - \mathbb{E}(Y | T = 0, Z=z, B=b)]$$*

Given a Causal DAG, we can find $$Z$$ using conditions like *d-separation* and *backdoor-criteria.*

# Methodology: Databases and Queries

- Let $$D$$ be a single-relation database over a schema $$\mathbb{A} = (A_1, \dots, A_s)$$. A database instance is a set of tuples $$t = (a_1, \dots, a_s)$$ where each $$a_i\in dom(A_i)$$.

- Let $$\mathcal{A}_{gb}\subseteq \mathbb{A}$$ be the *group attributes* and $$A_{avg}\in \mathbb{A}$$ be the *average attribute.*

- The result of the query for a particular database instance is $$Q(D)$$, which has size $$m$$.

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

# Methodology: Explanation Patterns

- A *predicate* is of the form $$\varphi = A_i op a_i$$ where $$op\in\{=, >, <, \leq, \geq\}$$. *Patterns* are conjunctive predicates (e.g. $$\varphi_1 \wedge\cdots\wede \varphi_k$$).

- **Grouping pattern ($$\mathcal{P}_g$$):** Captures a *well-defined* subset of groups in $$Q(D)$$. Well-defined meaning there is a functional dependency $$\mathcal{A}_{gb}\to W$$ for all $$W\in \mathcal{P}_g$$.

- **Treatment pattern ($$\mathcal{P}_t$$):** Partitions $$D$$ into a *treated* and *control* group based on pattern evaluation. Used to determine causal effects on $$A_{avg}$$

An **explanation pattern** is a pair **$$(\mathcal{P}_g, \mathcal{P}_t)$$**

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

# Methodology: Explainability

# References

[1] Brit Youngmann, Michael Cafarella, Amir Gilad, and Sudeepa Roy. 2024. *Summarized Causal Explanations For Aggregate Views.* Proc. ACM Manag. Data 2, 1 (SIGMOD), Article 71 (February 2024), 27 pages. https://doi.org/10.1145/3639328