Distributed Summarized Causal Explanations for Fairness Insights

Aditya Vadlamani and Ram Sai Ganesh

The Ohio State University

1 Introduction & Related Work

With the increase in database interactions by data analysts and practitioners, the need for insightful and automated explanations of the data increases. One of the most popular class of SQL queries are those with *group-by and average* queries. This class of queries provides information about how the average of a quantity of interest may vary across different sub-populations represented in the database in the form of aggregate views of the data. This sort of data is popular due to how easy it is to visually represent the aggregate view using bar charts that can be interpreted by folks on various backgrounds. However, it is difficult from these aggregate views and visualizations to determine the *causal reasons* for the high/low average values for different sub-populations. These causal reasons can then enable effective and data-driven decision-making. On this end, researchers have come up with methods to generate such causal explanations from these aggregate views using ideas from Causal Inference and Analysis [4, 7, 12, 13].

In addition to the increase in database interactions, the size of databases has also grown in recent years. This is especially true for databases hosted on cloud platforms (e.g. AWS, Google Cloud, and Azure). Due to the sheer size, it is common practice to partition the database across multiple nodes or servers, to enable distribution of storage and computation costs. It may also be that the database is partitioned to allow for privacy considerations. For example, the EU has different privacy laws [11] than the United States.

With the full database being partitioned, generating a full aggregate view becomes more expensive (or even impossible if privacy is a concern), and generating causal reason becomes even more challenging. As a result, the algorithms running on data at this scale must also run in a distributed setting.

The objective of this project was to extend existing work in Causal Fairness Summarization to applications in a distributed setting, allowing for efficient and decentralized computation of causal explainability from the data. This work focuses on extending CauSumX [12], a recent work in causal summarizations of *group-by and average* queries, to the distributed setting

2 Background

CauSumX [12] is a recent algorithm enabling the generation of causal explanations for aggregate data views. Building on prior work extending Pearl's causal model for relational databases [8], CauSumX provides fine-grained causal explanations for individual groups in an entire aggregate view, providing important treatments affecting outcomes.

We define some terms with context obtained from the StackOverflow [10] dataset:

- The broad goal of **causal inference** is to estimate the effect of a *treatment variable* T on an outcome variable Y. A popular measure of the efficacy of a causal explanation is the **Average Treatment Effect** (ATE). In a randomized experiment, the ATE is the difference in the average outcomes of the treated and control groups [5] given as

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

  Note that we assume that the treatment assigned to one unit does not affect the outcome of another unit. This assumption is called the Stable Unit Treatment Value Assumption (SUTVA) [6].

- In this particular scenario, involving generating explanations for SQL group-by-average queries (where the treatment with maximum effect may vary among different tuples in the query answer), we are interested in computing the **Conditional Average Treatment Effect (CATE)**, which measures the effect of a treatment on an outcome on a *subset of input units*.

  Given a subset of units defined by a vector of attributes B and their values b, we can compute $CATE(T,\ Y \mid B = b)$ as:

  $$CATE(T,\ Y|B = b) = \mathbb{E}[Y \mid \text{do}(T = 1),\ B = b] - \mathbb{E}[Y \mid \text{do}(T = 0),\ B = b]$$

- A **grouping pattern** ($\mathcal{P}_g$) illustrates a property or predicate on the groups for which this insight holds. For instance, causal explanations may be generated that are accurate at predicting patterns in *countries with high GDPs*.

- A **positive treatment pattern** ($\mathcal{P}_t^+$) is a predicate on the individuals from the above groups with a high positive treatment effect. For instance, considering the grouping pattern of high-GDP countries, we may observe a positive causal impact on income for *individuals with a Master's Degree*.

- A **negative treatment pattern** ($\mathcal{P}_t^-$) , similarly, is a predicate on individuals from a grouping pattern with a high *negative* treatment effect. In the StackOverflow dataset, we observe that *being over the age of 55 with a bachelor's degree* has the greatest adverse impact on annual impact.

- An **explanation pattern** is the tuple $(\mathcal{P}_g, \mathcal{P}_t^+, \mathcal{P}_t^-)$ or simply $(\mathcal{P}_g, \mathcal{P}_t)$
- The **explainability** of an explanation pattern is $CATE_{M_{\mathbb{A}}}(\mathcal{P}_t, A_{avg}|\mathcal{P}_g)$, where $M_{\mathbb{A}}$ is the underlying causal model and A_{avg} is the attribute which is being averaged.

With these terms, the problem setup and algorithm for CauSumX is as follows:

Problem: Given a database D, a causal background knowledge DAG, a group-by-average query Q and parameters k and θ, CauSumX generates a set of k *explanation patterns* (predicates) that explain at least θ fraction of the groups in *Q(D)*.

Algorithm:

1. Mine for frequent grouping patterns using Apriori algorithm [9]
2. Greedily mine for treatment patterns using a lattice approach
3. Solve an Integer Linear Program to maximize explainability.

We note an important gap in CauSumX's original formulation in that it requires the entire dataset to be available in the same computation node in order to generate accurate causal explanations. This results in CauSumX having limited real-world applicability for any size of dataset significantly larger than the paper's examples of the StackOverflow [10] and Adult [1] datasets.

3 Methodology

In this section, we present a distributed variation of CauSumX that can work on partitioned data and analyze the effect of different algorithms for mining grouping patterns.

3.1 Distributing CauSumX

To adapt CauSumX for the distributed setting, given a partitioned database, we independently run CauSumX on each partition to generate explanations for that partition. The choice of having independent executions comes from allowing the assumption of privacy between partitions. Once we have the explanations from each partition, we can then further aggregate the results. One method of doing so, would be to sum the *explainability* of the same explanations and then choose those higher total explainability to present to the end user. The intuition behind using the sum is that explainability is a conditional expectation which satisfies linearity. The aggregation of explanation patterns is left as future work.

3.2 Grouping Pattern Mining

Another part of CauSumX that we explored was its use of the Apriori Algorithm [9] to mine for frequent grouping patterns in the dataset. Due to the underlying principle of the Apriori algorithm, which is that itemsets (or grouping patterns) are frequent only if every subset of the itemset is also "frequent". This principle makes it difficult to select longer itemsets or grouping patterns which may hold a more diverse causal explanations. To this end, we integrated another mining algorithm, Localized Approximate Miner (LAM) [2], to mine for the initial grouping patterns. Unlike traditional mining methods which does a top-down search of the search-lattice, LAM looks at different parts of the lattice for its mining.

4 Experiments & Results

Equipped with the adaptations made to the CauSumX framework, we ran several experiments regarding the scalability and the effect of mining algorithms.

4.1 Datasets

We examine some of the datasets used in the original CauSumX [12] work:

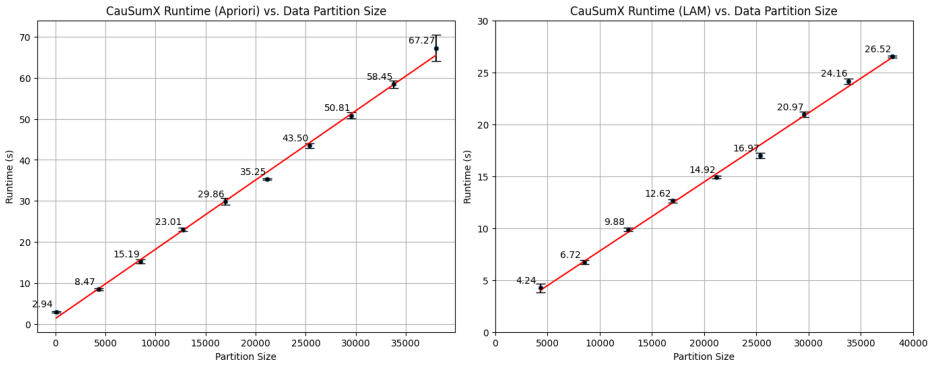
StackOverflow [10]: This dataset contains results from StackOverflow's Annual Developer Survey where users answer questions regarding themselves and their jobs. The causal DAG is from [14].
Adult [3]: This dataset contains demographic information of individuals along with their education, occupation, annual income, etc.
German [3]: This dataset contains details of bank account holders, including demographic and financial information, along with their credit risk.

4.2 Scalability

Figure 1 shows how CauSumX scales with varying partition sizes when using either the Apriori [9] or LAM [2] mining algorithms. In both cases linear scaling is observed and the LAM version runs faster. The latter is likely due to implementation since Apriori [9] is written in Python and LAM [2] in Java.

4.3 Group Pattern Mining

We attempted to improve upon the results from the paper by utilizing an improved Group Pattern Mining Algorithm. We investigated the use of the Locally Approximate Miner (LAM) [2]. Table 1 presents the explainability and coverage of CauSumX on the StackOverflow dataset. Because LAM can mine for longer grouping patterns, this makes it harder to achieve a higher coverage since the grouping pattern will refer to a more specific sub-population. In fact, we observe



(a) Apriori (b) LAM

Figure 1: CauSumX Runtime vs Partition Size for Different Group Pattern Mining Methods

	Apriori	LAM
Total Explainability (↑)	263246.447	334718.403
Coverage (↑)	1.0	0.45

Table 1: CauSumX performance with Apriori vs LAM on the StackOverflow 2021 Dataset

in other datasets CauSumX couldn't produce an explanation pattern (ILP had no solution) or one with zero explainability. It may be that the datasets chosen are too small for the benefits of mining long-term relationships to be apparent.

4.4 Grouping by Data Attributes

While partitioning datasets at random is useful to estimate the runtime of CauSumX on average, we can gain useful insights about the applicability in a data-distributed setting by instead choosing to subset data based on existing features.

Continuing our consideration of the StackOverflow dataset, we attempt to extract useful insights about income disparities *across Continents* in the data. Figure 2 shows that in grouping the StackOverflow dataset, we observe the responses are majorly skewed towards primarily English-speaking countries – North America and Europe; we would expect to have lower success with explainability in such areas with a large population (and consequently higher GINI scores). We observe similar explanation patterns in solutions across multiple continents. Specifically, the following insights are consistent across continents:

- In countries with predominantly Caucasian populations and smaller proportions of other races (Europe and North America), there is a strong

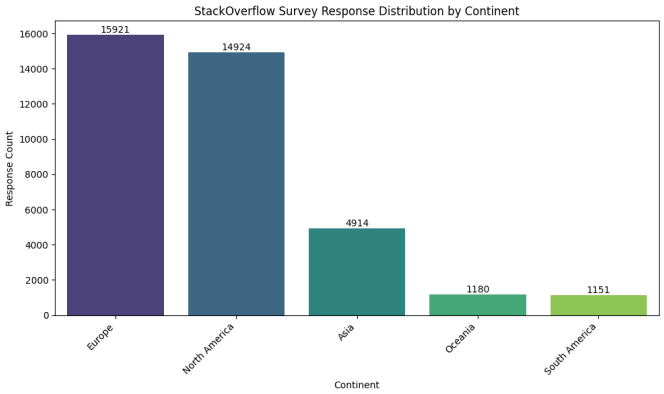


Figure 2: Distribution of StackOverflow 2021 survey responses by Continent

positive causal relationship between being White and earning a higher income.

- In four of five continents (other than Oceania), there is an explicit positive causal relationship between higher levels of education and reported income level. The exact treatment varies – Asia: Education="Master's Degree"; Europe: Education = "Bachelor's Degree".
- In four of five continents, we see a peak in income levels for people in the 35 to 44 year-old age bracket. This presents either as a strong positive treatment effect for people in that demographic, or as strong negative effects seen in people *outside* that range (ie. Europe: Age=35 - 44 years old, strong positive relationship; Oceania, Age=18-24 years old, strong negative relationship).

These results both corroborate existing beliefs of the prevalence of ageism and racism in hiring in the industry, as well as validate our methodology of extracting credible and useful fairness insights in a distributed setting.

5 Conclusion & Future Work

In this work, we present an adaptation of CauSumX that can operate on partitioned databases and allow for other mining methods to be used within the algorithm. We show that the distributed version allows for a linear speed-up with respect to the number (and size) of partitions.

Future work entails investigating methods of aggregating explanation patterns generated to preserve privacy across partitions and also running experiments on larger more complicated datasets and queries where we may see a more discernible difference in performance between LAM [2], Apriori [9], or other mining methods.

References

[1] Becker, B., and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[2] Buehrer, G., de Oliveira, R. L., Fuhry, D., and Parthasarathy, S. Towards a parameter-free and parallel itemset mining algorithm in linearithmic time. In *2015 IEEE 31st International Conference on Data Engineering* (2015), pp. 1071–1082.

[3] Chiappa, S. Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence 33*, 01 (Jul. 2019), 7801–7808.

[4] Ma, P., Ding, R., Wang, S., Han, S., and Zhang, D. Xinsight: explainable data analysis through the lens of causality, 2023.

[5] Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys 3*, none (2009), 96 – 146.

[6] Rubin, D. B. Causal inference using potential outcomes. *Journal of the American Statistical Association 100*, 469 (2005), 322–331.

[7] Salimi, B., Gehrke, J., and Suciu, D. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data* (New York, NY, USA, 2018), SIGMOD '18, Association for Computing Machinery, p. 1021–1035.

[8] Salimi, B., Parikh, H., Kayali, M., Roy, S., Getoor, L., and Suciu, D. Causal relational learning. *CoRR abs/2004.03644* (2020).

[9] Srikant, R., and Naughton, J. F. *Fast algorithms for mining association rules and sequential patterns*. PhD thesis, 1996. AAI9708697.

[10] Stack Overflow. Stack Overflow Developer Survey 2021, 2021.

[11] Voigt, P., and Bussche, A. v. d. *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[12] Youngmann, B., Cafarella, M., Gilad, A., and Roy, S. Summarized causal explanations for aggregate views. *Proc. ACM Manag. Data 2*, 1 (mar 2024).

[13] Youngmann, B., Cafarella, M., Moskovitch, Y., and Salimi, B. On explaining confounding bias. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (Los Alamitos, CA, USA, apr 2023), IEEE Computer Society, pp. 1846–1859.

[14] Youngmann, B., Cafarella, M., Salimi, B., and Zeng, A. Causal data integration. *Proc. VLDB Endow. 16*, 10 (jun 2023), 2659–2665.