LABORATORY RESEARCH PROJECT (PRL)

An Empirical Analysis on the Effect of the Outcome Variable Binarization and Bias on Causal Discovery

Sep-Dec 2024

Kenza Chaabouni [1]
Under the supervision of Dr. Karima Makhlouf [2]

[1]Email: kenza.chaabouni@polytechnique.com
[2]Email: karima.makhlouf@inria.fr

ÉCOLE POLYTECHNIQUE

IP PARIS

Abstract

This laboratory research project investigates the influence of outcome variable binarization on causal graph structures within the domain of causal discovery for fairness. It contributes to the work of the Comète team at Inria on *Causal Discovery on Biased Data* [10], which develops a framework for generating synthetic datasets and analyzes the effect of varying bias and outcome binarization on causal discovery. The study spans both synthetic and real-world datasets to assess how the distribution of outcome variables impacts causal discovery results. The findings demonstrate that outcome binarization affects the identification of causal relationships, particularly between the sensitive attribute and the outcome, while revealing variations in causal structures under different bias levels. These results are especially critical in the context of fairness, emphasizing the importance of thoughtful data preprocessing choices to ensure reliable and equitable machine learning applications.

Contents

1 Introduction

As Machine Learning (ML) is increasingly used in decision-making processes that deeply impact people's lives, such as job hiring, disease diagnosis and loan granting [6]. It is crucial to account for the social and ethical implications of ML-based decisions to ensure they do not discriminate against individuals or minorities. Fairness serves as a cornerstone for the safe and equitable application of ML systems, as it enables an equitable deployment of these ML systems [8]. Several notions of fairness have been defined and analyzed [9]. While most rely on correlation-based definitions of discrimination, newer concepts are based on causality, supporting the increasingly recognized opinion that understanding causal relationships is essential to tackle fairness issues effectively [9]. These causality-based approaches uncover deeper structural fairness concerns by identifying the causal relationships between sensitive attributes (e.g., race, gender) and decision outcomes (e.g., hiring decisions, loan approvals)[2]. Causal fairness methods offer significant advantages by effectively accounting for confounding and mediation effects, limitations often overlooked by conventional fairness metrics [2]. This laboratory research project contributes to the Comète team's work on *Causal Discovery on Biased Data* [10] which makes three significant contributions:

- A framework for generating synthetic datasets with a specified ground truth causal graph and a bias level.
- An empirical study of the impact of bias levels on causal discovery outputs.
- An empirical analysis of how the binarization threshold of the outcome variable affects the discovered causal graph.

The team introduces a novel mechanism for generating synthetic datasets, using input causal graphs (ground truth causal graphs) and a bias level. The approach incorporates causal structures (e.g., mediators, confounders, colliders) to simulate realistic bias propagation scenarios. By controlling bias levels through adjustable parameters, the framework models varying degrees of discrimination against privileged or unprivileged groups. These synthetic datasets are employed to examine two aspects:

- The influence of varying bias levels on the causal graph output by Causal Discovery algorithms.
- The sensitivity of causal graphs to different binarization thresholds of the outcome variable.

The study predominantly uses the Peter-Clark (PC) algorithm [13] for causal discovery and evaluates fairness using metrics such as Statistical Disparity (SD) and Disparate Impact (DI), two well-established fairness notions [5]. The research extends to benchmark datasets, including the *Adult* [4], *Compas* [1], and *Communities*[14] datasets, to validate the findings in real-world contexts. The findings highlight that as bias levels increase, causal discovery algorithms identify more causal relationships. However, when the data is unbiased, these algorithms often fail to identify critical causal edges, such as the direct edge between the sensitive attribute and the outcome. In addition, the analysis of the binarization threshold sensitivity reveals that the outcome distribution impacts significantly the output of Causal Discovery algorithms. The choice of threshold affects the discovery of key causal edges, with direct edges between the sensitive attribute and the outcome appearing or disappearing inconsistently based on the outcome distribution.

This laboratory research project contributes to Causal Discovery on Biased Data [10] by systematically analyzing the effects of outcome variable binarization and bias on causal discovery, using both synthetic and real-world datasets. The key contributions of this laboratory research project are as follows:

- Analysis of the Effect of the Outcome Variable Binarization: The project systematically studies how the binarization threshold of the outcome variable influences the causal graph generated by the PC algorithm, using both synthetic and real-world datasets.
- Evaluation of the Impact of Bias: The project explores how varying bias levels affect causal discovery, using synthetic data generated with different biases and amplifying bias in real-world data. It employs synthetic data with controlled bias levels and introduces bias into real-world datasets.

2 Preliminaries

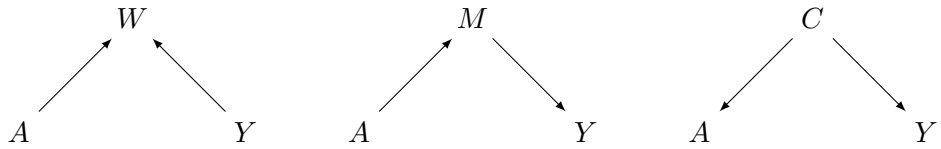
2.1 Notation

Consider an n-dimensional dataset $D = \{\mathbf{X}, Y\}$, then $\mathbf{X} = \{X_1, \ldots, X_{n-1}\}$ are the feature variables and Y the outcome variable, Y can either be continuous (e.g., price, income) or binary (e.g., hiring, granting a loan). In the latter case, $Y = 1$ represents a positive outcome (e.g., hiring, granting a loan), while $Y = 0$ represents a negative outcome (e.g., firing, refusing a loan). We denote by $A \in \mathbf{X}$ the sensitive attribute (e.g., the gender or race of an individual). We consider cases where A is binary, specifically, $A = 0$ designates the unprivileged group, while $A = 1$ represents the privileged group [3].

A directed graph is denoted $G = \{\mathbf{V}, \mathbf{E}\}$, where $\mathbf{V} = \{V_1, \ldots, V_n\}$ are the vertices, also called nodes, and $\mathbf{E} \subseteq V \times V$ is the set of directed edges between the vertices. An edge $E_{ij} \in \mathbf{E}$ exists if and only if there is direct connection $V_i \to V_j$, with $1 \leq i, j \leq n$ and $i \neq j$.

2.2 Causal Structures

Given a causal graph G where $A \in \mathbf{V}$ is the sensitive attribute and $Y \in \mathbf{V}$ is the outcome, we consider three causal structures described in Figure 1, Collider, Mediator and Confounder, denoted by W, M and C respectively.



(a) (b) (c)

Figure 1: Three basic causal structures: collider (a), mediator (b), and confounder (c).

A collider W is a common effect of A and Y. The causal structure is $A \to W \leftarrow Y$. Taking A the ethnicity and Y the salary, the collider W might be the choice of workplace. Ethnicity can influence the choice of workplace and the salary influences the choice of workplace.

A mediator M is a direct effect of A and a direct cause of Y. The causal structure is $A \to M \to Y$. For example, taking A as the gender of an individual, and Y the hiring decision, the M could be

[3] A privileged group refers to individuals who are historically or socially advantaged and tend to receive favorable outcomes, whereas an unprivileged group comprises those who face systemic disadvantages or biases, often leading to less favorable outcomes.

the education level of the individual. In this case, the gender influences the education level which in return influences the hiring decision.

A confounder C is a common cause between A and Y. The causal structure is $A \leftarrow C \rightarrow Y$. Taking A the region of residence of an individual and Y the access to healthcare services, the confounder C might be the availability of infrastructure. The availability of infrastructure (e.g., transportation, hospitals) influences both the region of residence (e.g., rural or urban) and access to healthcare.

2.3 Causal Discovery Algorithm

In this laboratory research project, we use the Peter-Clark (PC)[13] causal discovery algorithm, since it is well adapted for mixed data. The PC algorithm is a constraint-based algorithm that relies on conditional independence tests to identify causal relationships in the causal graph. This algorithm works in two steps:

1. Building the skeleton causal graph with undirected edges
2. Orienting the undirected edges to build a Partially Directed Acyclic Graph (PDAG)

Algorithm 1 PC Algorithm - Step 1: Build the Skeleton Graph

Require: Dataset D with variables V, significance level α
Ensure: Undirected graph G

Initialize G as a complete undirected graph over V.
for all pairs of variables $(X, Y) \in V$ **do**
  Initialize the conditioning set $\mathcal{Z} = \emptyset$.
  while $|\mathcal{Z}| \leq$ |Neighbors of $X \cup Y$| **do**
    Perform conditional independence test for X and Y given $\mathcal{Z}$ using Fisher's Z-test:
    Compute partial correlation $\rho_{XY|\mathcal{Z}}$.
    Convert $\rho_{XY|\mathcal{Z}}$ to a Z-score and calculate the p-value.
    if $p > \alpha$ **then**
      Remove edge (X, Y) from G.
      Add $\mathcal{Z}$ to the separating set for (X, Y).
    end if
    Update $\mathcal{Z}$ and repeat.
  end while
end for
return Undirected graph G

During the first step (Algorithm 1), the algorithm takes as input the data D with a set of variables V and a significance level α and outputs an undirected graph. The algorithm starts by connecting the graph fully, then iteratively removing edges between variables that are conditionally independent. For each edge $X \rightarrow Y$ and each variable $Z \in V$ neighboring both $X \in V$ and $Y \in V$, the algorithm checks if the variables X and Y are independent when conditioned on Z. Using the Fisher's Z-test, conditional independence is computed based on the partial correlation $\rho_{XY|Z}$ between variables X and Y given conditioning on Z. This partial correlation is then converted to a Z-score, which is used to compute a p-value. To test for independence, the p-value is compared to the threshold α, if $p > \alpha$ then the the two variables are considered independent and the edge is dropped.

Note that the higher the partial correlation, the higher the Z-score, and consequently, the lower the p-value. If the partial correlation is high enough, the algorithm retains a direct edge between X and Y.

Algorithm 2 PC Algorithm - Step 2: Orient Edges to Create a PDAG

Require: Undirected graph G and separating sets from Step 1
Ensure: Partially Directed Acyclic Graph (PDAG)

Identify all unshielded triples (X, C, Y) where:
  X and C are connected, C and Y are connected, but X and Y are not connected.
for all unshielded triples (X, C, Y) **do**
  if $C \notin$ separating set of (X, Y) **then**
    Orient the triple as $X \rightarrow C \leftarrow Y$.
  end if
end for
while further edges can be oriented **do**
  if $X \rightarrow C$ and $C - Y$ **then**
    Orient $C - Y$ as $C \rightarrow Y$.
  end if
  if $X \rightarrow C \rightarrow Y$ and $X - Y$ **then**
    Orient $X - Y$ as $X \rightarrow Y$.
  end if
  if $X \rightarrow C \leftarrow Y$ and $X - Y$ **then**
    Orient $X - Y$ as $X \rightarrow Y$.
  end if
end while
return PDAG

During the second step (Algorithm 2), the algorithm takes the undirected graph G and a set of edges E generated by the first step and outputs a PDAG, i.e. it orients all the edges E. The algorithm starts by finding unshielded triples (X, C, Y) in the graph, i.e. C is connected to both X and Y but X and Y are not directly connected. In this case, if C is not in the separating set of (X, Y) then the unshielded triple is oriented as a v-structure [4] $X \rightarrow C \leftarrow Y$. After identifying v-structures , the algorithm orients the remaining undirected edges by iteratively applying a set of logical rules. For example, if $X \rightarrow C$ and $C - Y$, then $X \rightarrow C \rightarrow Y$. This rule ensures logical consistency: if $C \leftarrow Y$ were true, it would form a new v-structure $X \rightarrow C \leftarrow Y$, which conflicts with the prior identification of v-structures. Therefore, $C \rightarrow Y$ is the only valid orientation under the algorithm's rules.

Following these 2 steps, the PC algorithms generates a causal graph given a data and a significance level.

3 Empirical Analysis of the Effect of the Outcome Variable Binarization on Causal Discovery

We study the effect of the outcome variable binarization on the causal graph generated using the PC algorithm. The empirical analysis encompasses synthetic and real-world datasets to provide a

[4] A v-structure is equivalent to a collider structure, such that $X \rightarrow C \leftarrow Y$.

comprehensive understanding of this effect.

3.1 Experimental Setting

We use the *Gcastle* library in Python for causal discovery, particularly the PC algorithm. We report the average result over 10 runs. We use synthetic and real-world datasets. Mainly, we generate the synthetic data using a framework developed by the team at Inria. Further, we analyse the *Adult* dataset, also known as the *Census Income dataset* [4] and the *Fragrances* dataset [7], previously identified as biased.

Table 1: Metadata of the datasets used in the experiments

Dataset	*Size*	*A*	*Dim.*	*Y*
Synthetic	1000	A	4	Y
Adult [4]	49531	gender	8	income
Fragrances [7]	539	gender	8	price

3.2 Methodology Overview

The analysis begins by constructing a baseline causal graph for each dataset without binarizing the outcome variable Y. The outcome variable Y is then binarized using a range of thresholds and the resulting causal structures are examined to determine the appearance or disappearance of causal edges, particularly the direct edge between the sensitive attribute and the outcome $A \to Y$. To quantify the relationship between A and Y at each step, we compute the Statistical Disparity (SD) [5]:

$$SD = P(Y = 1|A = 1) - P(Y = 1|A = 0)$$

While this metric is more applicable when A and Y are independent, it is still a good frame of reference to quantify the correlation between A and Y.

Recall that in the PC algorithm uses a conditional independence test based on the partial correlation $\rho_{AY|Z}$ between variables A and Y given conditioning on Z, for Z being a neighboring variable of A and Y.

$$\rho_{AY|Z} = \frac{\rho_{AY} - \rho_{AZ} \cdot \rho_{YZ}}{\sqrt{(1 - \rho_{AZ}^2)(1 - \rho_{YZ}^2)}}$$

where:

- ρ_{AY} is the correlation between A and Y.
- ρ_{AZ} is the correlation between A and Z.
- ρ_{YZ} is the correlation between Y and Z.

The statistical disparity SD is linked to ρ_{AY} through the following relationship [6]:

$$\rho_{AY} = \frac{\sqrt{c(1-c)} \cdot SD}{\sqrt{p(1-p)}}$$

[5] Since $A = 0$ is the unprivileged group, all statistical disparity values in this study are negative. For simplicity, we will discuss statistical disparity in terms of its absolute value. It is implied that the group discriminated against is $A = 0$.

[6] The derivation of this equation can be found in the Appendix.

where $c = P(A = 1)$ is constant and $p = P(Y = 1)$ varies when we solely change the threshold of the Y binarization. .

It is important to note that while statistical disparity offers insight into the relationship between A and Y, it does not account for external correlations and causal structures (e.g., mediators, confounders). Specifically, for a given variable Z neighboring A and Y, statistical disparity doesn't account for the correlation between A and Z and the one between Y and Z.

When solely changing the Y binarization threshold, both ρ_{AY} and ρ_{YZ} vary, and ρ_{AZ} stays constant. While the disparity gives an insight on ρ_{AY}, it doesn't account for the changes of ρ_{AZ}. As a result, the metric's utility is more pronounced in datasets with simpler causal structures. In datasets with more complex causal relationships, statistical disparity alone may fail to capture the true causal mechanisms.

By systematically varying the binarization threshold of Y, this methodology explores how such preprocessing decisions influence the causal graph and its interpretation in the context of fairness.

3.3 Empirical Findings from Synthetic Data

The synthetic dataset offers a controlled environment with a known ground truth causal graph, enabling direct comparison of the results.

We generate biased data ($BL = 1.1$) [7] with 1000 data points, a mediator, a confounder and a direct edge between A and Y, with the ground truth graph shown in Figure 2.

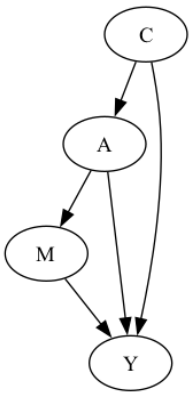


Figure 2: Ground Truth Causal Graph for biased synthetic data

We choose quantile thresholds for the Y variable binarization (e.g., 10%...90%). For example, at the threshold 10%, the Y variable is binarized such that all values of Y falling in the bottom 10% of its distribution are labeled as 0, and the remaining values are labeled as 1.

We obtain a baseline causal graph[8], before binarization, that lacks the edge $A \to M$. However, the ground truth graph is recovered at thresholds $20\%, 30\%, 40\%, 50\%$ and 60% while other thresholds fail to detect the direct edge $A \to Y$. Notably, the statistical disparity in this case is minimal, as the data exhibits low bias ($BL = 1.1$).

[7] BL stands for the *Bias Level*, where the higher the bias level, the higher the discrimination against $A = 0$. Note that for $BL = 1$, the data is unbiased.

[8] The causal graphs generated with synthetic data can be found in the Appendix.

3.4 Empirical Findings from Real-World Data

3.4.1 Fairness Benchmarck Dataset: The *Adult* Dataset

We extend this analysis to a fairness benchmark dataset, in particular the *Adult* dataset, also known as the *Census Income dataset* [4]. We consider the gender as the sensitive attribute A and the income as the outcome Y. This dataset, is particularly important for this study because when this data was released in 1994, the income was binarized using the threshold at $50K$, it became widely known for fairness studies as it exhibited strong biases. The non-binarized income values were made public only recently [3], providing an opportunity to analyze the dataset in its continuous form.

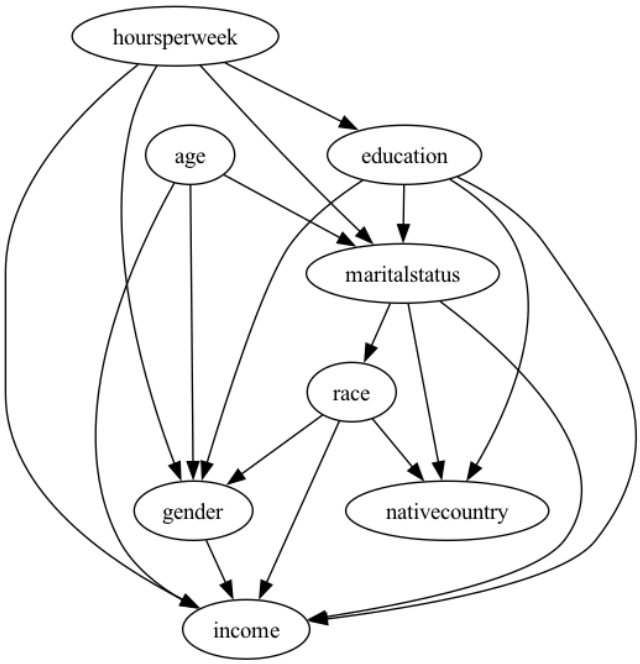


Figure 3: Baseline Causal Graph for Adult dataset

We can observe in the baseline causal graph [9] shown in Figure 3 that a direct edge $A \to Y$ is present, consistent with existing literature. Moreover, we can observe several causal structures. For example, age, race and hours per week are confounders.

We choose the binarization thresholds for Y at $10K...90K$, to be consistent with the literature on the subject. We observe that a direct edge $A \to Y$ only appears at the thresholds $10K$ and $50K$. This is highly relevant, as in the literature [4] the binarization level used was $50K$.

[9] The causal graphs generated with the *Adult* dataset can be found in the Appendix.

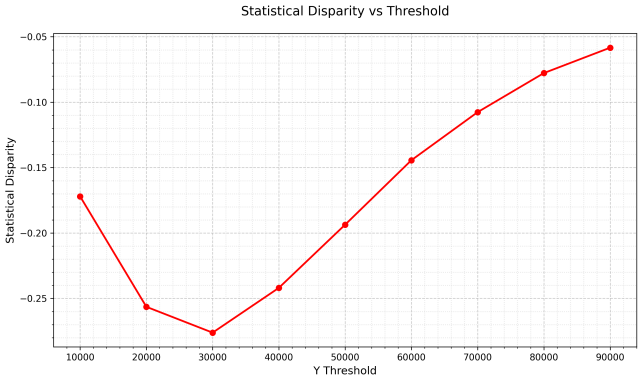


Figure 4: Statistical Disparity when varying the Threshold for the Adult Dataset

We can observe that the Statistical disparity is higher at the thresholds $20K$, $30K$ and $40K$ suggesting a higher correlation between A and Y and hence more discrimination, considering this fairness metric. Yet, the direct edge $A \to Y$ only appears at the thresholds $10K$ and $50K$ and doesn't appear at thresholds with higher statistical disparity. This happens because in this dataset, there are multiple causal structures around A and Y involving many neighboring variables that statistical disparity doesn't account for. For example, at the thresholds $20K$, $30K$ and $40K$, race, marital status and education are confounders. Therefore, statistical disparity doesn't give insight on a direct effect between A and Y, since they are obviously not independent.

The *Adult* dataset demonstrates that statistical disparity peaks do not always correspond to direct causal effects due to the dataset's complex causal structures.

3.4.2 The *Fragrances* Dataset

In a previous research project conducted as part of an internship at Columbia Business School, we investigated gender-based price discrimination in the product category of fragrances using the *Fragrances* dataset [7]. The analysis revealed a pink tax phenomenon, i.e. a price disparity between products marketed to men and those marketed to women, in the list price of fragrances. Building on this work, the present study further examines this dataset [10] to identify potential causal structures underlying this gender-based pricing bias.

Accordingly, we extend the analysis to the real-world *Fragrances* dataset [7], which was previously identified as biased. We consider the gender as the sensitive attribute A and the price per milliliter as the outcome Y. This dataset is particularly relevant for examining gender-based pricing disparities, also referred to as *pink tax* [11].

[10] Note that to be able to compare the results, we use the already preprocessed data from the previous research, and binarize the variables.

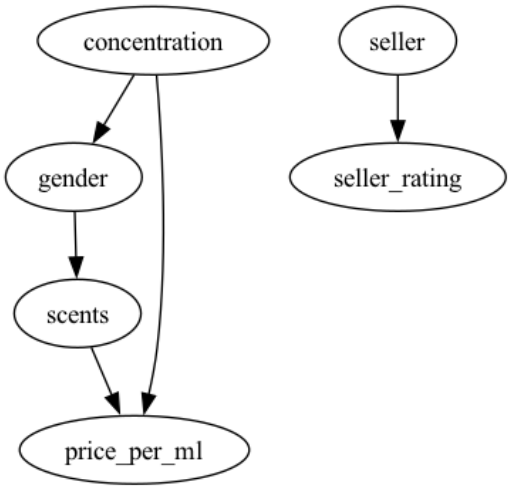


Figure 5: Baseline Causal Graph for the Fragrances dataset

The baseline graph[11] in Figure 5, reveals no direct edge between A and Y. However, we can observe that Concentration acts as a confounder, and Scents as a mediator, suggesting an indirect effect of gender on the price per milliliter.

We choose quantile thresholds for the Y variable binarization (e.g., 10%...90%). We observe that a direct edge $A \to Y$ only appears at the thresholds 60%, 70% and 80%. We further observe that Concentration is always present as a confounder at all thresholds and that Scents only appears as a mediator at the thresholds 30%, 40% and 80%.

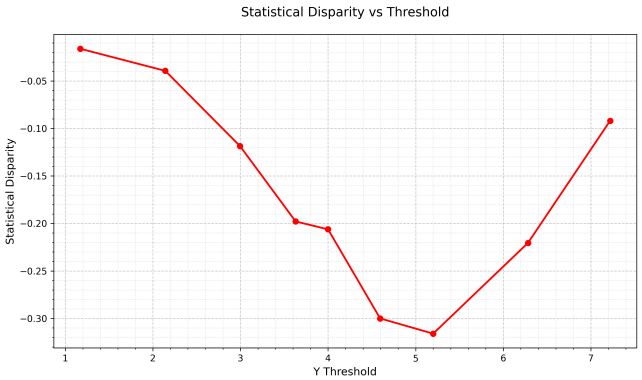


Figure 6: Statistical Disparity when varying the Threshold for the Fragrances dataset

We can observe that the statistical disparity reaches its maximum in magnitude at the threshold 70% [12], which is consistent with the appearance of the direct edge $A \to Y$ around this threshold. Notice that in this case, the data has less complicated causal structures around A and Y. This suggests that in datasets with simpler causal structures, statistical disparity can provide meaningful insights into direct causal effects.

[11] The causal graphs generated with the *Fragrances* dataset can be found in the Appendix.

[12] Since quantile thresholds are used, the x-axis represents the threshold values for Y. For instance, a threshold of 70% corresponds to a value of 5.20, indicating that all values in the bottom 70% of the distribution (i.e., less than 5.20) are labeled as 0, while the remaining values are labeled as 1.

However, in contrast with the previous analysis on this dataset, the variable Brand doesn't appear to have an effect on the price, while it appeared to be significantly impacting the price when using the regression analysis.

The *Fragrances* dataset analysis shows threshold-specific emergence of $A \to Y$ edges, correlating with peaks in statistical disparity. This suggest that binarization thresholds might reveal hidden causal structures, such as the direct influence of gender on pricing. In datasets with fewer confounding and mediating variables, statistical disparity aligns more closely with direct causal effects, underscoring its utility in simpler causal systems. That is, statistical disparity is equivalent to total effect (TE) [13] in the absence of causal structures (e.g., confounders, mediators, colliders) between A and Y.

4 Empirical Analysis of the Effect of Bias on Causal Discovery

We analyse the impact of bias by generating controlled synthetic data with varying bias levels and amplifying the existing bias in the real world dataset. This approach allows for a detailed examination of how different bias conditions and binarization thresholds affect the resulting causal structures.

4.1 Empirical Findings from Synthetic Data

4.1.1 Unbiased Synthetic Data

We start by generating an unbiased dataset ($BL = 1$) with 1000 data points, no direct edge between A and Y, a collider and a mediator, as shown in Figure 7.

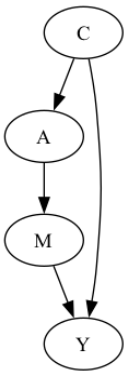


Figure 7: Ground Truth Causal Graph for unbiased synthetic data

The baseline causal graph[14], generated without binarization, lacks the edge $A \to M$

We choose a quantile thresholds for the Y variable binarization (e.g., 10%...90%). Across all thresholds, the resulting causal graphs remain consistent with the baseline, showing no direct edge $A \to Y$ and consistently missing the edge $A \to M$. Additionally, the statistical disparity is negligible, aligning with the absence of bias ($BL = 1$).

These results suggest that when data is unbiased, no direct discrimination is detected, and the causal structure remains stable across binarization thresholds.

[13] Total Effect[12] (TE) refers to the overall effect of the variable A on the variable Y. It is defined as: $TE = \mathbb{E}[Y \mid \text{do}(A = a_1)] - \mathbb{E}[Y \mid \text{do}(A = a_0)]$ where a_0 and a_1 are specific values of A (e.g., male and female) and $\text{do}(A = a)$ represents an intervention that sets A to the value a.

[14] The causal graphs generated with synthetic data can be found in the Appendix.

We further generate unbiased data ($BL = 1$), with 1000 data points, a mediator, a confounder and a direct edge between A and Y, as shown in Figure 2.

Despite the ground truth containing the direct edge $A \to Y$, it does not appear in the baseline causal graph or in any causal graphs obtained at any binarization thresholds. The edge $A \to M$ is also absent across all causal graphs. In addition, the statistical disparity remains negligible as the data is unbiased.

These findings, consistent with team's work findings, highlight that when the data is unbiased, the causal discovery algorithm may fail to detect certain causal structures such as the first edge in mediator structures $A \to M$, and the direct edge $A \to Y$, even when they are part of the ground truth. This shows limitations of the causal discovery algorithm in identifying key causal relationships in unbiased data.

4.1.2 Biased Synthetic Data

To analyze the effect of bias on causal discovery, we generate biased data with 1000 data points, a mediator, a confounder and a direct edge between A and Y, with the same ground truth graph shown in Figure 2. We choose different bias levels such that $BL \in \{1.1, 1.3, 1.5, 2, 3\}$, where higher BL values represent greater discrimination against $A = 0$. We use quantile thresholds for the Y variable binarization.

- $BL = 1.1$: The baseline causal graph lacks the edge $A \to M$. The ground truth graph is recovered at thresholds $20\%, 30\%, 40\%, 50\%$ and 60% while other thresholds fail to detect the direct edge $A \to Y$.
- $BL = 1.3$: The baseline causal graph deviates from the ground truth by missing the edge $C \to A$ and adding the edge $C \to M$. The ground truth graph is recovered at all thresholds except 80% and 90%, where the direct edge $A \to Y$ is missing.
- $BL = 1.5$: The baseline causal graph matches the ground truth, with consistent results across all thresholds, except at the threshold 90% where the direct edge $A \to Y$ disappears.
- $BL = 2$, The baseline causal graph deviates from the ground truth by missing the edge $C \to A$ and adding the edge $C \to M$. However, binarizing Y restores the ground truth graph across all thresholds. This indicates that binarization may provide a layered perspective, revealing causal structures absent in the baseline.
- $BL = 3$: Higher bias distorts some causal structures. The ground truth graph is recovered through the baseline causal graph and most thresholds. However, at thresholds $30\%, 40\%$ and 50%, the edge $C \to A$ disappears and at at threshold 90%, the direct edge $A \to Y$ disappears.

Observe that as the bias increases, the direct edge $A \to Y$ appears more frequently in both baseline and binarized causal graphs. However, at very high bias levels, causal structures, such as confounders and mediators, become distorted.

Interestingly, in some cases where the baseline causal graph fails to capture the full ground truth, the causal graphs obtained after binarization successfully restore it. This suggests that causal graphs after binarization can provide additional perspectives on the ground truth causal structure, particularly in the presence of bias.

4.1.3 Impact of Bias on Statistical Disparity

The statistical disparity is analyzed at each bias level. Since Y is not binary initially, the statistical disparity is evaluated only after binarization. Specifically, we compute the average and maximum (in magnitude) statistical disparities across thresholds.

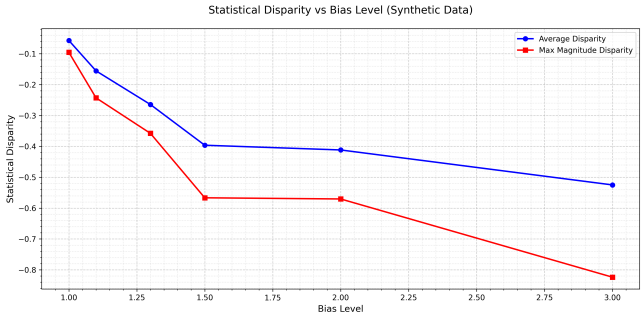


Figure 8: Statistical disparity when varying the bias level for the synthetic datasets

We observe that as the bias increases, both the average and the maximum statistical disparities increase in magnitude in (Figure 8).

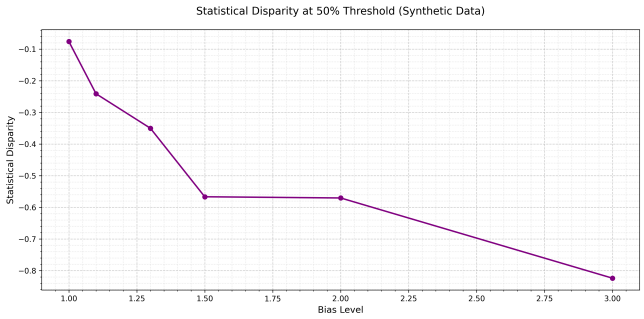


Figure 9: Statistical disparity vs bias for the synthetic datasets for a fixed Y binarization threshold at 50%

Figure 9 illustrates the statistical disparity at the fixed Y binarization threshold of 50%, further confirming that increased bias amplifies the disparity.

We conclude that as the bias increases, the statistical disparity increases in magnitude, and the direct edge $A \to Y$ appears more frequently in baseline and binarized causal graphs. At very high bias levels, some causal structures get distorted. These findings align with the team's findings [10].

4.2 Empirical Findings from Real-World Data: The *Fragrances* Dataset

4.2.1 Impact of Bias on the Causal Graph

We try to further analyse the effect of bias on the causal graph by amplifying the existing bias in the real world dataset.

We inject bias by multiplying the price per milliliter (Y) for women ($A = 0$) by α, then adding Gaussian noise to the price per milliliter as follows:

$$Y' = \begin{cases} \alpha \cdot Y + \epsilon_0, & \text{if } A = 0 \\ Y + \epsilon_1, & \text{if } A = 1 \end{cases}$$

where:

- α is the scaling factor applied to Y for $A = 0$.
- $\epsilon_0 \sim \mathcal{N}(0, \sigma_0^2)$ is Gaussian noise added for $A = 0$, with mean 0 and standard deviation σ_0.
- $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ is Gaussian noise added for $A = 1$, with mean 0 and standard deviation σ_1.
- σ_0 and σ_1 control the variability of the noise for $A = 0$ and $A = 1$, respectively.

This formulation introduces systematic scaling (α) and variability differences ($\sigma_0 \neq \sigma_1$) to model gender-based pricing bias. We take $\alpha \in \{1.1, 1.2, 1.5, 2, 3\}$, $\sigma_0 = 0.3$ and $\sigma_1 = 0.2$.

4.2.1.1 Comparing Baseline Causal Graphs when Amplifying Bias

- $\alpha = 1.1$: The baseline causal graph remains unchanged, matching the original baseline graph (Figure 5). In this graph, Scents acts as a mediator, Concentration as a confounder, and no direct edge $A \to Y$ is observed.
- $\alpha \in \{1.2, 1.5, 2, 3\}$: As bias increases, a direct edge $A \to Y$ emerges in the baseline causal graph. This indicates that higher bias amplifies the direct influence of the sensitive attribute on the outcome.

The results demonstrate a clear relationship between increasing bias and the appearance of a direct edge $A \to Y$. As α increases, the direct causal relationship between A and Y becomes more apparent, reflecting amplified gender-based pricing disparities.

4.2.1.2 Comparing Causal Graphs with Binarized Outcome when Amplifying Bias

- $\alpha = 1.1$: A direct edge $A \to Y$ appears at the thresholds $60\%, 70\%, 80\%$ and 90%. Concentration remains a confounder across all thresholds, while Scents acts as a mediator at the thresholds 40% and 80%.
- $\alpha = 1.2$: A direct edge $A \to Y$ appears at the thresholds $50\%, 60\%, 70\%, 80\%$ and 90%. Concentration remains a confounder across all thresholds and Scents becomes a mediator at the thresholds $50\%, 70\%$ and 80%.
- $\alpha = 1.5$: A direct edge $A \to Y$ appears at the thresholds $50\%, 60\%, 70\%, 80\%$ and 90%. Concentration remains a confounder across all thresholds and Scents becomes a mediator at the thresholds $60\%, 70\%, 80\%$ and 90%.
- $\alpha = 2, 3$: A direct edge $A \to Y$ appears at all thresholds except for 10%. Concentration remains a confounder across all thresholds and Scents becomes a mediator at the thresholds $40\%, 50\%, 60\%, 70\%, 80\%$ and 90%.

The results reveal that higher levels of injected bias increase the frequency of the direct edge $A \to Y$ across binarization thresholds. This suggests that as bias intensifies, the sensitive attribute A exerts a stronger direct influence on the outcome Y even after binarization.

4.2.2 Impact of Bias on Statistical Disparity

To quantify the effects of bias on statistical disparity, the dataset is analyzed at varying levels of injected bias $\alpha \in \{1, 1.1, 1.2, 1.5, 2, 3\}$. Statistical disparity is calculated for each binarization threshold of the outcome variable. Since Y is continuous in its original form, statistical disparity is defined only after binarization. We focus on its average and maximum magnitude across thresholds.

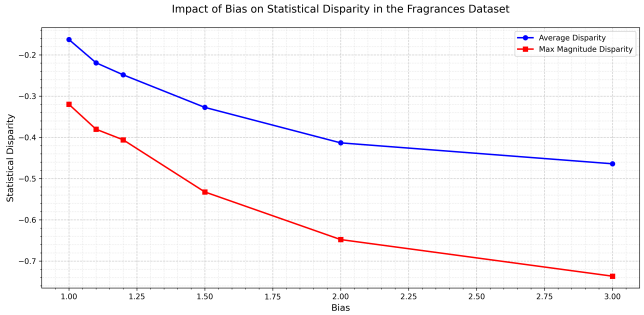


Figure 10: Statistical disparity when varying the bias for the Fragrances dataset

As the bias increases, statistical disparity (in magnitude) grows significantly. At very high bias levels, statistical disparity begins to converge, indicating a limit to the extent of disparity introduced by the scaling factor.

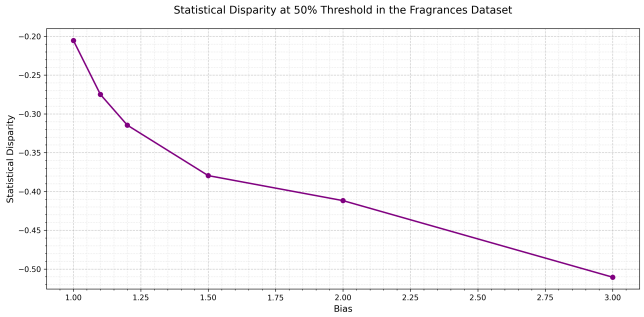


Figure 11: Statistical disparity vs bias for the Fragrances dataset for a fixed Y binarization threshold at 50%

A similar trend is observed when the binarization threshold is fixed at 50% (Figure 11). The statistical disparity increases with bias levels.

We conclude that higher bias amplifies the statistical disparity and the likelihood of a direct edge $A \to Y$ appearing in the causal graph. These findings are consistent with the findings on the controlled experiments with synthetic data.

5 Conclusion

We conclude that the binarization threshold of the outcome variable impacts the causal graph structure significantly, particularly the direct edge $A \to Y$. This effect varies across datasets, depending on the causal complexity surrounding the sensitive attribute and the outcome. Taking this into account, data

preprocessing choices, such as outcome binarization, should be carefully considered to avoid misleading fairness conclusions.

In simpler dataset, with less causal structures surrounding the sensitive attribute and the outcome, statistical disparity aligns more closely with the appearance of direct causal edges. However, in complex datasets, statistical disparity fails to account for confounding and mediating variables, leading to mismatches between observed correlations and actual direct effects.

Further, higher bias amplifies statistical disparity and the likelihood of discovering a direct causal edges between the sensitive attribute and the outcome. But with extreme biases, causal structures (e.g., mediators and confounders) get distorted. However, in unbiased datasets, the causal discovery algorithm fails to capture critical causal edges, including the direct edge between the sensitive attribute and the outcome. We therefore advise seeking a professional opinion when analyzing unbiased data to ensure accurate interpretation of the results.

Acknowledgment

I would like to express my sincere gratitude to the Comète Team, led by Dr. Catuscia Palamidessi, at Inria for their guidance. I am deeply grateful to Dr. Karima Makhlouf for her consitent feedback, invaluable for this laboratory research project.

References

[1] Julia Angwin et al. "Machine bias. ProPublica". In: *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

[2] Rūta Binkytė et al. "Causal Discovery for Fairness". In: *Proceedings of the Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*. Vol. 214. Proceedings of Machine Learning Research. PMLR, 2023. URL: https://proceedings.mlr.press/v214/binkyte23a.html.

[3] Frances Ding et al. "Retiring Adult: New Datasets for Fair Machine Learning". In: *Advances in Neural Information Processing Systems* 34 (2021).

[4] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[5] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[6] Ruocheng Guo et al. "A survey of learning causality with data: Problems and methods". In: *ACM Computing Surveys (CSUR)* 53.4 (2020), pp. 1–37.

[7] Kaggle. *Noon Perfume Dataset*. https://www.kaggle.com/datasets/monirahabdulaziz/noon-perfume. 2021.

[8] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "Machine learning fairness notions: Bridging the gap with real-world applications". In: *Information Processing & Management* 58.5 (2021), p. 102642.

[9] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "Survey on causal-based machine learning fairness notions". In: *arXiv preprint arXiv:2010.09553* (2020).

[10] Karima Makhlouf et al. "Causal Discovery of Biased Data". Unpublished manuscript. 2024.

[11] Sarah Moshary, Anna Tuchman, and Natasha Vajravelu. "Gender-Based Pricing in Consumer Packaged Goods: A Pink Tax?" In: *Marketing Science* Articles in Advance (2023), pp. 1–14. DOI: 10.1287/mksc.2023.1452.

[12] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[13] Peter Spirtes and Clark Glymour. "An algorithm for fast recovery of sparse causal graphs". In: *Social Science Computer Review* 9.1 (1991), pp. 62–72.

[14] US LEMAS survey. *Communities and crime dataset*. 1990. URL: https://archive.ics.uci.edu/dataset/183/communities+and+crime (visited on 12/07/2009).

Appendix

Derivation of the relationship between ρ_{AY} and SD:

Let A and Y be binary variables and Z a neighboring variable of A and Y.

We define statistical disparity as follows:

$$SD = P(Y = 1|A = 1) - P(Y = 1|A = 0)$$

The correlation between A and Y is expressed as:

$$\rho_{AY} = \frac{\text{Cov}(A, Y)}{\sigma_A \sigma_Y} = \frac{\text{Cov}(A, Y)}{\sqrt{Var(A)}\sqrt{Var(Y)}}$$

The covariance $\text{Cov}(A, Y)$ for binary A and Y can be derived as:

$$\text{Cov}(A, Y) = P(Y = 1, A = 1) - P(Y = 1).P(A = 1)$$

Since $P(A = 1)$ remains constant, we consider $P(A = 1) = c$, with c constant.

$$\text{Cov}(A, Y) = P(Y = 1, A = 1)(1 - c) - P(Y = 1, A = 0).c$$

$$\text{Cov}(A, Y) = (P(Y = 1|A = 1) - P(Y = 1|A = 0)).(1 - c)c$$

$$\text{Cov}(A, Y) = SD.(1 - c)c$$

where SD is the statistical disparity, and $c = P(A = 1)$.

Thus, the correlation ρ_{AY} between A and Y can be expressed as:

$$\rho_{AY} = \frac{\sqrt{c(1 - c)} \cdot SD}{\sqrt{p(1 - p)}}$$

where $p = P(Y = 1)$ and $c = P(A = 1)$.

Causal Graphs

Causal Graphs for Synthetic Data

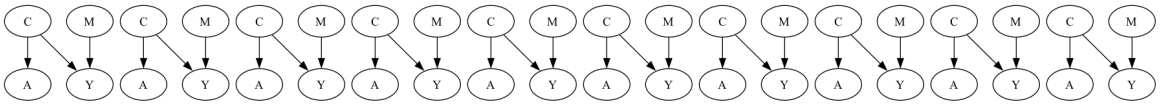


Figure 12: Causal graphs for unbiased synthetic data ($BL = 1$) with no direct edge $A \to Y$: Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

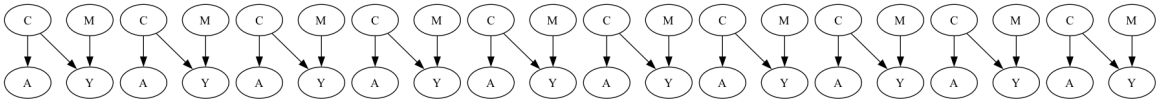


Figure 13: Causal graphs for unbiased synthetic data ($BL = 1$) with a direct edge $A \to Y$: Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

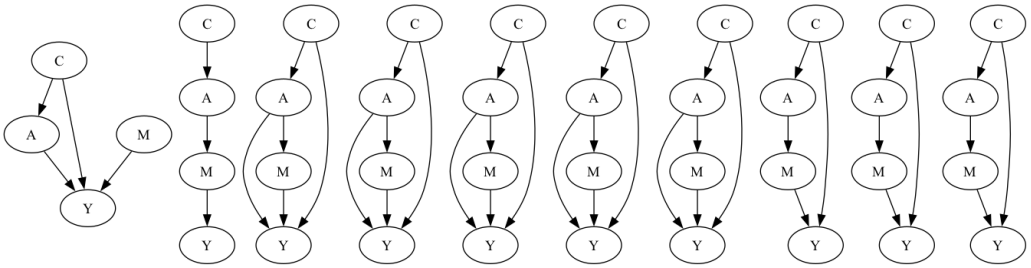


Figure 14: Causal graphs for biased synthetic data ($BL = 1.1$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

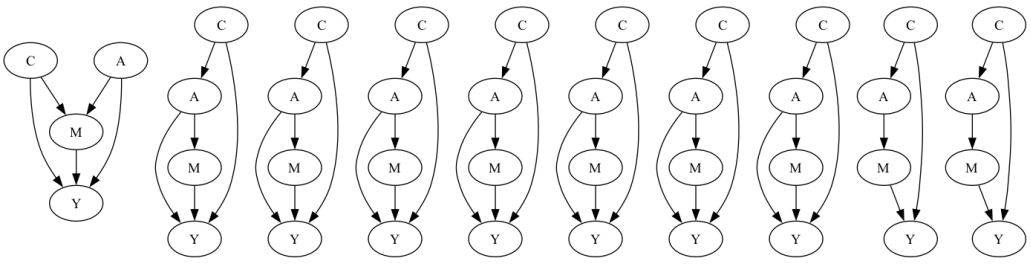


Figure 15: Causal graphs for biased synthetic data ($BL = 1.3$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

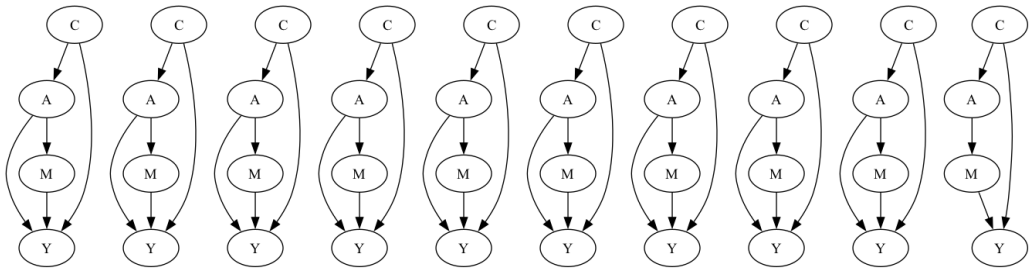


Figure 16: Causal graphs for biased synthetic data ($BL = 1.5$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

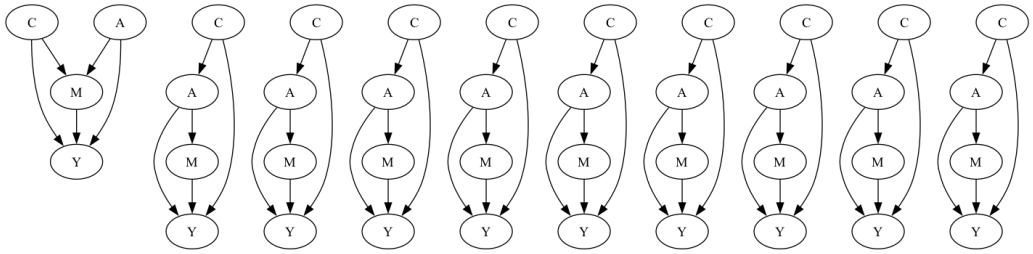


Figure 17: Causal graphs for biased synthetic data ($BL = 2$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

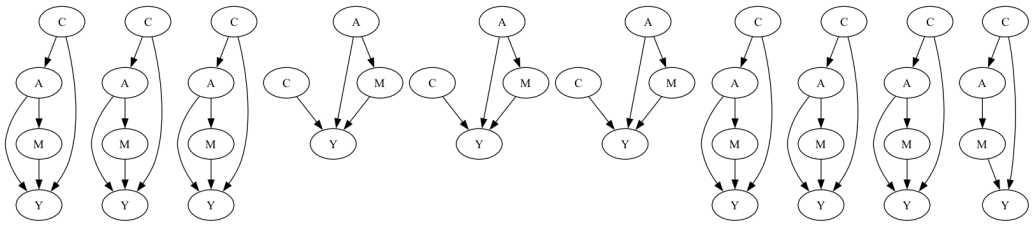


Figure 18: Causal graphs for biased synthetic data ($BL = 3$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Causal Graphs for Real World Data

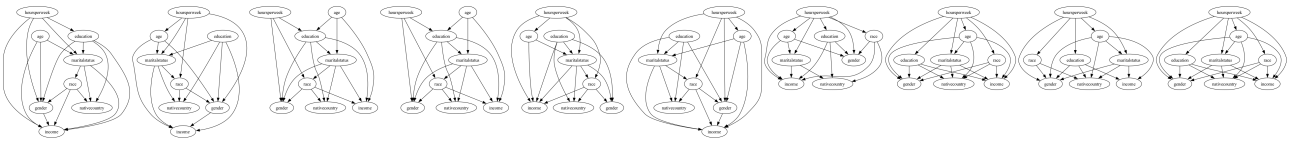


Figure 19: Causal graphs for the Adult dataset: Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

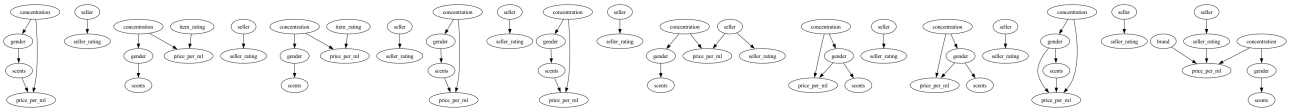


Figure 20: Causal graphs for the Fragrances dataset: Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Figure 21: Causal graphs for the Fragrances dataset with injected bias ($\alpha = 1.1$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Figure 22: Causal graphs for the Fragrances dataset with injected bias ($\alpha = 1.2$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Figure 23: Causal graphs for the Fragrances dataset with injected bias ($\alpha = 1.5$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Figure 24: Causal graphs for the Fragrances dataset with injected bias ($\alpha = 2$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively

Figure 25: Causal graphs for the Fragrances dataset with injected bias ($\alpha = 3$): Baseline causal graph, causal graph at threshold 10%, ..., causal graph at threshold 90%, respectively