Causal Discovery for Fairness

Rūta Binkytė RUTA.BINKYTE@INRIA.FR
Karima Makhlouf KARIMA.MAKHLOUF@INRIA.FR
Carlos Pinzón CARLOS.PINZON@INRIA.FR
Sami Zhioua ZHIOUA@LIX.POLYTECHNIQUE.FR
Catuscia Palamidessi CATUSCIA@LIX.POLYTECHNIQUE.FR
Inria, École Polytechnique, IPP Paris, France

Abstract

Fairness guarantees that the ML decisions do not result in discrimination against individuals or minority groups. Identifying and measuring reliably fairness/discrimination is better achieved using causality which considers the causal relation, beyond mere association, between the sensitive attribute (e.g. gender, race, religion, etc.) and the decision (e.g. job hiring, loan granting, etc.). The big impediment to the use of causality to address fairness, however, is the unavailability of the causal model (typically represented as a causal graph). Existing causal approaches to fairness in the literature do not address this problem and assume that the causal model is available. In this paper, we do not make such an assumption and we review the major algorithms to discover causal relations from observable data. This study focuses on causal discovery and its impact on fairness. In particular, we show how different causal discovery approaches may result in different causal models and, most importantly, how even slight differences between causal models can have significant impact on fairness/discrimination conclusions.

Keywords: Causal discovery, fairness, causal effect, fairness metrics

1. Introduction

Several fairness criteria have been introduced in the literature to assess discrimination (statistical parity Darlington (1971), equal opportunity Hardt et al. (2016), calibration Chouldechova (2017), etc.) Makhlouf et al. (2021). The most recent fairness criteria, however, are causal-based Makhlouf et al. (2020a) and reflect the now widely accepted idea that causality is necessary to appropriately address the problem of fairness. There are at least three benefits of using causality to assess fairness. First, in the presence of a common cause (confounder) between the sensitive attribute A (e.g. gender) and the decision Y (e.g., job hiring) as in Figure 1(a), using conditional probability $\mathbb{P}(Y|A)$ leads to wrong conclusions about the dependence of Y on A. Confounders are the reason why we say that "correlation is different than causation". A more reliable measure of the dependence between Y and A is the causal effect of A on Y which is typically computed by adjusting on confounders. Second, causality is well equipped to carry out mediation analysis, that is, distinguishing the different paths of causal effects as shown in Figure 1(b), A causal effect between two variables A and Y can be classified as direct ($A \to Y$), indirect ($A \to R \to Y$ and $A \to E \to Y$), or a path-specific effect (only through $A \to R \to Y$ or $A \to E \to Y$). This is very relevant to fairness as a direct effect is always unfair, while an indirect or a path-specific effect may be unfair or fair depending on the mediator variable: an indirect effect through a redlining/proxy variable (R) is unfair, while an indirect effect through an explaining variable (E) is acceptable (fair). Third, in some legal liability frameworks such as disparate treatment Barocas and Selbst (2016), discrimination claims require the plaintiff to demonstrate a causal connection between the challenged decision (e.g. hiring, firing, admission) and the sensitive attribute (e.g. gender, race, age). The main impediment to causal inference is the unavailability of the true causal graph which indicates the causal relations between variables. Causal graphs can be set manually by experts in the field, but are very often generated using experiments (also called interventions). The process of identifying the causal graph from the data is called causal discovery or structure learning.

A large number of causal discovery algorithms exist in the literature. The majority of causal discovery algorithms fall into three categories: constraint-based, score-based, and procedures that exploit semi-parametric assumptions. In the constraint-based category, algorithms rely mainly on the (conditional) independencies present in the data to discover causal relations between variables. Therefore their efficiency depends on the reliability of the conditional independency test procedure. Score-based algorithms rely instead on goodness-of-fit tests. They learn causal graphs by maximizing a scoring criterion such as the Bayesian Information Criterion (BIC) Schwarz (1978) which trades-off accuracy (fitness of graph to the data) with complexity (the number of parameters in the model). Algorithms in the third category use additional assumptions to learn causal relations more efficiently and in more details. The most common assumptions relevant to the third category are linearity of the model and non-gaussianity of the regression residuals.

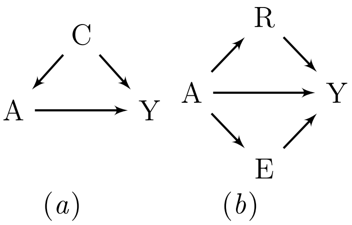


Figure 1: Causal graphs illustrating confounding (a) and mediation analysis (b).

This paper studies the problem of discovering causal graphs[1] to be used to assess fairness of machine learning-based decision systems. As reference, we provide, in Appendix A, an intuitive explanation of the major causal discovery algorithms (PC Spirtes et al. (1999) and its FCI extension Spirtes et al. (1999), GES Hauser and Bühlmann (2012), directLiNGAM Shimizu et al. (2011), and the fairness related discovery algorithm SBCN Bonchi et al. (2017)). This is helpful to understand why different discovery approaches may generate different causal structures. The main contribution of the paper is to use real-world fairness benchmark datasets to illustrate the consequences of slight differences in causal structures on fairness conclusions.

2. Causal Structure and Fairness

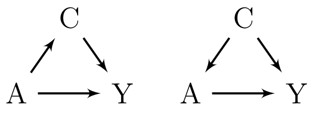
Several fairness notions rely on causality to assess fairness and hence require a causal graph. The most basic causal-based fairness notion is total effect (TE) Pearl (2009)[2] which considers the overall effect of a variable A on a variable Y. Assume that the sensitive variable A can take two possible values a_0 (e.g. female) and a_1 (e.g. male) and that the positive outcome is y^+ (e.g. hiring), $TE_{a_1,a_0}(y^+)$ is defined as $\mathbb{P}[Y{=}y^+ \,|\, do(A{=}a_1)] - \mathbb{P}[Y{=}y^+ \,|\, do(A{=}a_0)]$ which

[1] Interested readers can find in Appendix A quick descriptions of commonly used causal discovery algorithms

[2] Known also as average causal effect (ACE).

measures the effect of the change of A from a_1 to a_0 on $Y = y^+$ along all the causal paths from A to Y. $\mathbb{P}[Y{=}y \,|\, do(A{=}a)]$ denotes the probability of $Y = y$ after an intervention $do(A = a)$. This is equivalent to the probability of $Y = y$ after forcing all individuals in the population to have value $A = a$. $\mathbb{P}[Y{=}y \,|\, do(A{=}a)]$ is denoted $\mathbb{P}[y_a]$ for short.

Direct effect (DE) is another fairness notion which focuses exclusively on the direct path $A \rightarrow Y$ (ignoring all indirect paths between A and Y). The most general formulation of DE is natural direct effect (NDE) Pearl (2001) defined as $NDE_{a_1,a_0}(y^+) = \mathbb{P}[y^+_{a_1,\mathbf{Z}_{a_0}}] - \mathbb{P}[y^+_{a_0}]$ where $\mathbf{Z}$ is the set of mediator variables and $\mathbb{P}[y^+_{a_1,\mathbf{Z}_{a_0}}]$ is the probability of $Y = y^+$ had A been a_1 and had $\mathbf{Z}$ been the value it would naturally take if $A = a_0$. Indirect effect (IE), which focuses rather on the indirect paths from A to Y, can be computed using the natural indirect effect (NIE) Pearl (2001) formula $NIE_{a_1,a_0}(y^+) = \mathbb{P}[y^+_{a_0,\mathbf{Z}_{a_1}}] - \mathbb{P}[y^+_{a_0}]$. Using the identifiability theory of causal inference Shpitser and Pearl (2008); Makhlouf et al. (2022), the above expressions of fairness notions, involving interventions and counterfactuals, can be expressed in terms of observable probabilities.



(a) C is a mediator (b) C is a confounder

Figure 2: Two simple causal graphs differing only on the direction of the edge between A and C.

To see the impact of the causal graph structure on how these fairness notions can be identified and computed, consider the two simple graphs in Figure 2. Although both graphs differ in the orientation of a single edge between A and C, they lead to significantly different expressions for the causal fairness notions.

TE is identifiable in both graphs. In the left graph, since there is no confounder, an intervention on A ($do(A = a)$) coincides with conditioning on $A = a$. Hence, $TE_{a_1,a_0}(y^+) = \mathbb{P}[y^+_{a_1}] - \mathbb{P}[y^+_{a_0}] = \mathbb{P}[y^+ \,|\, A{=}a_1] - \mathbb{P}[y^+ \,|\, A{=}a_0]$ which coincides with total variation (TV) Makhlouf et al. (2020a). However, in the slightly different graph on the right, C is a confounder, and hence TE is computed by adjusting on C [3]: $TE_{a_1,a_0}(y^+) = \mathbb{P}[y^+_{a_1}] - \mathbb{P}[y^+_{a_0}] = \sum_{c \in dom(C)} \left(\mathbb{P}[Y{=}y^+ \,|\, a_1, c] - \mathbb{P}[Y{=}y^+ \,|\, a_0, c]\right) \mathbb{P}[C{=}c]$. For NDE, it is computed the same way in both graphs since it requires blocking all non-direct paths which is achieved by adjusting on variable C: $NDE_{a_1,a_0}(y^+) = \sum_{c \in dom(C)} \mathbb{P}[C{=}c] \left(\mathbb{P}[Y{=}y^+ \,|\, a_1, c] - \mathbb{P}[Y{=}y^+ \,|\, a_0, c]\right)$. For the indirect effect, NIE is equal to zero in the right graph since there is no causal indirect path between A and Y, while for the left graph, it is equal to: $NIE_{a_1,a_0}(y^+) = \sum_{c \in dom(C)} \mathbb{P}[y^+ \,|\, a_0, c](\mathbb{P}[c \,|\, a_1] - \mathbb{P}[c \,|\, a_0])$.

The disparity of identifying causal fairness notions due to slight differences in the causal graphs holds also for other fairness notions Makhlouf et al. (2020a). This is further illustrated in the following experimental analysis section.

3. Experimental Analysis

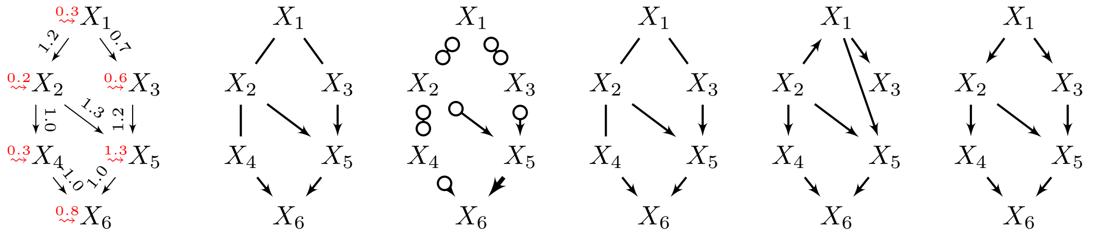
We apply the different causal discovery algorithms on one synthetic dataset and three real-world fairness benchmark datasets. We use Tetrad Ramsey et al. (2018) implementation of PC, FCI and GES algorithms with a significance threshold (α) set to 0.01 for conditional

[3] We are considering the discrete case.

independence testing. For SBCN discovery, we use the same implementation as the original paper Bonchi et al. (2017). The only background knowledge we use in this study is temporal order using tiers. Variables are split into a set of ordered tiers (tier 1, tier 2, ... tier n) which imply the following constraints. A variable in tier i can be the cause of variables in the same tier or in subsequent tiers $(i+1\dots n)$ but not of variables in previous tiers $(1\dots i-1)$. We use five different causal-based fairness notions, namely, ATE_IPW, TE, DE, ID, and ED which correspond, respectively, to average total effect using inverse propensity weighting Imbens and Rubin (2015), total effect, direct effect, indirect discrimination, and explainable discrimination. Indirect discrimination and the explainable discrimination compute both the indirect causal effect between the sensitive variable and the outcome. However, indirect discrimination measures the path-specific effect with a proxy/redlining variable while the explainable discrimination considers the path-specific effect with an explaining variable. Thus, while the first is discriminatory, the second is legitimate and hence should be removed from the causal effect estimation. These and other causal-based fairness notions are described by Makhlouf et al. Makhlouf et al. (2020b). The *paths* package implementation Zhou and Yamamoto (2020) is used to estimate TE, DE, ID, and ED. Computing (or estimating) discrimination consists in subtracting the probability of positive (desirable) output (e.g. hiring, granting a loan, etc.) for the protected group (e.g. female) from the probability of positive output of the privileged group (e.g. male) as expressed in Section 2. This leads to values in the range $[-1,+1]$. A value of 0 means the outcome is fair (no discrimination), a positive value indicates a discrimination *against* the protected group, and a negative value indicates a discrimination *in favor* of the protected group. Estimating discrimination using all the above measures requires the knowledge of the confounder and mediator variables. However, PC, FCI, and GES algorithms can output partially directed graphs (PDAG) which do not guarantee to tell if a certain variable is a confounder or a mediator since some edges are left undirected. In such cases, we consider all possible ways of directing the (typically few) undirected edges (as long as they do not introduce a v-structure). For instance, if there are two undirected edges $X-W$ and $Z-Y$, there are 4 ways of directing them: $X\rightarrow W$ and $Z\rightarrow Y$, $X\leftarrow W$ and $Z\rightarrow Y$, $X\rightarrow W$ and $Z\leftarrow Y$, and $X\leftarrow W$ and $Z\leftarrow Y$. For each combination, we compute the discrimination and finally we report the range of values. This can be seen as bounding the discrimination value. Interpreting variables differently (explaining vs redlining) can also result in a range of values in mediation analysis (computing DE, ID, and ED).

3.1. Synthetic linear dataset

We generated a continuous linear dataset with a very simple causal structure (Figure 3(a)) to illustrate the main properties of the causal discovery algorithms discussed in this article. In general, synthetic datasets are crucial for testing causal discovery algorithms systematically because, unlike real-world datasets, the ground truth graph is known. Each edge in Figure 3(a) has a weight that was chosen randomly, and each node has a gaussian source of noise whose standard deviation is indicated in red. The value of a node is the weighted sum of the values of the parents plus the noise. For instance, the values of X_1 and X_5 are generated as $X_1=\mathcal{N}(0,0.3)$ and $X_5=1.3X_2+1.2X_3+\mathcal{N}(0,1.3)$ respectively. Figure 3 shows the graphs generated by each algorithm for 100000 samples. On the one hand, PC, FCI, and



(a) Ground truth (b) PC (c) FCI (d) GES (e) LiNGAM (f) LiNGAM*

Figure 3: Generated causal graphs for the synthetic dataset with gaussian noise and *uniform noise.

GES generate the same CPDAG structure with 4 (correctly) directed edges, and the Markov equivalence class of this CPDAG contains indeed the ground truth DAG in Figure 3(a). On the other hand, LiNGAM leaves no edge undirected and produces a DAG. Since the residuals of this experiment are the gaussian noises, LiNGAM is unable to perform correctly (Figure 3(e)). However, if the experiments are repeated using uniform sources of noise centered at zero, then LiNGAM is able to recover the causal structure[4], as shown in Figure 3(f). Importantly, the 3 undirected edges in Figures 3(b), 3(c) and 3(d) can not be directed by any of the three algorithms regardless of the number samples used. This occurs because it is possible to invert some of the arrows, e.g. $X_4 \to X_2 \to X_1$, and tune the weights of the noises and the edges so as to get the same joint probability distribution. As a consequence, the only way to direct the CPDAG undirected edges is by means of background knowledge. For instance, if it was known that X_1 preceded X_2 temporally, this constraint would force the algorithms to direct $X_1 \to X_2$ as well as $X_2 \to X_3$ because a v-structure would appear at X_2 otherwise.

DirectLiNGAM, however, could generate the correct skeleton as well as the correct directions of the edges successfully. This is possible because the first dataset satisfies exactly the assumptions for the applicability of LiNGAM. That is, functional relations between variables are linear, values are continuous, and the noise distribution is non-Gaussian (uniform).

3.2. Compas

The *Compas* dataset includes data from Broward County, Florida, initially compiled by ProPublica Angwin et al. (2016) and the goal is to predict the two-year violent recidivism. That is, whether a convicted individual would commit a violent crime in the following two years (1) or not (0). We consider race as sensitive feature. Five variables are used for the structural learning, namely: race, sex, age, priors and recidivism. Three tiers in the partial order for temporal priority are used: race, sex and age are defined in the first tier, priors is in the second tier and recidivism is defined in the third tier. When found to be mediator, variables age and sex are considered as redlining, whereas priors as explaining. As Compas dataset Angwin et al. (2016) contains mixed data, LiNGAM is not applied. Figure 4 shows the generated causal graphs for PC, FCI and GES. Figure 4(d) shows the SBCN for the protected group (non-white defendants)[5].

[4] Using the threshold $\alpha = 0.05$. Smaller values lead to extra false edges, e.g. $\alpha = 0.03$ detects $X_3 \to X_6$.

[5] Note that Figure 4(d) shows a subgraph of the generated SBCN graph of the Compas dataset including solely the causal paths between the protected group and the outcome.

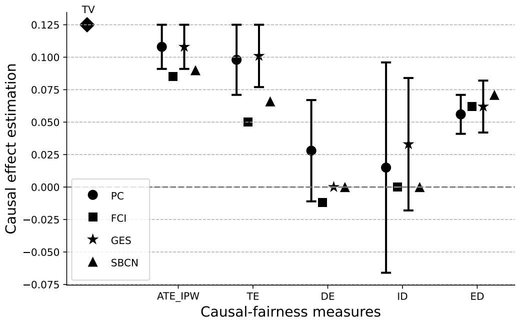


Figure 5: Estimation of causal effects of the Compas dataset.

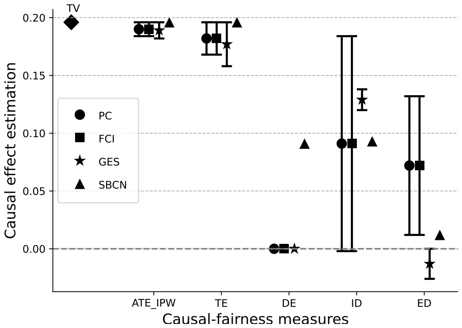
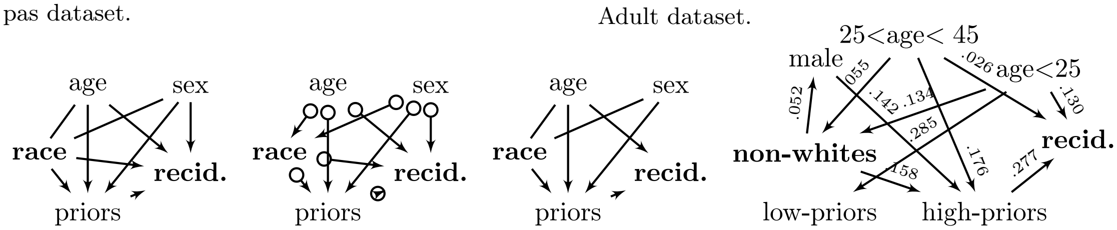


Figure 6: Estimation of causal effects of the Adult dataset.



(a) PC (b) FCI (c) GES (d) SBCN of non-whites

Figure 4: Generated causal graphs and SBCN for the Compas dataset. Recid. stands for recidivism.

It is important to mention that the obtained graphs for Compas dataset do not agree on the direct edge from the sensitive attribute (race) to the outcome variable (recidivism). There is such an edge according to PC and FCI, but not according to GES and SBCN. This is of crucial importance to fairness as the direct effect is always discriminatory.

Figure 5 shows the different discrimination measures using the different graphs. Both TE and ATE_IPW produce positive values which indicate a discrimination against non-white defendants. Considering the PC CPDAG (Figure 4(a)), the highest value of TE is obtained when there are no confounders (the two undirected edges are directed as $race \rightarrow age$ and $race \rightarrow sex$). In such graph, TE coincides with TV which is equal to 0.125. The same high value of TE is obtained with GES CPDAG (Figure 4(c)) when the undirected edge is directed as $race \rightarrow age$. In such no confounding case, the presence or absence of the direct edge $race \rightarrow recidivism$ does not matter for TE. The smallest value for TE (0.050) is only obtained in FCI PAG (Figure 4(b)) where both age and sex variables are confounders. This implies that the total effect is going through only two paths $race \rightarrow recidivism$ and $race \rightarrow priors \rightarrow recidivism$. Such low TE value cannot be obtained in PC nor in GES CPDAGs because the edges $age \rightarrow race$ and $sex \rightarrow race$ will create a new v-structure, and hence, lead to a causal graph outside the Markov equivalence class. ID is highest (0.096) with PC when both age and sex are mediators ($race \rightarrow age$ and $race \rightarrow sex$). This is inline with GES as ID is highest (0.084) with the same directions of the edges ($race \rightarrow age$ and $race \rightarrow sex$). This is expected as this yields two redlining paths ($race \rightarrow age \rightarrow recidivism$ and $race \rightarrow sex \rightarrow recidivism$). Surprisingly, when age is a confounder while sex remains a redlining, the indirect discrimination *against* blacks (0.096) becomes indirect discrimination *in favor* of blacks (−0.064). This is an example of Simpson's paradox Simpson (1951); Bickel et al. (1975) when conditioning on a variable changes significantly the statistical conclusions. In the case where the edges are directed as $race \rightarrow age$ and $sex \rightarrow race$, both PC and GES graphs produce the same ID value (−0.018). The case that leads the highest discrepancy in

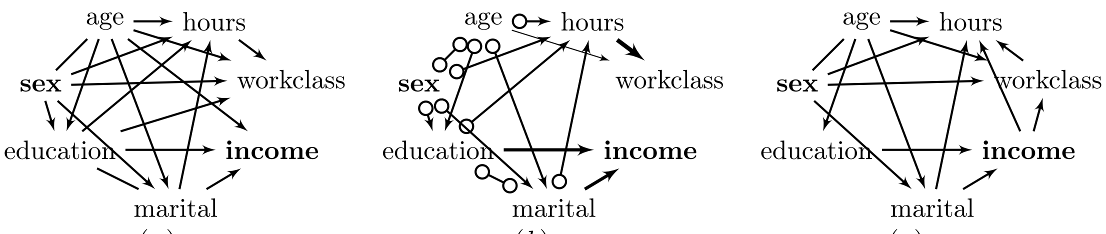
ID values between PC and GES is $age \rightarrow race$ and $race \rightarrow sex$ (age is a confounder and sex is a mediator). In such setup, according to PC, ID is lowest (-0.064) while according to GES, ID is zero as there are no redlining paths between race and recidivism. For GES, ID is highest when age is a mediator (redlining) and the indirect discrimination is conveyed through the path $race \rightarrow age \rightarrow recidivism$. If age is confounder ($age \rightarrow race$), there is no redlining variables, and hence ID is zero. It is important to mention here that if a causal path is going through redlining and explaining variables (e.g. $race \rightarrow sex \rightarrow priors \rightarrow recidivism$), it is considered as part of explained discrimination. The rule of thumb is that any path containing at least one explaining variable is considered as part of explained discrimination[6]. ID is zero for FCI and SBCN for the same reason (absence of redlining paths). ED values according to all graphs are comparable as all explained discrimination is going through the single explaining variable (priors). For ED, there are three possible paths from race to recidivism: $race \rightarrow priors \rightarrow recidivism$, $race \rightarrow sex \rightarrow priors \rightarrow recidivism$, and $race \rightarrow age \rightarrow priors \rightarrow recidivism$. For PC, all paths are possible. For FCI, only the first path is possible, while for the others, only the first two are possible. Therefore ED for GES and SBCN are equal. As a summary, estimating TE using graphs generated by different causal discovery algorithms can lead to a significant inconsistency ($0.125 - 0.050 = 0.075$) in assessing the amplitude of the discrimination against non-white defendants. Moreover, graphs generated by the same discovery algorithms (belong to the same Markov equivalence class), can lead to very different discrimination values (ID goes from a positive discrimination of 0.096 to a negative one (-0.064) due to reversing the direction of a single edge) which can be seen as a form of Simpson's paradox. Finally, the value of the threshold to decide about causal relations can have important consequences on fairness conclusion as well (missing $race \rightarrow recidivism$ edge in GES and SBCN).

3.3. Adult

The *Adult* dataset[7] consists of $32,561$ samples and the goal is to predict the income of individuals: $\leq 50K$ (negative outcome) or $> 50K$ (positive outcome). Only 7 variables are used for structural learning namely: age, sex, education level, marital status, work-class and number of working hours per week. Three tiers in the partial order for temporal priority are used: age and sex are defined in the first tier, education and marital status in the second tier, and work-class, number of working hours per week and the income are defined in the last tier. When found to be mediators, variables age and marital status are considered as redlining, whereas education as explaining. The causal graphs generated by PC, FCI and GES are shown in Figure 7. Figure 8 shows the SBCN for females. As in the Compas dataset, LiNGAM cannot be used as data is mixed as well.

[6] This interpretation can be justified by considering the simple path $race \rightarrow priors \rightarrow recidivism$. Such path is clearly part of explained discrimination as priors is explaining variable. However, it contains also a "redlining" variable which is the sensitive attribute race!

[7] https://archive.ics.uci.edu/ml/datasets/adult.



(*a*) PC (*b*) FCI (*c*) GES

Figure 7: Generated causal graphs for the Adult dataset.

There are two important notes about the generated graphs. First, only SBCN exhibits a direct edge between sex and income. Second, all remaining graphs have undirected edges (in particular, between sex and age). This leads to variability in the fairness measures as shown in Figure 6. For instance, although all TE and ATE_{IPW} values are positive which indicates a discrimination against females, there is some variability in the extent of this discrimination. The highest discrimination can be seen in the GES CPDAG (Figure 7(*c*)) where *sex* → *age* (age is a mediator) yields to $TE = 0.196$ whereas *age* → *sex* (age is a confounder) yields to $TE = 0.157$. DE is zero according to all graphs except for SBCN since it is the only one with a direct edge between sex and income. For PC and FCI graphs (having the same structure with two undecided edges), ID value ranges between -0.003 and 0.184 where the former is obtained with *age* → *sex* and *education* → *marital* and the latter is obtained with *sex* → *age* and *education* → *marital*. This is expected as *sex* → *age* opens an additional redlining path *sex* → *age* → *income*. In other words, having only one redlining path *sex* → *marital* → *income* shows a very small indirect discrimination *in favor* of females. Opening the other redlining path (through age) turns that into a clear indirect discrimination *against* females. A possible explanation is that young married women tend to have low income due to motherhood responsibilities, while older married women passed that part of their life and are more available for their professional careers. Notice that, the lowest value of ID in GES (0.119 obtained with *age* → *sex*) is significantly higher than the lowest ID value in PC and FCI (-0.003). The reason is that in GES, there is only one indirect path *sex* → *marital* → *recidivism* while in PC and FCI, there are three different paths (*sex* → *marital* → *income*, *sex* → *education* → *income*, and *sex* → *education* → *marital* → *income*). Hence, the causal effect between sex and income in GES is only conveyed through the redlining path. Whereas in PC and FCI, it is split between the redlining path and also the two other explained discrimination paths. For ED, the highest value (0.132) is obtained in PC and FCI when age is a confounder (*age* → *sex*) and marital status is a mediator between education and income (*education* → *marital*). The smallest value (-0.027) is obtained in GES when age is a mediator (*sex* → *age*) which indicates a small explained discrimination *in favor* of females through the path *sex* → *age* → *education* → *income*. This path is only possible as a single explaining path in GES CPDAG. In all the graphs obtained by the other algorithms, such path is possible but along other explaining paths, in particular, *sex* → *education* → *income*. This explains why the discrimination in favor of females is only observable with GES. It is interesting to notice that in PC and FCI graphs, the explained discrimination through *sex* → *education* → *income* is slightly positive (0.016) whereas in GES graph, adding another mediator *sex* → *age* → *education* → *income* yields a slightly negative explained discrimination. As there is no overlap between the ranges of ED values in PC and FCI graphs on one hand and GES on another, and that values (although small) have different signs (positive vs negative), the explained discrimination conclusions depend on which algorithm is used to discover causal relations.

Compared to Compas dataset, the mediation analysis on adult dataset reveals two additional fairness relevant observations. First, a specific causal path can be discovered by several causal discovery algorithms. However, the causal effect through that path may significantly differ depending on the presence of other causal paths not necessarily with the same interpretation (redlining or explaining paths). Second, even with the same causal path (e.g. *sex* → *education* → *income*), considering a mediator (e.g. age) can reverse the type of the discrimination (e.g. *sex* → *age* → *education* → *income*).

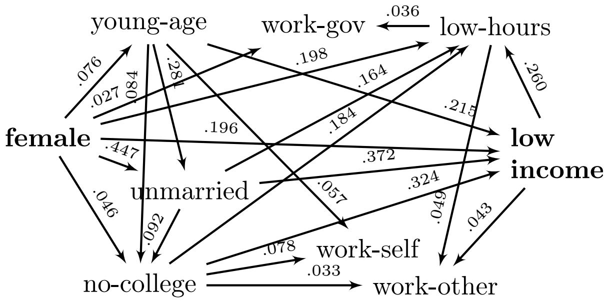


Figure 8: SBCN of Females in the Adult dataset.

There is a significant variability in the values of ED for PC and FCI because there are three possible paths for explaining discrimination that depends on the direction of the undirected edges, namely, *sex* → *education* → *income*, *sex* → *education* → *marital* → *income*, *sex* → *age* → *education* → *income*, and *sex* → *age* → *education* → *marital* → *income*. For GES, ED is either zero (when age is a confounder) or indicates a small discrimination through a single path, *sex* → *age* → *education* → *income*. Note that the edge *education* − *marital* in PC and FCI graphs should be oriented as *education* → *marital*. Otherwise, a new collider (*education*) will be created.

4. Conclusion

The main contributions of the paper are two-fold. First, we show how the subtle differences between the causal discovery algorithms can explain why they generate different causal graphs. Second, and foremost, we demonstrate how slight differences between causal graphs may have significant impact on fairness/discrimination conclusions. Most of the causal approaches to fairness in the literature do not tackle the causal graph generation task. With this study, we hope to raise the awareness about the importance of this step in the fairness assessment and enforcement pipeline as any difference in the structure of the graph may lead to very different fairness conclusions. A natural follow-up work after this study is to design a new causal discovery algorithm specifically tuned for fairness. It looks promising, for instance, to focus around the sensitive attributes by using ideas from local discovery algorithms Gupta et al. (2022). Another future direction would be to study the impact of pre-processing transformations on the structure of the generated graph and consequently on the fairness conclusions.

Acknowledgments

This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21, 2017.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Richard B Darlington. Another look at "cultural fairness". *Journal of educational measurement*, 8(2):71–82, 1971.

Dorit Dor and Michael Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*, 1992.

Juan Gamella. Greedy equivalence search (ges) algorithm for causal discovery. https://github.com/juangamella/ges, 2021. Accessed: 2022-03-16.

Shantanu Gupta, David Childers, and Zachary Chase Lipton. Local causal discovery for estimating causal effects. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, Spain, 2016.

Dominique MA Haughton. On the choice of a model to fit data from an exponential family. *The annals of statistics*, pages 342–355, 1988.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012. URL https://jmlr.org/papers/v13/hauser12a.html.

Christopher Hitchcock. Probabilistic causation. *Stanford Encylopedia of Philosophy (archive)*, 2002.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan): 111–152, 2013.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020a.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020b.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58 (5):102642, 2021.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Identifiability of causal-based fairness notions: A state of the art. *arXiv preprint arXiv:2203.05900*, 2022.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420, 2001.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Peter Spirtes, Christopher Meek, and Thomas Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:211–252, 1999.

Patrick Suppes. A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4), 1973.

Xiang Zhou and Teppei Yamamoto. Tracing causal paths from experimental and observational data. *SocArXiv. January*, 11, 2020.