Identifiability of Causal-based Fairness Notions: A State of the Art

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi

Identifiability of Causal-based Fairness Notions: A State of the Art

Karima Makhlouf
Inria, École Polytechnique, IPP
Paris, France
makhlouf@lix.polytechnique.fr

Sami Zhioua
Inria, École Polytechnique, IPP
Paris, France
zhioua@lix.polytechnique.fr

Catuscia Palamidessi
Inria, École Polytechnique, IPP
Paris, France
catuscia@lix.polytechnique.fr

ABSTRACT

Machine learning algorithms can produce biased outcome/prediction, typically, against minorities and under-represented sub-populations. Therefore, fairness is emerging as an important requirement for the large scale application of machine learning based technologies. The most commonly used fairness notions (e.g. statistical parity, equalized odds, predictive parity, etc.) are observational and rely on mere correlation between variables. These notions fail to identify bias in case of statistical anomalies such as Simpson's or Berkson's paradoxes. Causality-based fairness notions (e.g. counterfactual fairness, no-proxy discrimination, etc.) are immune to such anomalies and hence more reliable to assess fairness. The problem of causality-based fairness notions, however, is that they are defined in terms of quantities (e.g. causal, counterfactual, and path-specific effects) that are not always measurable. This is known as the identifiability problem and is the topic of a large body of work in the causal inference literature. This paper is a compilation of the major identifiability results which are of particular relevance for machine learning fairness. The results are illustrated using a large number of examples and causal graphs. The paper would be of particular interest to fairness researchers, practitioners, and policy makers who are considering the use of causality-based fairness notions as it summarizes and illustrates the major identifiability results.

KEYWORDS

Fairness, machine learning, causal-based, identifiability.

1 INTRODUCTION

Machine learning is being used to inform decisions with critical consequences on human lives such as job hiring, college admission, loan granting, and criminal risk assessment. Unfortunately, these automated decision systems have been found to consistently discriminate against certain individuals or sub-populations, typically minorities. Because the discrimination is very often unintentional, discovering and addressing it is a challenging task. The most commonly used fairness notions are observational and rely on mere correlation between variables. For example, statistical parity [5] requires that the proportion of positive outcome (e.g. granting loans) is the same for all sub-populations (e.g. male and female groups). Equal opportunity [8] requires that the true positive rate (TPR) is the same for all sub-populations. The main problem of correlation-based fairness notions is that they fail to detect discrimination in presence of statistical anomalies such as Simpson's paradox [27] and Berkson's paradox [2, 11]. A famous example of the Simpson's paradox is the gender bias in 1973 Berkley admission [3, 13]. In that year, 44% of male applicants were admitted against only 34% of female applicants. While this looks like a bias against female candidates, when the same data has been analyzed by department, acceptance rates were approximately the same.

One way to address this limitation is to consider how data is generated in the first place which leads to causal-based fairness notions. Because this new breed of fairness notions is immune to statistical paradoxes, it is now widely accepted that causality is necessary to appropriately address the problem of fairness [13]. Examples of causal-based fairness notions include total effect [18], interventional fairness [21], counterfactual fairness [12], counterfactual effects [36], and path-specific counterfactual fairness [4, 35]. These notions are defined in terms of non-observable quantities such as causal, counterfactual, and path-specific effects. As they are non-observable, these quantities cannot always be estimated based on observable data. This is known as the *identifiability* problem and is the topic of a large body of work in the causal inference literature. For example, the identifiability of causal effects can be decided using a set of three causal inference rules called do-calculus [16, 18]. Based on the do-calculus, Shpitser and Pearl [24] proposed a complete identification algorithm for causal effects. The algorithm (ID) was independently shown to be complete by Shpitser and Pearl [24] and Huang and Valtorta [10]. Using the do-calculus for identifiability has two main issues. First, it is typically a manual process. Second, it is not clear in which order the rules should be applied [31]. On the other hand, using the ID algorithm[1], can produce unnecessarily complex expressions that can lead to inefficient, and even biased, estimates when data is missing feature values [33]. A more intuitive alternative for deciding about identifiability is to rely on graphical criteria, that is, recognizing common graph structures that produce identifiable effects. Graphical criteria is an efficient and intuitive approach to the identifiability of all types of effects (causal, counterfactual, and path-specific) and is more easy to use than the do-calculus or the identifiability algorithms (e.g. ID, ID*, etc.).

This paper summarizes the main identifiability results as they relate to the specific problem of discrimination discovery with an emphasis on graphical criteria. These results fall into the following categories: causal effect (intervention) identifiability [7, 9, 18, 24, 26, 28–30], counterfactual identifiability [22, 25, 26, 34], direct/indirect identifiability [17], and path-specific effect identifiability [1, 14, 22, 37]. Section 2 provides necessary background concepts. Then, instead of repeating the definition of identifiability (Definition 3.2.3 in [18]), Section 3 gives an intuitive explanation of the identifiability problem through the teacher firing example. Sections 4, 5, and 6 compile the common identifiability results of causal, counterfactual, and path-specific effects, respectively.

[1]Implemented in the **causaleffect** R package [32].

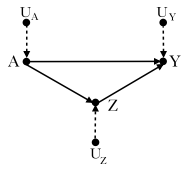
2 PRELIMINARIES AND NOTATION

Variables are denoted by capital letters. In particular, A is used for the sensitive variable (e.g., gender, race, age) and Y is used for the outcome of the automated decision system (e.g., hiring, admission, releasing on parole). Small letters denote specific values of variables (e.g., $A = a'$, $W = w$). Bold capital and small letters denote a set of variables and a set of values, respectively.

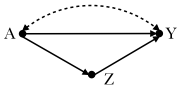
A structural causal model [18] is a tuple $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where:

- $\mathbf{U}$ is a set of exogenous variables which cannot be observed or experimented on but constitute the background knowledge behind the model.
- $\mathbf{V}$ is a set of observable variables which can be experimented on.
- $\mathbf{F}$ is a set of structural functions where each f_i is mapping $\mathbf{U} \cup \mathbf{V} \rightarrow \mathbf{V} \backslash \{V_i\}$ which represents the process by which variable V_i changes in response to other variables in $\mathbf{U} \cup \mathbf{V}$.
- $P(\mathbf{u})$ is a probability distribution over the unobservable variables $\mathbf{U}$.

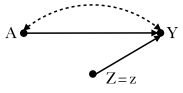
Causal assumptions between variables are captured by a causal diagram G which is a directed acyclic graph (DAG) where nodes represent variables and directed edges represent functional relationships between the variables. Directed edges can have two interpretations. A probabilistic interpretation where the edge represents a dependency among the variables such that the direction of the edge is irrelevant. A causal interpretation where the edge represents a causal influence between the corresponding variables such that the direction of the edge matters. Unobserved variables $\mathbf{U}$, which are typically not represented in the causal diagram, can be either mutually independent (Markovian model) or dependent from each others. In case the unobserved variables can be dependent and each $U_i \in \mathbf{U}$ is used in at most two functions in F, the model is called semi-Markovian. In causal diagrams of semi-Markovian models, dependent unobservable variables (unobserved confounders) are represented by a dotted bi-directed edge between observable variables. Graphs $G5$ (Table 1) and $G16$ (Table 2) show causal graphs of Markovian and semi-Markovian models, respectively.



(a) Causal graph of a Markovian model.



(b) Causal graph of a semi-Markovian model.



(c) Causal graph of a semi-Markovian model after intervention: $do(Z = z)$.

Figure 1

An intervention, noted $do(V = v)$, is a manipulation of the model that consists in fixing the value of a variable (or a set of variables) to a specific value regardless of the corresponding function f_v. Graphically, it consists in discarding all edges incident to the node corresponding to variable V. Figure 1(c) shows the causal diagram of the manipulated model after intervention $do(Z = z)$ denoted $M_{Z=z}$ or M_z for short. The intervention $do(V = v)$ induces a different distribution on the other variables. For example, in Figure 1(c), $do(Z = z)$ results in a different distribution on Y, namely, $P(Y|do(Z = z))$. Intuitively, while $P(Y|Z = z)$ reflects the population distribution of Y among individuals whose Z value is z, $P(Y|do(Z = z)$ reflects the population distribution of Y if *everyone in the population* had their Z value fixed at z. The obtained distribution $P(Y|do(Z = z)$ can be considered as a *counterfactual* distribution since the intervention forces Z to take a value different from the one it would take in the actual world. Such counterfactual variable is noted $Y_{Z=z}$ or Y_z for short[2]. $P(Y = y|do(Z = z)) = P(Y_{Z=z} = y) = P(Y_z = y) = P(y_z)$ is used to define the causal effect of z on Y. The term counterfactual quantity is used for expressions that involve explicitly multiple worlds. In Figure 1(b), consider the expression $P(y_{a'}|Y = y, A = a) = P(y_{a'}|y, a)$. Such expression involves two worlds: an observed world where $A = a$ and $Y = y$ and a counterfactual world where $Y = y$ and $A = a'$ and it reads "the probability of $Y = y$ had A been a' given that we observed $Y = y$ and $A = a$". In the common example of job hiring, if A denotes race (a :white, a':non-white) and Y denotes the hiring decision (y:hired, y':not hired), $P(y_{a'}|y, a)$ reads "given that a white applicant has been hired, what is the probability that the same applicant is still being hired had he been non-white". Nesting counterfactuals can produce complex expressions. For example, in the relatively simple model of Figure 1(b), $P(y_{a,z_{a'}}|y'_{a'}) = P(y(a, z(a'))|y'(a'))$ reads the probability of $Y = y$ had (1) A been a' and (2) Z been z when A is a', given that an intervention $A = a'$ produced y'. This expression involves three worlds: a world where $A = a$, a world where $Z = z_{a'}$, and a world where $A = a'$. Such complex expressions are used to characterize direct, indirect, and path-specific effects.

3 EXPLAINING IDENTIFIABILITY THROUGH AN EXAMPLE

Consider the example of an automated system for deciding whether to fire a teacher at the end of the academic year. Deployed teacher evaluation systems have been suspected of bias in the past. For example, IMPACT is a teacher evaluation system used in the city of Washington, D.C., and have been found to be unfair against teachers from minority groups [15, 19, 20]. Assume that the system takes as input one feature, namely, the initial[3] average level of the students assigned to that teacher (A). The outcome is whether to fire the teacher ($\hat{Y}$). Assume that these two variables are confounded by a third unobservable variable U which represents a socioeconomic status related to the school neighborhood.

Assume also that all 3 variables are binary with the following values: If the initial average level of the students assigned to the teacher is high, $A = 1$, otherwise (initial level is low), $A = 0$. Firing a teacher corresponds to $\hat{Y} = 1$, while retaining her corresponds to $\hat{Y} = 0$. If the school is located in a high-income neighborhood, $U = 1$, otherwise (the school is located in a low-income neighborhood), $U = 0$. The level of students in a given class can be influenced by several variables, but in this example, assume that it is only influenced by the socioeconomic status of the school; students in

[2]The notations $Y_{Z \leftarrow z}$ and $Y(z)$ are used in the literature as well.
[3]At the beginning of the academic year.

high-income neighborhoods are more advantaged and typically perform better in school.

The relationships between the variables A, U, and Y can be graphically represented using the causal directed acyclic graph (DAG) in Figure 2[4]. Notice that the edges $U \rightarrow A$ and $U \rightarrow Y$ are dotted because they are emanating from an unobservable variable (U).

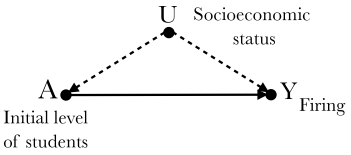


Figure 2: Causal graph of the teacher firing example.

Assume that the automated decision system is suspected to be biased by the level of students assigned to the teacher. That is, it is claimed that the system is more likely to fire teachers who have been assigned classes with low level students at the beginning of the academic year, which is clearly unfair. The bias in the outcome ($\hat{Y}$) due to the sensitive variable A can be assessed by computing the total variation:

$$TV_{a_1,a_0}(y) = P(y \mid a_1) - P(y \mid a_0) \tag{1}$$

which coincides with statistical parity [5] and measures the difference between the conditional distributions of $\hat{Y}$ when we (passively) observe A changing from a_0 to a_1 (e.g. from 0 to 1 in our example). The main limitation of TV is that it is purely statistical and may be fooled by statistical anomalies such as Simpson's and Berkson's paradoxes. Total effect (TE) [18] is the causal version of TV and is defined in terms of experimental probabilities as follows:

$$\begin{aligned} TE_{a_1,a_0}(y) &= P(Y = y|do(A = a_1)) - P(Y = y|do(A = a_0)) \\ &= P(y_{a_1}) - P(y_{a_0}) \end{aligned} \tag{2}$$

While TV is expressed in terms of observable probabilites ($P(y|a_1)$ and $P(y|a_0)$) and hence can always be computed from observable data, TE is not. The question is can TE be expressed in terms of observable probabilities and hence computed from observable data? If the answer is yes, TE is said to be identifiable. Otherwise, it is not identifiable. Pearl gives a formal definition of identifiability [18], Page 77, Definition 3.2.3. Intuitively, given a dataset D (which can be generated by different causal models), a quantity (e.g. $P(Y_{A=a_1} = y)$) is identifiable if it keeps the same value regardless of the causal model which generated the dataset D. For example, in the teacher firing scenario, $P(\hat{Y}_{A=0} = 1)$ is not identifiable since it is possible to come up with two causal models that can generate the same data, and hence $P(\hat{Y}_{A=0} = 1)$ cannot be uniquely computed based only on observable data. For illustration, consider the two following causal models ($\mathbf{M_1}$ and $\mathbf{M_2}$) expressed in terms of all three variables A, $\hat{Y}$, and U[5]:

Causal model $\mathbf{M_1}$

$P(\hat{Y} = 1 \mid A = 0, U = 0) = 0.25$	$P(A = 0 \mid U = 0) = 0.6$
$P(\hat{Y} = 1 \mid A = 1, U = 1) = 0.25$	$P(A = 0 \mid U = 1) = 0.4$
$P(\hat{Y} = 1 \mid A = 0, U = 1) = 0.02$	$P(A = 1 \mid U = 0) = 0.4$
$P(\hat{Y} = 1 \mid A = 1, U = 0) = 0.02$	$P(A = 1 \mid U = 1) = 0.6$

Causal model $\mathbf{M_2}$

$P(\hat{Y} = 1 \mid A = 0, U = 0) = 0.24$	$P(A = 0 \mid U = 0) = 0.65$
$P(\hat{Y} = 1 \mid A = 1, U = 1) = 0.24$	$P(A = 0 \mid U = 1) = 0.35$
$P(\hat{Y} = 1 \mid A = 0, U = 1) = 0.01$	$P(A = 1 \mid U = 0) = 0.35$
$P(\hat{Y} = 1 \mid A = 1, U = 0) = 0.01$	$P(A = 1 \mid U = 1) = 0.65$

It is easy to show that both causal models generate the same joint distribution $P(\hat{Y}, A)$. Using the chain rule,

$$P(A, \hat{Y}) = \sum_{u \in \{0,1\}} P(\hat{Y}|\ A, U = u)P(A|\ U = u)P(U = u) \tag{3}$$

both $\mathbf{M_1}$ and $\mathbf{M_2}$ generate the same observable distribution:

$$\begin{aligned} P(\hat{Y} = 1\ ,\ A = 1) &= 0.079 \\ P(\hat{Y} = 0\ ,\ A = 0) &= 0.42 \\ P(\hat{Y} = 1\ ,\ A = 0) &= 0.079 \\ P(\hat{Y} = 0\ ,\ A = 1) &= 0.42 \end{aligned}$$

$P(\hat{Y}_{A=0} = 1)$ is not an observable quantity. However, since we assumed the existence of an oracle with knowledge about all model parameters, it can be computed using the back-door formula (Equation 6) as follows:

$$P(\hat{Y}_{A=0} = 1) = \sum_{u \in \{0,1\}} P(\hat{Y} = 1|\ A = 0, U = u)P(U = u) \tag{4}$$

For causal model $\mathbf{M_1}$,

$$P(\hat{Y}_{A=0} = 1) = (0.25 \times 0.5) + (0.02 \times 0.5) = 0.135$$

whereas for causal model $\mathbf{M_2}$,

$$P(\hat{Y}_{A=0} = 1) = (0.24 \times 0.5) + (0.01 \times 0.5) = 0.125$$

Hence, $\mathbf{M_1}$ and $\mathbf{M_2}$ are two different causal models that generate the same observable data but yield two different values for the quantity $P(\hat{Y}_{A=0} = 1)$ which is consequently not identifiable from observational data. In other words, in this situation, it is not possible to use observations to tell whether A is actually a cause of $\hat{Y}$.

Since total variation TV is defined in terms of observable probabilities, it can be computed based on the observable data. Total effect TE, however, cannot be computed based on observable data as $P(\hat{Y}_{A=0} = 1)$ is not identifiable.

Notice that, in this example, both models $\mathbf{M_1}$ and $\mathbf{M_2}$ share the same graph structure (Figure 2). This is not always the case. That is, it is possible to have two causal models with different graph structures coinciding on the observable joint distribution. Hardt et al. [8] illustrate this case with an example. Tikka [31] presents another non-identifiable example defined using the XOR logic operator.

Based on the causal inference literature, the next sections compile a list of identifiability criteria for the different types of non-observable quantities: causal, counterfactual, direct, indirect, and path-specific effects.

[4]The structure of this graph is known as the bow structure in the literature.

[5]U is an exogenous variable and it is not observable by definition. But to illustrate the identifiability concept, assume there is an oracle with a knowledge about all model parameters including U.

4 IDENTIFIABILITY OF CAUSAL EFFECTS

The natural way to estimate the causal effect of a variable (the sensitive attribute A) on another (the outcome variable Y) is to carry out real experiments using RCT (Randomized Controlled Trial) [6]. If possible, RCT drops the need for identiability altogether. However, in the context of machine learning fairness, RCT is often not an option as experiments can be too costly to implement or physically impossible to carry out (e.g. changing the gender of a job applicant).

As an alternative, intervention using the do-operator can be used to compute the causal effect. Without loss of generality, this section focuses on the identifiability of $P(Y = y|do(A = a)) = P(y_a)$, that is, the causal effect of the sensitive attribute A on the outcome variable Y. The computation of $P(y_a)$ uses a "surgically altered" graph in which all arrows into A are deleted and the value of A is fixed at a, but the rest of the graph remains unchanged.

Whether it is possible to express $P(y_a)$ only in terms of observable probabilities (identifiability) depends on the structure of the causal graph (which captures how data is generated). A first important result is that any causal effect is identifiable in a Markovian model (where all unobservable variables are independent). In semi-Markovian models, however, the causal effect is not always identifiable.

4.1 Identifiability in Markovian models

Table 1 shows different Markovian models involving various patterns of causal relationships along with the corresponding expression in terms of observable probabilities.

Graphs $G1 - G5$ illustrate the simplest cases where no confounding between A and Y exists. In that case, the causal effect matches the conditional probability regardless of any mediator M as follows:

$$P(y_a) = P(y|a) \tag{5}$$

4.1.1 Back-door adjustment. In case there are confounders involving A and Y, the causal effect can be identified by finding a set of variables C that block all back-door paths from A to Y. This is called the back-door criterion[6]. This criterion necessitates the existence of a set of covariates C which blocks all the indirect paths from A to Y, but keeps all the direct paths open. C satisfies the back-door criterion when (1) C blocks every back-door path between A and Y, and (2) no node in C is a descendant of A. Graphs $G6 - 12$ illustrate examples where C (or $\{C_1,C_2\}$) meets the back-door criterion. In presence of an observable confounder C, $P(y_a)$ is identifiable by adjusting[7] on that confounder using back-door formula:

$$P(y_a) = \sum_C P(y|a, c)\ P(c) \tag{6}$$

where the summation is on values c in the domain (sample space) of C denoted as $dom(C)$. Note that $G4$ and $G5$ contain a collider (W). Marginalizing over the collider variable disproves the equality in Eq. 6 as it might open back-door paths between A and Y and consquently create a dependency between these two variables.

Despite the fact that $G11$ involves two confounders C_1 and C_2, no adjustment is required because of the presence of the collider W. Hence $P(y_a)$ can be computed using Eq. 5. Alternatively, controlling on: C_1, C_2, $\{W, C_1\}$,$\{W, C_2\}$ or $\{W, C_1, C_2\}$ is possible using Eq. 6.

[6]Called also adjustment formula or stratification.

[7]The terms adjusting, controlling, and marginalizing are used interchangeably.

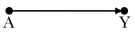
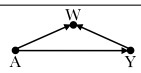
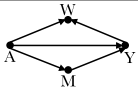
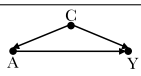
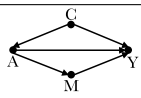
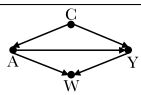
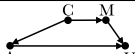
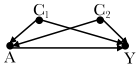
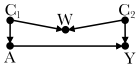
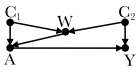
	Causal graph	$P(y_a)$
$G1$		$P(y\|a)$
$G2$		
$G3$		
$G4$		
$G5$		
$G6$		$\sum_C P(y\|a, c)\ P(c)$
$G7$		
$G8$		
$G9$		
$G10$		$\sum_{C_1C_2} P(y\|a, c_1, c_2)\ P(c_1, c_2)$
$G11$		$P(y\|a)$ $\sum_{C_1} P(y\|a, c_1)\ P(c_1)$ $\sum_{C_2} P(y\|a, c_2)\ P(c_2)$ $\sum_{WC_1} P(y\|a, w, c_1)\ P(w, c_1)$ $\sum_{WC_2} P(y\|a, w, c_2)\ P(w, c_2)$ $\sum_{WC_1C_2} P(y\|a, w, c_1, c_2)\ P(w, c_1, c_2)$
$G12$		$\sum_{C_2} P(y\|a, c_2)\ P(c_2)$ $\sum_{WC_1} P(y\|a, w, c_1)\ P(w, c_1)$ $\sum_{WC_2} P(y\|a, w, c_2)\ P(w, c_2)$ $\sum_{WC_1C_2} P(y\|a, w, c_1, c_2)\ P(w, c_1, c_2)$

Table 1: $P(y_a)$ of some Markovian models.

Table 1 shows all possible formulas that can be used to calculate $P(y_a)$ for $G11$. $G12$ presents another case with two confounders ($C1$ and C_2) and the two following back-door paths between A and Y: $A \leftarrow W \leftarrow C_2 \rightarrow Y$ and $A \leftarrow C_1 \rightarrow W \leftarrow C_2 \rightarrow Y$. The former

must be blocked by either W or C_2 or both while the latter doesn't need any controlling because of the presence of the collider: W. Thus, the set of variables sufficient to control for confounding are: C_2, $\{C_1, W\}$, $\{W, C_2\}$ or $\{W, C_1, C_2\}$ but not W or C_1 (the minimum to control for is: C_2). That is, any one of these equations can be used to calculate the causal effect of A on Y.

4.1.2 Truncated factorization formula. An alternative way to measure the causal effect $P(y_a)$ in Makovian models is to use the truncated factorization formula [18]:

$$P(y_a) = \sum_{\substack{\mathbf{V}\setminus\{A,Y\} \\ Y=y}} \prod_{V\in\mathbf{V}\setminus\{A\}} P(v|\mathbf{Pa}_V) \tag{7}$$

where $\mathbf{Pa}_V$ denotes the parent variables of V. For instance, applying the truncated factorization formula on $G12$ leads to the following equality:

$$P(y_a) = \sum_{WC_1C_2} P(y|\ a, c_2)\ P(w|\ c_1, c_2)\ P(c_1)\ P(c_2) \tag{8}$$

Note that Eq. 6 and the last result of applying the back-door criterion for $G12$ in Table 1 are equivalent. This can be easily demonstrated as follows:

$$\begin{aligned} &\sum_{WC_1C_2} P(y|a, w, c_1, c_2)\ P(w, c_1, c_2) \\ &= \sum_{WC_1C_2} P(y|a, w, c_1, c_2)\ P(w|\ c_1, c_2)P(c_1)\ P(c_2) &(9)\\ &= \sum_{WC_1C_2} P(y|a, c_2)\ P(w|\ c_1, c_2)P(c_1)\ P(c_2) &(10)\end{aligned}$$

$P(y|a, w, c_1, c_2)$ in (9) is replaced by $P(y|a, c_2)$ in (10) due to the fact that $Y \perp\!\!\!\perp W|\ A$[8] and $Y \perp\!\!\!\perp C_1|\ A$.

As a summary, the only type of variables that have an impact on the identifiability of $P(y_a)$ in Markovian models is the confounder. To compute the causal effect in presence of confounding, adjusting using the back-door formula (Eq. 6) is required. However, adjusting should not be used in presence of a collider variable since this might open back-door paths between A and Y and hence, create a dependency between them. Mediator variables, on the other hand, have no impact on the identifiability of causal effects in Markovian models.

4.2 Identifiable semi-Markovian models

Causal effects are not always identifiable in semi-Markovian models. This subsection focuses on causal models where the causal effect of A on Y is identifiable. The following subsection gives a graphical criteria of causal models where the causal effect is not identifiable.

In the causal model, the measurement of causal effects is assisted by interventions following a set of inference rules introduced by Pearl [18] known as: *do-calculus*. These rules tend to link the interventional quantities of causal effects to simple statistical distributions based solely on observational data. As an alternative way of assessing causal effects, relevant graphical patterns will be presented in the remainder of this section.

[8] Y and W are independent given A.

4.2.1 do-calculus inference rules. do-calculus [16, 18] is a set of three inference rules that can be used to express an interventional expression of the form $P(y_a)$ in terms of subscript-free (observable) quantities. The rules are:

- **Rule 1 (Insertion/Deletion of Observations):** $P(y_a|\mathbf{c}, w) = P(y_a|\mathbf{c})$ provided that the set of variables $\mathbf{C}$ blocks all back-door paths from W to Y after all arrows leading to A have been deleted.
- **Rule 2 (Action/Observation Exchange):** $P(y_a|\mathbf{c}) = P(y|a, \mathbf{c})$ provided that the set of variables $\mathbf{C}$ blocks all back-door paths from A to Y.
- **Rule 3 (Insertion/Deletion of Actions):** $P(y_a) = P(y)$ provided that there are no causal paths between A and Y.

Although the do-calculus is proven to be complete for identifying causal effects[9] [10, 24], the completeness is not immediately apparent from the rules themselves. The other issue is that do-calculus is typically used manually and hence it is not obvious in which order the rules should be used to reach the subscript-free expression. An example of using the do-calculus is detailed in [18] Section 3.4.3. Deciding about the identifiability of causal effect is not easy with the do-calculus. A more intuitive approach would be to use graphical criteria. The rest of the subsection lists the most common graphical criteria for the identifiability of causal effects in semi-Markovian models.

4.2.2 Graphical criteria. The simplest case where the causal effect of A on Y is identifiable in semi-Markovian models is when A is not connected to any unobserved confounder, that is, no bi-directed edge is connected to A. Graphs $G13$ and $G14$ in Table 2 satisfy this criterion. Now, depending on the existence (or absence) of back-door paths connecting A to Y, the calculation of the causal effect varies. Then, in case all the pathways connecting A to Y are front-door from A (start with an outgoing edge from A), the causal effect coincides with the conditional probability (Eq. 5). $G13$ illustrates an example of such situation. On the other hand, in case A is connected to some observed confounders (there is a pathway from A to Y staring with an edge into A), the back-door formula (Eq. 6) is needed to compute the causal effect. $G14$ presents such pattern. This matches Theorem 3.2.5 in [18] which states that if all parents of a cause variable A are observable, the causal effect of that variable is identifiable. Hence, the back-door formula can be generalized as follows (Theorem 3.2.5 [18]):

$$P(y_a) = \sum_{pa_A} P(y|a, \mathbf{pa}_A)\ P(\mathbf{pa}_A) \tag{11}$$

where $\mathbf{pa}_A$ is the set of values of the parents of A. $G16$ is a more complex causal model that satisfies Eq. 11 where the causal effect of A on Y is identifiable and can be computed as:

$$P(y_a) = \sum_{c_1,c_2} P(y|a, c_1, c_2)\ P(c_1, c_2)$$

Note that, in $G15$, despite the fact that A is involved in confounding, Eq. 6 also applies as C blocks all the back-door paths including the unobserved one: $A \leftrightarrow C$.

[9] If an interventional expression cannot be converted into subscript-free quantity, it means the expression is not identifiable.

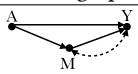
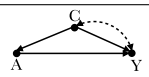
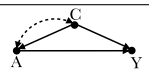
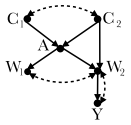
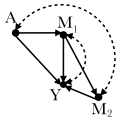
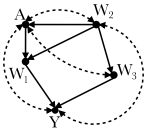
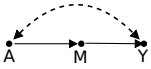
	Causal graph	$P(y_a)$
$G13$		$P(y\|a)$
$G14$		$\sum_C P(y\|a,c)\ P(c)$
$G15$		
$G16$		$\sum_{c_1,c_2} P(y\|a,c_1,c_2)\ P(c_1,c_2)$
$G17$		$\sum_{m_1,m_2}\ P(y\|m_1,m_2,a)\ P(m_1\|a)$ $\times \sum_{a'}\ P(m_2\|m_1,a')\ P(a')$
$G18$		$\sum_{w_1}\sum_{w_2}\sum_{a'}\ P(y\|w_1,w_2,a')$ $\times P(a'\|w_2)\ P(w_1\|w_2,a)P(w_2)$
$G19$		$\sum_m\ P(m\|a)\ \sum_{a'} P(y\|m,a')\ P(a')$

Table 2: $P(y_a)$ of some semi-Markovian models.

$G17$ and $G18$ are more complex causal models, but still identifiable due to a more specific graphical criterion: A is not connected through bi-directed and dashed paths to any of its children that are at the same time ancestors of Y. Under such criterion, the causal effect of a single variable A on all the other variables $\mathbf{V}\backslash\{A\}$ in the model denoted as $P(\mathbf{v}_a) = P_a(\mathbf{v})$ is identifiable and is given by Theorem 2 [29]:

$$P_a(\mathbf{v}) = \left(\prod_{i|V_i\in \mathbf{ch_A}} P(v_i|\mathbf{pa}_i)\right) \sum_{a'\in dom(A)} \frac{P(v)}{\prod_{i|V_i\in \mathbf{ch_A}} P(v_i|\mathbf{pa}_i)} \quad (12)$$

where ch_A is the set of the children of the node A while $\mathbf{pa_i}$ is the set of values of the parents of the variable V_i.

Now, since our aim is to assess the effect of the sensitive attribute A on a single variable (the outcome Y), Eq. 12 should be adapted. To illustrate that, consider the example of the causal graph $G18$. Applying Eq. 12 to $G18$ leads to:

$$P_a(w_1,w_2,w_3,y) = P(w_1|\ a,w_2) \sum_{a'\in dom(A)} \frac{P(w_1,w_2,w_3,y)}{P(w_1|\ a',w_2)}) \quad (13)$$

$$= P(w_1|\ a,w_2) \sum_{a'\in dom(A)} P(y,w_3|\ a',w_1,w_2) \times P(a',w_2) \quad (14)$$

where Eq. 14 is obtained by applying the Bayes' rule. Adapting Eq. 12 in order to compute $P(y_a)$ (causal effect on the single variable Y) requires summing over the possible values of variables W_1, W_2 and W_3 as follows. Starting from Eq. 14, summing over W_1 gives:

$$P_a(w_2,w_3,y) = \sum_{w_1'\in dom(W_1)} P(w_1'|\ a,w_2) \times \sum_{a'\in dom(A)} P(y,w_3|\ a',w_1',w_2)P(a',w_2)$$

Similarly, summing over W_2 and W_3 (and omitting $dom()$ for conciseness), leads to:

$$P(y_a) = \sum_{w_1'}\sum_{w_2'}\sum_{a'} P(y|w_1',w_2',a')\ P(a'|w_2')\ P(w_1'|w_2',a)\ P(w_2') \quad (15)$$

Note that W_3 is omitted from Eq. 15 as this variable is not concerned by the causal effect of A on Y [18].

Pearl [18] obtained the same result (Eq. 15) using *do-calculus* (Section 4.2.1). Thus,

$$P(y_a) = \sum_{w_1}\sum_{w_2}\sum_{w_3} P(y|w_1,w_3)\ P(w_2)\ P(w_1,y|\ do(a)) \quad (16)$$

$$= \sum_{w_1}\sum_{w_2} P(y|w_1,w_2)\ P(w_2)\ P(w_1,y|\ do(a)) \quad (17)$$

$$= \sum_{w_1}\sum_{w_2} P(y|w_1,w_2)\ P(w_2)\ P(w_1|\ a,w_2) \times \sum_{a'} P(y|\ w_1,a')\ P(a'|\ w_2) \quad (18)$$

$$= \sum_{w_1}\sum_{w_2}\sum_{a'} P(y|w_1,w_2,a')\ P(w_2)\ P(w_1|\ a,w_2)\ P(a'|\ w_2) \quad (19)$$

Note that Eq. 19 is exactly the same as Eq. 15. For the same reason stated earlier, w_3 is ommitted in (17). The term $P(w_1, y|\ do(a))$ in (17) is replaced by $P(w_1|\ a,w_2)\sum_{a'} P(y|\ w_1,a')\ P(a'|\ w_2)$ after applying Rule 2 followed by Rule 3 of *do-calculus* (symbolic derivation of Causal Effects: Eq. 3.43 [18]). Since W_2 blocks all back-door paths between A and Y, we apply the back-door formula (Eq. 6) to adjust on W_2 in (19).

4.2.3 C-component factorization. C-component factorization [26] aims to express the observational distribution $P(v)$ as a product of factors $P_{v\backslash s}(s)$, where each s represents the set of vertices included in a c-component. A c-component is a set of vertices in the graph such that every pair of vertices are connected by a confounding edge. The c-components are very important in measuring the causal effect of A on Y since they help in decomposing the identification problem into smaller sub-problems. In other words, variables in the graph can be partitioned into a disjoint set of c-components in order to calculate $P(y_a)$. For example, the graph $G17$ is partitioned into two c-components: $S_1 = \{A, M_2\}$ and $S_2 = \{M_1, Y\}$ while the c-components of $G18$ are: $S_1 = \{A, W_2, Y, W_3\}$ and $S_2 = \{W_1\}$.

Shpitser and Pearl [26] designed an algorithm called **ID** which aims to decompose the identification problem into smaller sub-problems based on the *c-component factorization* property. This algorithm provides a complete solution for computing all identifiable causal effects.

Note that as long as there is no confounding path connecting A to any of its direct children, $P(y_a)$ is identifiable and can be computed as [29]:

$$P(y_a) = \frac{P(y)}{Q^A} \sum_{a'} Q^A \tag{20}$$

where Q^A is the c-factor of the c-component containing A (S^A) computed as follows:

$$Q^A = \prod_{v \in S^A} P(v|\mathbf{v}^{-1}) \tag{21}$$

where $\mathbf{v}^{-1}$ is the set of values of all previous variables to V, assuming a topological order $V_1 \prec V_2 \prec \ldots \prec V_n$ over all variables. For instance, in $G17$, $W_2 \prec A \prec M_1 \prec M_2 \prec Y$ is a valid topological order. This criterion can be slightly generalized to be: $P(y_a)$ is identifiable if there is no confounding path connecting A to any of its children in $G_{An(Y)}$ which is the subgraph of G composed only of ancestors of the outcome variable Y.

To illustrate the c-component factorization property, consider the causal graph $G17$. Hence, applying Eq. 20 to $G17$ leads to:

$$P(y_a) = \frac{P(y)}{P(a)P(m_2|\, a, m_1)} \sum_{a'} P(a')P(m_2|\, a', m_1) \tag{22}$$

Note that the causal effect $P(y_a)$ is not identifiable in a causal graph composed of a single c-component. Graphs $G20$, $G23$, $G25$, $G26$ and $G27$ in Table 3 illustrate such situation. These are discussed further in Section 4.3.

4.2.4 Front-door adjustment. In case a bi-directed edge between the sensitive attribute A and the outcome Y exists, all the above approaches will fail. However, $P(y_a)$ can still be measured using another criterion called the front-door criterion. The graph $G19$ satisfies this criterion. In fact, the back-door criterion cannot be used because of the unobserved confounder (impossible to control for) however, due to the presence of the mediators M, the front-door criterion can be applied to identify the causal effect as follows:

$$P(y_a) = \sum_{m} P(m|a) \sum_{a'} P(y|m, a')\, P(a') \tag{23}$$

More generally, the front-door adjustment can be applied if the the following conditions hold:

(1) all of the direct paths from A to Y pass through M.
(2) there are no back-door paths from A to M,
(3) all back-door paths from M to Y are blocked by A.

Back-door and front-door adjustments are the main ingredients of the *do-calculus*(Section 4.2.1).

4.2.5 Instrumental variables. Consider graphs $G21$ and $G22$ in Table 3. For non-parametric causal models, the causal effects $P(y_a)$ are not identifiable. However, for linear models, the causal effects for both graphs are identifiable using instrumental variables. A variable I is said to be instrumental when it affects A, and only affects Y by influencing A. That is, I is an instrumental variable (relative to the pair (A, Y) if (1) I is independent of all variables (including unobserved variables) that have an influence on Y that is not mediated by A and (2) I is not independent of A [18]. For graph $G21$, assuming a linear model, identifying the causal effect of A on Y is equivalent to identifying the coefficient $b = \frac{r_{IY}}{r_{IA}}$ on the edge: $A \rightarrow Y$. where: r_{IY} is the slope of the regression line of I on Y and r_{IA} is the slope of the regression line of I on A. Hence, instrumental variables provide an efficient way to calculate the causal effect of A on Y without the need of controlling on the unobserved confounders. I might be used to infer the causal effect $P(y_a)$ when some additional observed variable Z can be used to control on such that: (1) $(I \not\perp\!\!\!\perp A|\, Z)_G$, (2) $(I \perp\!\!\!\perp Y|\, Z)_{G_{\overline{A}}}$. For example, in graph $G22$, I is an instrument variable and by making back-door adjustments for M, we can identify $P(y_i)$ and $P(a_i)$. Since all the causal influence of I on Y must be channeled through A, we have:

$$P(y_i) = \sum_{a} P(y_a)\, P(a_i) \tag{24}$$

Thus, the causal effect of A on Y is identifiable whenever Eq. 24 can be solved for $P(y_a)$ in terms of $P(y_i)$ and $P(a_i)$.

4.3 Non-identifiability: The hedge criterion

In some causal graphs neither the back-door, nor the front-door criteria are satisfied. The simplest graph where $P(y_a)$ cannot be calculated is the bow graph (graph $G20$ in Table 3). The back-door criterion fails since the confounder variable is unobserved, while the front-door criterion fails since no intermediate variables between A and Y exist in the graph.

Shpitser and Pearl [26] have constructed a list of more generalized graphs inheriting the difficulty of the bow structure. Such graphs are called C-trees. A C-tree is a graph that is at the same time a tree[10] and a c-component (defined earlier in Section 4.1). A tree is a graph such that each vertex (variable) has at most one child, and only one vertex (called the root) has no children. Graphs $G20-25$ are examples of Y-rooted C-trees (C-trees having as root the variable Y). Shpitser and Pearl proved that the causal effect of a Y-rooted C-tree is always unidentifiable (Theorem 12 in [26]). More generally, they demonstrated that all the unidentifiable cases of the causal effect $P(y_a)$ boil down to a general graphical structure called: *hedge* which is defined as follows [26]:
F and F' (sub-graphs in G) form a hedge for $P(y_a)$ if:

- F and F' are $\mathbf{R}$-rooted C-forests in G ($\mathbf{R} \in An(\mathbf{Y})_{G_{\overline{A}}}$)
- F' is a sub-graph of F
- A only occurs in F

where a C-forest is a graph G which is both a c-component and a forest. And a forest is a graph G such that each vertex has at most one child. Note that any Y-rooted C-tree and its root node Y form a hedge.

Table 3 presents other patterns where $P(y_a)$ is not identifiable due to the existence of a hedge structure in the graph. The graphs ($G26-30$) possess multiple c-components by contrast to $G20-25$. So, under such situations, the identifiability of the causal graph turns into the identifiability of each one of the c-components constituting this graph. In other words, each c-component is examined for a potential presence (or absence) of a hedge structure. It is sufficient to

[10]Notice that the direction of the arrows between nodes is reversed compared to the usual tree structure.

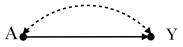
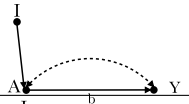
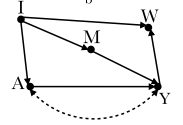
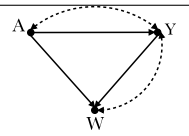
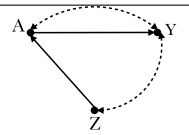
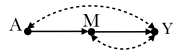
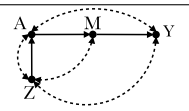
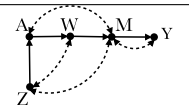
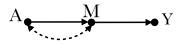
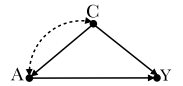
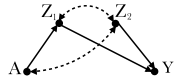
	Causal graph	Hedge
$G20$		$F : \{A, Y\}$ $F' : \{Y\}$
$G21^*$		$F : \{A, Y\}$ $F' : \{Y\}$
$G22^*$		$F : \{A, Y\}$ $F' : \{Y\}$
$G23$		$F : \{A, W, Y\}$ $F' : \{W, Y\}$
$G24$		$F : \{A, Z, Y\}$ $F' : \{Z, Y\}$
$G25$		$F : \{A, M, Y\}$ $F' : \{M, Y\}$
$G26$		$F : \{A, M, Z, Y\}$ $F' : \{M, Z, Y\}$
$G27$		$F : \{A, W, M, Z, Y\}$ $F' : \{W, M, Z, Y\}$
$G28$		$F : \{A, M\}$ $F' : \{M\}$
$G29$		$F : \{A, C\}$ $F' : \{C\}$
$G30$		$F : \{A, Z_1, Z_2\}$ $F' : \{Z_1, Z_2\}$

Table 3: Non-identifiable semi-Markovian models. * indicates that the graph is identifiable in linear models.

discover a hedge structure in one of the c-components to conclude that the whole causal graph is not identifiable (Theorem 19 [26]). For example, the graph $G30$ is composed of two maximal c-components: $S1 = \{A, Z_1, Z_2\}$ and $S2 = \{Y\}$. Starting by the former, it is easy to recover a hedge structure: $F = \{A, Z_1, Z_2\}$ and $F' = \{Z_1, Z_2\}$ leading to the unidentifiability of the whole causal graph. Table 3 shows the sets F and F' for each graph.

Apart from measuring the causal effect for an identifiable causal graph, the **ID** algorithm [26] tells why $P(y_a)$ is not identifiable. That is, in case a particular graph is not identifiable, the algorithm returns the two sub-graphs F and F' that form a hedge in that graph.

5 IDENTIFICATION OF COUNTERFACTUAL EFFECTS

While causal effects (Section 4) interpret the effect of actions as downward flow, counterfactual effects require more complex reasoning. Basically, counterfactual effects measure fairness based on multiple worlds: the actual world and other hypothetical (or counterfactual) worlds. The actual world is represented by a causal model M in its actual (normal) state without any interventions, while the counterfactual worlds are represented by sub-models: M_a where the intervention $do(a)$ forces the actual state to change to an alternative state.

Note that in Markovian, as well as semi-Markovian models, if all parameters of the causal model are known (including $P(\mathbf{u})$), any counterfactual is identifiable and can be computed using the three steps abduction, action, and prediction (Theorem 7.1.7 in [18]). However, this method is usually infeasible in real-world scenarios due to the lack of the complete knowledge of the causal model (more specifically the knowledge of the background variables U).

By contrast to causal effects, the calculation of counterfactual effects cannot only rely on the observational data summarized by $P(v)$ and the consensual causal graph G. In fact, counterfactual effects need, in addition, as input a set of experiments denoted as P_* (called also: interventional probabilities). These experiments are formally defined as: $P_* = \{P_{\mathbf{x}}|\mathbf{X} \subseteq \mathbf{V},\ \mathbf{x}\ x\ value\ assignment\ of\ \mathbf{X}\}$ and are possible to perform in principle on a given causal model. Thus, the identifiability of counterfactuals depends on the identifiability of P_* which in turn depends on $P(v)$ and G. Shpitser and Pearl [26] designed **ID*** and **IDC*** algorithms to evaluate counterfactuals given a causal graph and an observational data. In case these algorithms fail to uniquely measure a counterfactual quantity, the situation is referred to as an unidentifiable situation.

5.1 Counterfactual graph

Given a causal graph G of a Markovian model and a counterfactual expression $\gamma = v_a|e$ with e some arbitrary set of evidence, measuring $P(\gamma)$ requires to construct a counterfactual graph which combines parallel worlds. Every world is represented by a counterfactual sub-model M_a. For example, Figure 3 shows a causal graph for the firing example (Figure 3(a)) along with its corresponding counterfactual graph (Figure 3(b)). Thus, Figure 3(b) combines two worlds: the actual world where the teacher has actually $A = a_0$ and the counterfactual world where the same teacher is assigned A^*[11] $= a_1$. As shown in the figure, the two worlds share the same unobserved background variable: U_Y that highlights the interaction between these worlds. Note that no bi-directed edges are connected to the node $A^* = a_1$. The reason for that is that the intervention $do(a^* = a_1)$ removes all the incoming arrows to A^*. Thus, in order to calculate the counterfactual expression $P(Y^*_{a*=a_1} | A = a_0)$ of the

[11]The subscript * is added to nodes belonging to the counterfactual world for graph legibility.

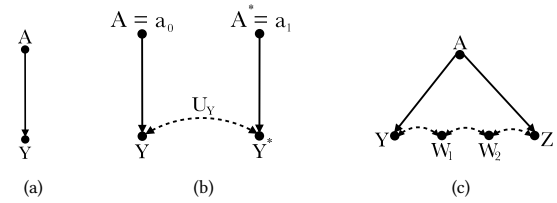


Figure 3: (a) A causal graph for the firing example (b) A corresponding counterfactual graph for the query $P(Y^*_{a*=a_1} \mid A = a_0)$(c) Zig-zag pattern.

simple Markovian graph in Figure 3(a), we need to construct the semi-Markovian graph in Figure 3(b). The **make-cg** algorithm [26] automates this procedure. Basically, **make-cg** algorithm starts by combining the two causal graphs (actual and counterfactual) and makes them share the same background variable **U** (as shown in Figure 3(b)). Then, it discards the duplicated endogenous nodes which are not affected by $do(a)$.

One typical unidentifiable counterfactual quantity is $P(y_{a'}, y'_a)$ which is called the probability of necessity and sufficiency. The corresponding counterfactual graph is the W-graph that has the same structure as to Figure 3(b). This simple criterion can be generalized to the zig-zag graph (Figure 3(c)) where the counterfactual $P(y_a, w_1, w_2, z'_x)$ is not identifiable.

5.2 C-component factorization

Now, in order to measure the counterfactual effects of a certain constructed counterfactual graph, all the graphical criteria listed in Section 4 hold. For instance, the c-component factorization approach can be used to decompose the counterfactual graph into a set of disjoint sub-graphs (or c-components). Thus, the joint distribution of all variables in the counterfactual graph can be factorized as the product of the conditional distribution of each c-component. Thus, if a certain causal effect is not identifiable in a particular c-component, the counterfactual quantity of the whole model is not identifiable as well.

To illustrate how counterfactual quantities are measured, consider another simple scenario for the firing example. As shown in Figure 4(a), we assume now that the confounder variable (location of school) is observable. Consider the counterfactual query: $P(y_{a_1}|a_0)$ which reads the probability of firing a teacher who is assigned a class with a high initial level of students (a_0) had she been assigned a class with a low initial level of students (a_1). Figure 4(b) shows the two parallel-worlds graph[12] for the query while Figure 4(c) presents the final constructed counterfactual graph using **make-cg** algorithm. Note that in Figure 4(c), C and C^* are merged as a single node C (by applying Lemma 24 [26]). The main reason for that is that these nodes are not descendants of A. Then, C inherits all the children of both nodes C (the old node in the previous graph) and C^*. Finally, U_C is omitted since any unobserved variable that possesses a single child should be removed [26].

Now, having constructed the counterfactual graph for the counterfactual expression $P(y_{a_1}|a_0)$, we can turn to the identifiability of this expression. Note that the obtained counterfactual graph (Figure 4(c)) has three c-components: $\{C\}$, $\{A\}$, $\{Y, Y^*_{a_1}\}$ thus, applying algorithm **IDC*** [26] results in:

$$P(y_{a_1}|a_0) = \frac{\sum_{y,c}\ Q(c)\ Q(a_0)\ Q(y, y_{a_1})}{P(a_0)} \tag{25}$$

where $Q(\mathbf{v}) = P(\mathbf{v}|pa(\mathbf{V}))$ in the counterfactual graph.

Hence,

$$P(y_{a_1}|a_0) = \frac{\sum_{y,c}\ P(c)\ P(a_0|c)\ P(y, y_{a_1}|c)}{P(a_0)}$$

$$= \frac{\sum_{c}\ P(c)\ P(a_0|c)\ P(y_{a_1}|c)}{P(a_0)} \tag{26}$$

$$= \frac{\sum_{c}\ P(c)\ P(a_0|c)\ P(y|a_1, c)}{P(a_0)} \tag{27}$$

y in Eq. 26 is cancelled by summation while $P(y_{a_1}|c)$ in the same equation is transformed into $P(y|a_1, c)$ in Eq. 27 using rule (2) of the do-calculus (Section 4.2).

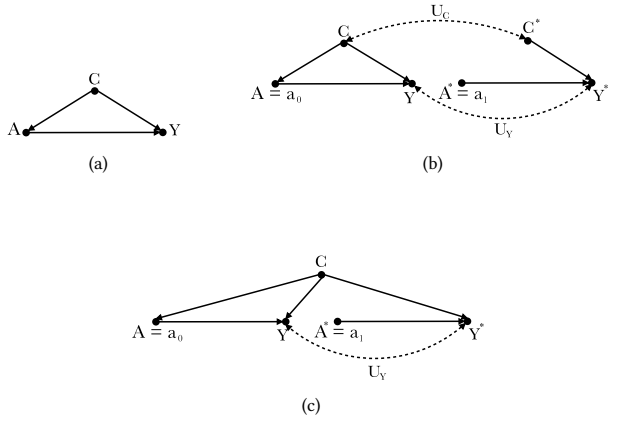


Figure 4: (a) Original causal graph for the firing example (b) Parallel worlds graph for $P(y_{a_1}|a_0)$ (c) Counterfactual graph for $P(y_{a_1}|a_0)$.

Table 4 presents various examples of identifying the counterfactual quantities (column 3) of some causal graphs (first column) after obtaining their corresponding counterfactual graphs (column 2).

For example, $G33$ includes in addition to the confounder C a mediator M. As shown in the corresponding counterfactual graph, the nodes M and $M^*_{a_1}$ are not merged as they differ on their A-derived parents by contrast to the node C. Similarly, in $G34$, the pair of nodes M, $M^*_{a_1}$ and W, $W^*_{a_1}$ are not merged for the same reason.

6 IDENTIFICATION OF PATH-SPECIFIC EFFECTS

Identifying path-specific effects in a fairness context arises when one is interested in measuring the causal effect of the sensitive attribute A on the outcome Y only along certain pathways in the causal graph. That is, only the paths of interest are kept while all the other paths in the graph are excluded and not considered in the analysis.

[12]This graph is called *twin network* graph since it includes only two hypothetical graphs [26].

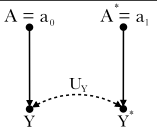
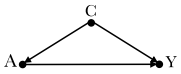
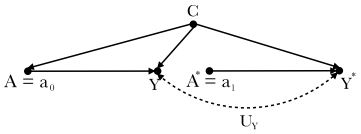
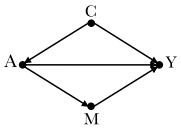
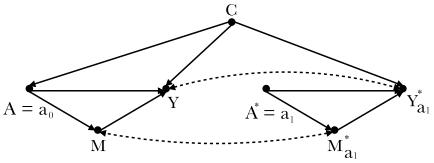
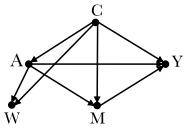
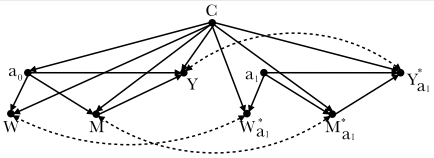
	Original Graph	Counterfactual Graph	$P(y_{a_1} \mid a_0)$
G31			unidentifiable if $y \neq y^*$
G32			$\frac{\sum_c P(c)\, P(a_0 \mid c)\, P(y \mid a_1, c)}{P(a_0)}$
G33			$\frac{\sum_{c,m} P(c)\, P(a_0 \mid c)\, P(y \mid m, m', a_1, c)\, P(m' \mid a_0, c, a_1)}{P(a_0)}$
G34			$\frac{\sum_c P(c)\, P(a_0 \mid c)\, P(w \mid a_0, c)\, P(m' \mid a_1, c)\, P(y' \mid a_1, c, m')}{P(a_0)}$

Table 4: Identifiability of counterfactual effects.

Direct and indirect effects are the simplest cases of path-specific effects. While the direct effect isolates the effect of A on Y along the direct path $A \rightarrow Y$, indirect effect considers the indirect causal paths between A and Y ($A \rightarrow \cdots \rightarrow Y$).

A more general and complex case is when one wants to isolate the effect of A on Y along a specific group of paths. Such case is called path-specific effect.

6.1 Identification of direct and indirect effects

Average natural direct effect *NDE* and Average natural indirect effect *NIE* have been introduced by Pearl in [17]. *NDE* measures the total effect of A on Y that is not mediated by other variables M in the causal model. In other words, *NDE* evaluates the sensitivity of Y to variations in A while fixing all the other variables of the model. For example, the average natural direct effect in Figure 5(a) is written as:

$$NDE_{a_1,a_0}(Y) = \mathbf{E}[y_{a_1}, M_{a_0}] - \mathbf{E}[y_{a_0}] \tag{28}$$

where $\mathbf{E}[.]$ is the expectation of a random variable over all data inputs.

Considering the simple job hiring example (with $A = a_1$ = female, $A = a_0$ = male, y = hiring, M= education level). Eq. 28 measures the expected change in male hiring ($\mathbf{E}[y_{a_0}]$) had A been a_1 (female), while mediators $\mathbf{M}$ are kept at the level they would take had A been a_0 (e.g. male), in particular for the individuals $A = a_1$ (e.g. female).

NIE, on the other hand, measures the effect of the mediator M at levels M_{a_0} and M_{a_1} on Y had A been a_0 (Figure 5(b)) and is defined as:

$$NIE_{a_1,a_0}(Y) = \mathbf{E}[y_{a_0}, M_{a_1}] - \mathbf{E}[y_{a_0}] \tag{29}$$

Considering the same hiring example, Eq. 29 measures the expected change in male hiring $\mathbf{E}[y_{a_0}]$, if males had equal education levels ($M = m$) as those of females ($A = a_1$). Note that in the context of discrimination discovery, *NIE* has been used under the assumption that A has no parent node in the causal diagram (no spurious discrimination) [36].

In Markovian models, *NDE* and *NIE* are identifiable from observational data and can be calculated as follows [17]:

$$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \sum_{\mathbf{c}} \Big(\mathbf{E}[Y|a_1, \mathbf{m}] - \mathbf{E}[Y|a_0, \mathbf{m}]\Big) P(\mathbf{m}|a_0, \mathbf{c}) P(\mathbf{c}) \tag{30}$$

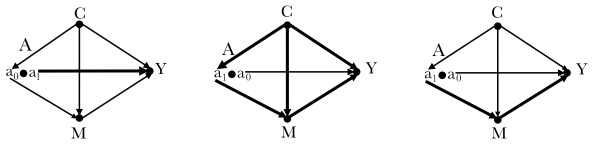


Figure 5: Simple Markovian model that measures (a) *NDE* (b) *NIE* and (c) *PSE* through the heavy edges.

$$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \sum_{\mathbf{c}} \mathbf{E}[Y|a_0, \mathbf{m}] \left(P(m|a_1, c) - P(m|a_0, c)\right) P(c) \tag{31}$$

Where $\mathbf{M}$ is a set of mediator variables and $\mathbf{C}$ stands for any set satisfying the back-door criterion between the sensitive attribute A and the outcome Y. For instance, the graph $G37$ in Table 5 illustrates a simple Markovian model for which the *NDE* is given by Eq. 30 and the *NIE* is given by Eq. 31. In case the effect of A on M is not confounded (graphs $G35$ and $G36$), *NDE* is calculated using the following simplified equality:

$$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \Big(\mathbf{E}[Y|a_1, \mathbf{m}] - \mathbf{E}[Y|a_0, \mathbf{m}]\Big) P(\mathbf{m}|a_0) \tag{32}$$

while *NIE* is measured as follows:

$$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \mathbf{E}[Y|a_0, \mathbf{m}] \Big(P(m|\ a_1) - P(m|\ a_0)\Big) \tag{33}$$

In semi-Markovian models, *NDE* and *NIE* are not generally identifiable. However, if there exists a set Z of covariates, non-descendants of A or M, such that, for all values of a and m we have the following conditions (Theorem 2 [17]):

$$(1) \quad Y_{a_1 m} \perp\!\!\!\perp M_{a_0} |\ Z$$

$$(2) \quad P(Y_{a_1 m} = y|\ Z = z) \quad \text{is identifiable}$$

$$(3) \quad P(M_{a_0} = m|\ Z = z) \quad \text{is identifiable}$$

then, *NDE* is identifiable and is calculated as follows:

$$NDE_{a_1,a_0}(Y) = \sum_{m,z} \Big(\mathbf{E}[Y_{a_1,m}|\ z] - \mathbf{E}[Y_{a_0,m}|\ z]\Big)\ P(M_{a_0} = m|\ z)\ P(z) \tag{34}$$

Similarly, *NIE* is identifiable in semi-Markovian models using observational data and is given by (Theorem 4 [17]):

$$NIE_{a_1,a_0}(Y) = \sum_{m,z} \mathbf{E}[Y_{a_0,m}|\ z] \Big(P(M_{a_1} = m|\ z) - P(M_{a_0} = m|\ z)\Big) P(z) \tag{35}$$

if the following expressions are satisfied for all values of a and m:

$$(1) \quad Y_{a_0,m} \perp\!\!\!\perp M_{a_1} |\ Z$$

$$(2) \quad \mathbf{E}[Y_{a_0,m}|\ z] \quad \text{is identifiable}$$

$$(3) \quad P(M_{a_1} = m|\ z) \quad \text{is identifiable}$$

$$(3) \quad P(M_{a_0} = m|\ z) \quad \text{is identifiable}$$

6.2 Identification of path-specific effects

One of the challenges of discrimination discovery is to distinguish between two types of indirect effects of the sensitive attribute A on the outcome Y, namely: the indirect discrimination (unfair effect) and the explainable effect (fair effect). While the simple *NIE* (explained in the previous section) considers all the indirect paths between A and Y regardless of whether they are fair or not, path-specific effect provides a more fine-grained way to consider effects along a selected subset of paths between A and Y. In other words, path-specific effect makes the distinction between the two types of indirect effects possible. Figure 5(c) shows an example of a causal graph where only the heavy path ($A \rightarrow M \rightarrow Y$) is selected for effect analysis. Given a path set π, the π-specific effect [17] is defined as:

$$PSE^{\pi}_{a_1,a_0}(y) = P(y_{a_1|_{\pi},a_0|_{\overline{\pi}}}) - P(y_{a_0}) \tag{36}$$

where $P(y_{a_1|_{\pi},a_0|_{\bar{\pi}}})$ is the probability of $Y = y$ in the counterfactual situation where the effect of A on Y with the intervention ($A = a_1$) is transmitted along π, while the effect of A on Y without the intervention ($A = a_0$) is transmitted along paths not in π (denoted by: $\bar{\pi}$).

The identifiability of $PSE^{\pi}_{a_1,a_0}(y)$ depends on the identifiability of the term $P(y_{a_1|_{\pi},a_0|_{\pi}})$. Avin et al. [1] provided the necessary and sufficient condition for $P(y_{a_1|_{\pi},a_0|_{\bar{\pi}}})$ to be identifiable in Markovian models, namely, the recanting witness criterion.

Given a path π in G pointing from A to Y, the recanting witness criterion is satisfied for π-specific effect if and only if there exists a variable R (known as witness) in G such that: (1) there exists a path from A to R in π, (2) there exists a path from R to Y in π, and (3) there exists another path from R to Y not in π. The graphical pattern of this criterion is called the "kite" pattern and is shown in Figure 6(a). In this graph, the witness variable is R thus, $\pi = A \rightarrow R \rightarrow Z \rightarrow Y$ while $\overline{\pi} = A \rightarrow R \rightarrow Y$.

Hence, if the recanting witness criterion is satisfied, $PSE^{\pi}_{a_1,a_0}(y)$ is not identifiable. Figure 6(a) illustrates a simple example where $PSE^{\pi}_{a_1,a_0}(y)$ is identifiable as the recanting witness criterion is not satisfied. Under such setting, $P(y_{a_1|_{\pi},a_0|_{\bar{\pi}}})$ can be measured by applying the following steps:

(1) Express $P(y_{a_0})$ using the truncated factorization formula according to Eq. 7.
(2) Segregate $Ch(A)$ other than Y ($Ch(A)\backslash Y$) into two sets: $\mathbf{S_1}$ and $\mathbf{S_2}$ (where: $\mathbf{S_1} \cap \mathbf{S_2} = \emptyset$). The nodes $\in \mathbf{S_1}$ belong to edges in π while nodes $\in \mathbf{S_2}$ belong to edges not in π or not included in any path from A to Y.
(3) Replace values a_0 with a_1 for the terms corresponding to nodes in $\mathbf{S_1}$, and keep values a_0 unchanged for the terms corresponding to nodes in $\mathbf{S_2}$ (Theorem 2 [23]).

For example, in Figure 6(a), $S_1 = \{R\}$ and $S_2 = \emptyset$. Thus, $P(y_{a_1|_{\pi},a_0|_{\bar{\pi}}})$ is identifiable and is given by:

$$\sum_{r,w,z} P(r|\ a_1) P(z|\ r)\ P(w|\ r)\ P(y|\ z)$$

Table 6 presents other examples of Markovian models where π-specific effect is identifiable. The formula of how the quantity: $P(y_{a_1|_{\pi},a_0|_{\bar{\pi}}})$ is calculated for each example and is shown in the third column.

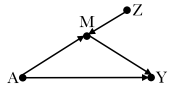
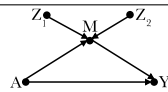
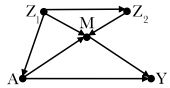
	Causal graph	Identifiability Formula
G35		$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \Big(\mathbf{E}[Y\|a_1,\mathbf{m}] - \mathbf{E}[Y\|a_0,\mathbf{m}]\Big)P(\mathbf{m}\|a_0)$
G36		$NIE_{a_1,a_0}(Y) = \sum_{\mathbf{m}} \mathbf{E}[Y\|a_0,\mathbf{m}]\Big(P(m\|\ a_1) - P(m\|\ a_0)\Big)$
G37		$NDE_{a_1,a_0}(Y) = \sum_{\mathbf{z_2}}\sum_{\mathbf{m}} \Big(\mathbf{E}[Y\|a_1,\mathbf{m}] - \mathbf{E}[Y\|a_0,\mathbf{m}]\Big)P(\mathbf{m}\|a_0,\mathbf{z_2})P(\mathbf{z_2})$ $NIE_{a_1,a_0}(Y) = \sum_{\mathbf{z_2}}\sum_{\mathbf{m}} \mathbf{E}[Y\|a_0,\mathbf{m}]\Big(P(m\|a_1,z_2) - P(m\|a_0,z_2)\Big)\ P(z_2)$

Table 5: Identifiability of NDE and NIE.

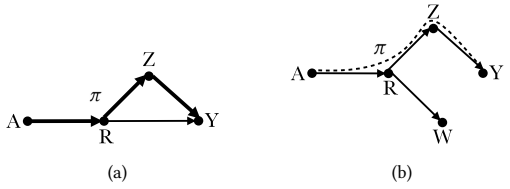


Figure 6: (a) The recanting witness criterion satisfied (the "kite" pattern). (b) illustrates an example of identifiable π-specific effect graphs. π path is represented by a heavy line.

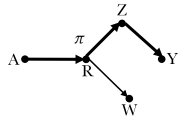
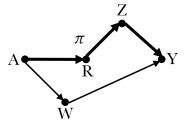
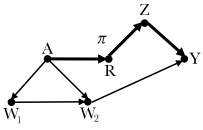
	Causal graph	$P(y_{a_1\|_\pi, a_0\|_{\overline{\pi}}})$
G38		$\sum_{r,w,z} P(r\|\ a_1)P(z\|\ r)\ P(w\|\ r)$ $P(y\|\ z)$
G39		$\sum_{r,w,z} P(w\|\ a_0)\ P(r\|\ a_1)\ P(z\|\ r)$ $P(y\|\ z,w)$
G40		$\sum_{w_1,w_2,r,z} P(w_1\|\ a_0)\ P(w_2\|\ a_0)$ $P(r\|\ a_1)P(z\|\ r)\ P(y\|\ z,w_2)$

Table 6: Examples of Markovian graphs where PSE is identifiable due to the absence of the recanting witness.

Shpitser [23] extended the recanting witness criterion to deal with semi-Markovian models. Thus, the identifiablity of $PSE^{\pi}_{a_1,a_0}(y)$ depends on the existence (or absence) of a pattern in the graph called the recanting district. In other words, if the recanting district exists, $PSE^{\pi}_{a_1,a_0}(y)$ is not identifiable and cannot be measured from observational data.

Given a graph G and two sets of nodes: **A** and **Y** in G. Let π be a path in G starting with a node in **A** and ending with a node in **Y**. Let **V** be the set of nodes not in **A** which are ancestral of **Y** via a directed path which does not intersect **A**. Then a district D[13] in G is called a recanting district for the π-specific effect of A on Y if there exist nodes $v_i, v_j \in D$, $a_i \in \mathbf{A}$, and $y_i, y_j \in \mathbf{Y}$ such that there is a path: $a_i \rightarrow v_i \cdots y_i$ in π and an another path $a_i \rightarrow v_j \cdots y_j$ not in π.

As an example, consider the two semi-Markovian graphs of Figure 7. In either of these graphs, *the recanting district* exists. In Figure 7(a), the district $\{M_1, M_2, M_3, Y\}$ is recanting since the path $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ is the π path of interest while there exists another path not in π connecting A and Y, namely, $A \rightarrow M_3 \rightarrow Y$, and M_1 and M_3 belong to the same district: $\{M_1, M_2, M_3, Y\}$. Similarly, in Figure 7(b), the district $\{M_3\}$ is recanting since there exist two paths connecting A and Y: the first is π itself ($A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$) and the second is not in π ($A \rightarrow M_3 \rightarrow Y$), and M_3 is its own district.

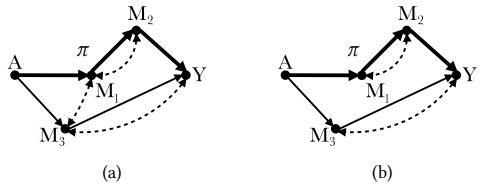


Figure 7: The recanting district criterion satisfied. The heavy line illustrates the π path of interest. The recanting district for Figure 7(a) is $\{M_1, M_2, M_3, Y\}$ while the recanting district of Figure 7(b) is $\{M_3\}$

Now, if the recanting district criterion is not satisfied, the π-specific effect is identifiable and $PSE^{\pi}_{a_1,a_0}(y)$ can be calculated from observational data as follows (Theorem3 [23]):

$$\sum_{\mathbf{V}\setminus\{Y\}} \prod_{D} P(D = d\|\ do(E_D = e_D)) \tag{37}$$

[13] A district is merely a c-component (introduced in Section 4.2).

where V is the set of nodes not in A which are ancestral of Y via a directed path which does not intersect A. D ranges over all districts in the graph G_V. E_D refers to nodes with directed arrows pointing into D but which are themselves not in D, and value assignments d and e_D are assigned as follows: if any element a in A occurs in E_D in a term $P(D = d|\ do(E_D = e_D))$, then it is assigned a baseline value if the arrows from a to elements in D are not in π, and an active value if the arrows from a to elements in D are in π[14]. Let the baseline and the active values are a_0 and a_1, respectively. All other elements in E_D and D are assigned values consistent with the values indexed in the summation.

Table 7 shows examples of semi-Markovian models where the π-specific effect is identifiable. Note that for all these graphs, it is never the case that both an arrow in π and an arrow not in π emanating from the node A to nodes in the same district such that these nodes are ancestors of Y. This implies there is no recanting district for the effect of A on Y hence, the π-specific effect is indeed identifiable. For instance, $G41$ involves a single district: $\{M_1, M_2, M_3, Y\}$. Applying Eq. 37 yields to:

$$\sum_{m_1,m_2,m_3} P(m_1, m_2, m_3, y|\ do(a_1))$$

Laveraging the general theory of identification of interventional probabilities, the above expression can be transformes as follows:

$$\sum_{m_1,m_2,m_3} P(m_1|\ a_1, m_3)\ P(m_2|\ m_1)\ P(m_3|\ m_1, y)\ P(y|\ m_3, m_2)$$

$G43$, at the other hand, presents a more sophisticated case including three districts, namely, $\{M_1\}$, $\{M_2\}$, and $\{M_3, M_4, Y\}$. Thus,

$$\sum_{m_1,m_2,m_3,m_4} P(m_1|\ do(a_1, m_3))\ P(m_3, m_4, Y|\ do(a_0, m_1, m_2))$$
$$P(m_2|\ do(m_1, m_4))$$

7 CONCLUSION

A typical goal of causal inference in the context of discrimination discovery is establishing the causal effect of the sensitive attribute A on the outcome Y. Unfortunately, this may not be possible due to the identifiability problem. This paper studied the problem of identifiability as it relates to discrimination discovery. We made use of the large-scale body of work on identifiability theory to summarize the main results found in the literature. Based on various graphical patterns, we discussed and assessed whether the causal effect of A on Y is identifiable. The main identifiability results fall into three main types, namely the causal effect (intervention), the counterfactual effect and the path-specific effect. Finally, we note that When identification is not possible, it may still be possible to bound causal effects. The development of bounds for non-identifiable quantities is called partial identifiability.

8 ACKNOWLEDGEMENTS

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.

[14]We use the term "active" to designate the value assigned to the sensitive attribute A along the causal paths we are interested in and the term "baseline" to designate the value assigned to A along all the other causal paths

REFERENCES

[1] Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of path-specific effects. In *Proceedings of the 19th international joint conference on Artificial intelligence*. 357–363.

[2] Joseph Berkson. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* 2, 3 (1946), 47–53.

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

[4] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[6] Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*. Springer, 66–70.

[7] David Galles and Judea Pearl. 1995. Testing identifiability of causal effects. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. 185–195.

[8] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[9] Yimin Huang and Marco Valtorta. 2006. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the national conference on artificial intelligence*, Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 1149.

[10] Yimin Huang and Marco Valtorta. 2006. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 217–224.

[11] H Kim and J Perl. 1983. A computational model for combined causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI)*. Morgan-Kaufmann, San Mateo, CA.

[12] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017), 4066–4076.

[13] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[14] Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. 2019. A Potential Outcomes Calculus for Identifying Conditional Path-Specific Effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 3080–3088.

[15] Catherine O'Neill. 2016. Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy* (2016).

[16] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

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

[18] Judea Pearl. 2009. *Causality*. Cambridge university press.

[19] Kimberly Quick. 2015. The Unfair Effects of IMPACT on Teachers with the Toughest Jobs. *The Century Foundation* (2015).

[20] Michelle Rhee. 2019. IMPACT: The DCPS Evaluation and Feedback System for School-Based Personnel.

[21] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.

[22] Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* 37, 6 (2013), 1011–1035.

[23] Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* 37, 6 (2013), 1011–1035.

[24] Ilya Shpitser and Judea Pearl. 2006. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 437–444.

[25] Ilya Shpitser and Judea Pearl. 2007. What counterfactuals can be tested. In *23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*. 352–359.

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

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

[28] Jin Tian. 2004. Identifying linear causal effects. In *Proceedings of the 2004 AAAI Conference on Artificial Intelligence*. 104–111.

[29] Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *Proceedings of the 2002 AAAI Conference on Artificial Intelligence*. 567–573.

[30] Jin Tian and Ilya Shpitser. 2003. On the identification of causal effects. (2003).

[31] Santtu Tikka. 2018. Improving identification algorithms in causal inference. *Report/University of Jyväskylä, Department of Mathematics and Statistics* 168 (2018).

[32] Santtu Tikka and Juha Karvanen. 2017. Identifying Causal Effects with the R Package causaleffect. *Journal of Statistical Software* 76 (2017).

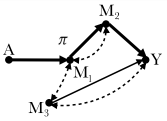
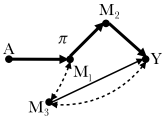
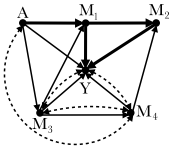
	Causal graph	$P(y_{a_1\|_\pi, a_0\|_{\overline{\pi}}})$
G41		$\sum_{m_1,m_2,m_3} P(m_1\mid a_1, m_3)\ P(m_2\mid m_1)\ P(m_3\mid m_1, y)\ P(y\mid m_3, m_2)$
G42		$\sum_{m_1,m_2,m_3} P(m_1\mid a_1, m_3)\ P(m_3\mid m_1, y)\ P(y\mid m_3, m_2)\ P(m_2\mid m_1)$
G43		$\sum_{m_1,m_2,m_3,m_4} P(m_1\mid a_1, m_3)\ \ P(m_2\mid m_1, m_4)\ P(m_3\mid a_0)\ P(m_4\mid a_0, m_3)\ P(y\mid a_0, m_1, m_2)$

Table 7: Examples of semi-Markovian graphs where PSE is identifiable due to the absence of the recanting district.

[33] Santtu Tikka and Juha Karvanen. 2017. Simplifying probabilistic expressions in causal inference. *The Journal of Machine Learning Research* 18, 1 (2017), 1203–1232.

[34] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual Fairness: Unidentification, Bound and Algorithm.. In *IJCAI*. 1438–1444.

[35] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*. 3404–3414.

[36] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the... AAAI Conference on Artificial Intelligence*.

[37] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (2017), 1–16.