Identifiability of Causal-based ML Fairness Notions

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi

Identifiability of Causal-based ML Fairness Notions

Karima Makhlouf
INRIA, Ecole Polytechnique
Palaiseau, France
makhlouf@lix.polytechnique.fr

Sami Zhioua
INRIA, Ecole Polytechnique
Palaiseau, France
zhioua@lix.polytechnique.fr

Catuscia Palamidessi
INRIA, Ecole Polytechnique
Palaiseau, 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 safe 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. The first contribution of this paper is a compilation of the major identifiability results which are of particular relevance for machine learning fairness. To the best of our knowledge, no previous work in the field of ML fairness or causal inference provides such systemization of knowledge. The second contribution is more general and addresses the main problem of using causality in machine learning, that is, how to extract causal knowledge from observational data in real scenarios. This paper shows how this can be achieved using identifiability.**

***Index Terms*—Causality, Machine Learning, Fairness**

I. 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 [4] 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 [7] 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 [22] and Berkson's paradox [1], [9]. A famous example of the Simpson's paradox is the gender bias in 1973 Berkley admission [2], [11]. 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 [11]. Examples of causal-based fairness notions include total effect [14], interventional fairness [17], counterfactual fairness [10], counterfactual effects [29], and path-specific counterfactual fairness [3], [28]. 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 [13], [14].

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 two categories: causal effect (intervention) identifiability [6], [8], [14], [19], [21], [23]–[25] and counterfactual identifiability [18], [20], [21], [27]. Section II provides necessary background concepts. Then, instead of repeating the definition of identifiability (Definition 3.2.3 in [14]), Section III gives an intuitive explanation of the identifiability problem through the teacher firing example. Sections IV and V compile the common identifiability results of causal and counterfactual effects, respectively. Section VI concludes.

II. 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

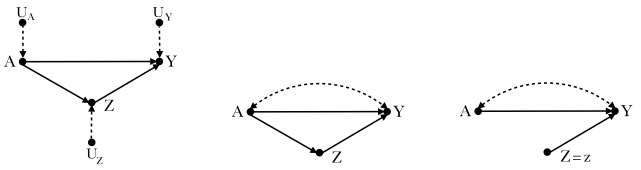
This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294.

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 [14] 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 I) and $G16$ (Table II) show causal graphs of Markovian and semi-Markovian models, respectively.



(a) Markovian model. (b) Semi-Markovian model. (c) Intervention: $do(Z = z)$.

Fig. 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. 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[1]. $P(Y = y|do(Z = z)) = P(Y_{Z=z} = y) = P(Y_z = y) = P(y_z)$ 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.

III. 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 [12], [15], [16]. Assume that the system takes as input one feature, namely, the initial[2] 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

[1]The notations $Y_{Z \leftarrow z}$ and $Y(z)$ are used in the literature as well.

[2]At the beginning of the academic year.

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 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[3]. Notice that the edges $U \rightarrow A$ and $U \rightarrow Y$ are dotted because they are emanating from an unobservable variable (U).

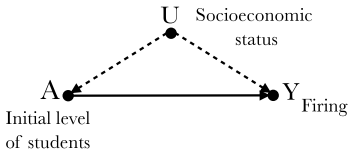


Fig. 2: Causal graph of the teacher firing example.

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 [4] 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) [14] 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 [14], 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[4]:

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 $\hat{Y}$.

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

[4]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.

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. [7] illustrate this case with an example. Tikka [26] 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.

IV. 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) [5]. 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.

Table I 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:

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

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[5]. 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[6] on that confounder using back-door

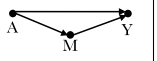
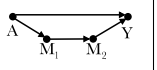
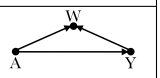
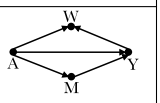
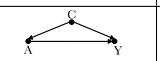
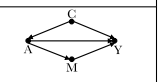
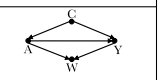
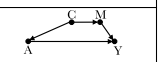
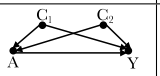
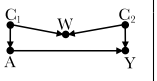
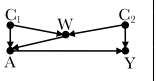
	Causal graph	$P(y_a)$
$G1$		$P(y\|a)$
$G2$		
$G3$		
$G4$		
$G5$		
$G6$		$\sum_C P(y\|a,c)\ P(c)$
$G7$		
$G8$		
$G9$		
$G10$		$\sum_{C_1C_2} P(y\|a,c_1,c_2)\ P(c_1,c_2)$
$G11$		$P(y\|a)$ $\sum_{C_1} P(y\|a,c_1)\ P(c_1)$ $\sum_{C_2} P(y\|a,c_2)\ P(c_2)$ $\sum_{WC_1} P(y\|a,w,c_1)\ P(w,c_1)$ $\sum_{WC_2}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

TABLE I: $P(y_a)$ of some Markovian models.

[5]Called also adjustment formula or stratification.

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

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. Table I 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. 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.

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 [14] 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.

2) do-calculus inference rules: do-calculus [13], [14] 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.

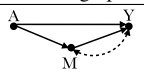
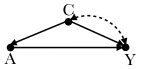
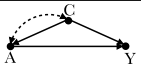
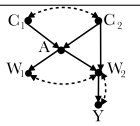
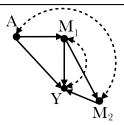
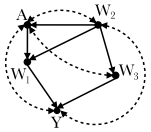
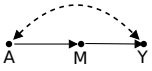
	Causal graph	$P(y_a)$
$G13$		$P(y\|a)$
$G14$		$\sum_C P(y\|a,c)\ P(c)$
$G15$		
$G16$		$\sum_{c_1,c_2} P(y\|a,c_1,c_2)\ P(c_1,c_2)$
$G17$		$\sum_{m_1,m_2}\ P(y\|m_1,m_2,a)\ P(m_1\|a)$ $\times \sum_{a'}\ P(m_2\|m_1,a')\ P(a')$
$G18$		$\sum_{w_1}\sum_{w_2}\sum_{a'}\ P(y\|w_1,w_2,a')$ $\times\ P(a'\|w_2)\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

TABLE II: $P(y_a)$ of some semi-Markovian models.

As an example, consider the graph $G18$. The causal effect $P(y_a)$ can be identified as follows:

$$P(y_a) = \sum_{w_1}\sum_{w_2}\sum_{w_3}\ P(y|w_1,w_3)\ P(w_2)\ P(w_1, y|\ do(a)) \tag{7}$$

$$= \sum_{w_1}\sum_{w_2}\ P(y|w_1,w_2)\ P(w_2)\ P(w_1, y|\ do(a)) \tag{8}$$

$$= \sum_{w_1}\sum_{w_2}\ P(y|w_1,w_2)\ P(w_2)\ P(w_1|\ a, w_2)$$
$$\times\ \sum_{a'} P(y|\ w_1, a')\ P(a'|\ w_2) \tag{9}$$

$$= \sum_{w_1}\sum_{w_2}\sum_{a'} P(y|w_1,w_2,a')\ P(w_2)\ P(w_1|\ a, w_2)$$
$$\times\ P(a'|\ w_2) \tag{10}$$

The term $P(w_1, y|\ \ do(a))$ in (8) 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 [14]). 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 (10).

3) C-component factorization: C-component factorization [21] aims to express the observational distribution $P(v)$ as a product of factors $P_{v\setminus s}(s)$, 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 $S_2 = \{W_1\}$.

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 [24]:

$$P(y_a) = \frac{P(y)}{Q^A} \sum_{a'} Q^A \tag{11}$$

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

$$Q^A = \prod_{v \in S^A} P(v|\mathbf{v}^{-1}) \tag{12}$$

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. 11 to $G17$ leads to:

$$P(y_a) = \frac{P(y)}{P(a)P(m_2|\ a, m_1)} \sum_{a'} P(a')P(m_2|\ a', m_1) \tag{13}$$

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:

$$P(y_a) = \sum_m\ P(m|a)\ \sum_{a'} P(y|m, a')\ P(a') \tag{14}$$

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 IV-2).

V. Identification of counterfactual effects

While causal effects (Section IV) 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 [14]). 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).

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^*[7] $= a_1$. 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 simple Markovian graph in Figure 3(a), we need to construct the semi-Markovian graph in Figure 3(b). The **make-cg** algorithm [21] 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 $do(a)$.

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

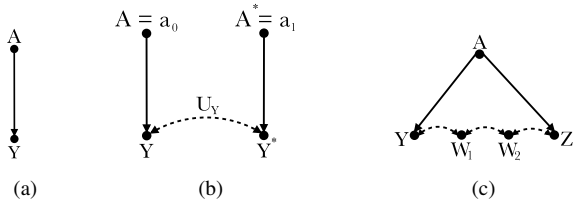


Fig. 3: (a) A causal graph for the firing example (b) A corresponding counterfactual graph for the query $P(Y^*_{a*=a_1} \mid A = a_0)$(c) Zig-zag pattern.

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.

A. C-component factorization

To illustrate how counterfactual quantities are measured, consider the same firing example (Figure 2 where the latent variable U is now replaced with an observable variable C. 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(a) shows the two parallel-worlds graph[8] for the query while Figure 4(b) presents the final constructed counterfactual graph using **make-cg** algorithm. Note that in Figure 4(b), C and C^* are merged as a single node C (by applying Lemma 24 [21]). 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 [21].

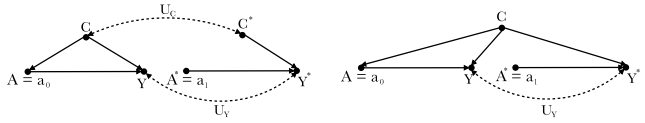


Fig. 4: (a) Parallel worlds graph for $P(y_{a_1}|a_0)$ (b) Counterfactual graph for $P(y_{a_1}|a_0)$.

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(b)) has three c-components: $\{C\}, \{A\}, \{Y, Y^*_{a_1}\}$ thus:

$$P(y_{a_1}|a_0) = \frac{\sum_{y,c}\ Q(c)\ Q(a_0)\ Q(y, y_{a_1})}{P(a_0)} \tag{15}$$

[8]This graph is called *twin network* graph since it includes only two hypothetical graphs [21].

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

$$P(y_{a_1}|a_0) = \frac{\sum_{y,c}\ P(c)\ P(a_0|c)\ P(y, y_{a_1}|c)}{P(a_0)}$$
$$= \frac{\sum_{c}\ P(c)\ P(a_0|c)\ P(y_{a_1}|c)}{P(a_0)} \tag{16}$$
$$= \frac{\sum_{c}\ P(c)\ P(a_0|c)\ P(y|a_1, c)}{P(a_0)} \tag{17}$$

Table III 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.

VI. 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 two types, namely the causal effect (intervention) and the counterfactual effect. Finally, we note that in the case 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.

References

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

[2] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

[3] S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[5] R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

[6] D. Galles and J. Pearl. Testing identifiability of causal effects. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 185–195, 1995.

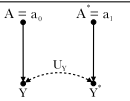
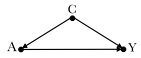
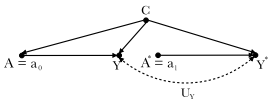
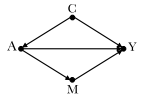
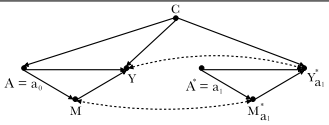
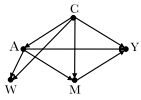
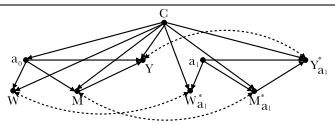
	Original Graph	Counterfactual Graph	$P(y_{a_1} \mid a_0)$
G31			unidentifiable if $y \neq y^*$
G32			$\frac{\sum_c P(c)\ P(a_0 \mid c)\ P(y \mid a_1, c)}{P(a_0)}$
G33			$\frac{\sum_{c,m} P(c)\ P(a_0 \mid c)\ P(y \mid m, m', a_1, c)\ P(m' \mid a_0, c, a_1)}{P(a_0)}$
G34			$\frac{\sum_c P(c)\ P(a_0 \mid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

TABLE III: Identifiability of counterfactual effects.

[7] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

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

[9] H. Kim and J. Perl. 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, 1983.

[10] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30:4066–4076, 2017.

[11] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[12] C. O'Neill. Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy*, 2016.

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

[14] J. Pearl. *Causality*. Cambridge university press, 2009.

[15] K. Quick. The unfair effects of impact on teachers with the toughest jobs. *The Century Foundation*, 2015.

[16] M. Rhee. Impact: The dcps evaluation and feedback system for school-based personnel, 2019.

[17] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

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

[19] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444, 2006.

[20] I. Shpitser and J. Pearl. What counterfactuals can be tested. In *23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*, pages 352–359, 2007.

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

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

[23] J. Tian. Identifying linear causal effects. In *Proceedings of the 2004 AAAI Conference on Artificial Intelligence*, pages 104–111, 2004.

[24] J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the 2002 AAAI Conference on Artificial Intelligence*, pages 567–573, 2002.

[25] J. Tian and I. Shpitser. On the identification of causal effects. 2003.

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

[27] Y. Wu, L. Zhang, and X. Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *IJCAI*, pages 1438–1444, 2019.

[28] Y. Wu, L. Zhang, X. Wu, and H. Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414, 2019.

[29] J. Zhang and E. Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the... AAAI Conference on Artificial Intelligence*, 2018.