Int J Data Sci Anal (2017) 3:1–21
DOI 10.1007/s41060-016-0040-z

REGULAR PAPER

Exposing the probabilistic causal structure of discrimination

Francesco Bonchi[1] · **Sara Hajian**[2] · **Bud Mishra**[3] · **Daniele Ramazzotti**[4]

Received: 15 May 2016 / Accepted: 6 December 2016 / Published online: 23 January 2017

Abstract *Discrimination discovery* from data is an important data mining task, whose goal is to identify patterns of illegal and unethical discriminatory activities against protected-by-law groups, e.g., ethnic minorities. While any legally valid proof of discrimination requires evidence of causality, the state-of-the-art methods are essentially correlation based, albeit, as it is well known, correlation does not imply causation. In this paper, we take a principled causal approach to discrimination detection following Suppes' *probabilistic causation theory*. In particular, we define a method to extract, from a dataset of historical decision records, the causal structures existing among the attributes in the data. The result is a type of constrained Bayesian network, which we dub *Suppes-Bayes causal network* (SBCN). Next, we develop a toolkit of methods based on random walks on top of the SBCN, addressing different anti-discrimination legal concepts, such as direct and indirect discrimination, group and individual discrimination, genuine requirement, and favoritism. Our experiments on real-world datasets confirm the inferential power of our approach in all these different tasks.

Keywords Discrimination discovery · Algorithmic discrimination · probabilistic causation · constrained Bayesian network · Random walks

1 Introduction

1.1 The importance of discrimination discovery

At the beginning of 2014, as an answer to the growing concerns about the role played by data mining algorithms in decision making, USA President Obama called for a 90-day review of big data collecting and analyzing practices. The resulting report[1] concluded that *"big data technologies can cause societal harms beyond damages to privacy"*. In particular, it expressed concerns about the possibility that decisions informed by big data could have discriminatory effects, even in the absence of discriminatory intent, further imposing less favorable treatment to already disadvantaged groups. It further expressed alarm about the threats of an "opaque decision-making environment" guided by an "impenetrable set of algorithms."

Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Human rights laws prohibit discrimination on several grounds, such as gender, age, marital status, sexual orientation, race, religion or belief, membership in a national minority, disability,

Our software together with the datasets used in the experiments are available at http://bit.ly/1GizSIG.

✉ Sara Hajian
sara.hajian@eurecat.org

Francesco Bonchi
francesco.bonchi@isi.it

Bud Mishra
mishra@nyu.edu

Daniele Ramazzotti
daniele.ramazzotti@disco.unimib.it

1 Algorithmic Data Analytics Lab, ISI Foundation, Turin, Italy

2 Eurecat-Technology Centre of Catalonia, Barcelona, Spain

3 New York University, New York, NY, USA

4 Milano-Bicocca University, Milan, Italy

[1] http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

or illness. Anti-discrimination authorities (such as equality enforcement bodies, regulation boards, consumer advisory councils) monitor, provide advice, and report on discrimination compliances based on investigations and inquiries. A fundamental role in this context is played by *discrimination discovery in databases*, i.e., the data mining problem of unveiling discriminatory practices by analyzing a dataset of historical decision records.

1.2 Discrimination is causal

According to current legislation, discrimination occurs when a group is treated "less favorably" [20] than others, or when "a higher proportion of people not in the group is able to comply" with a qualifying criterion [21]. Although these definitions do not directly imply causation, as stated in [9] all discrimination claims require plaintiffs to demonstrate a causal connection between the challenged outcome and a protected status characteristic. In other words, in order to prove discrimination, authorities must answer the counterfactual question: what would have happened to a member of a specific group (e.g., nonwhite), if he or she had been part of another group (e.g., white)?

"The Sneetches," the popular satiric tale[2] against discrimination published in 1961 by Dr. Seuss, describes a society of yellow creatures divided into two races: the ones with a green star on their bellies and the ones without. The Star-Belly Sneetches have some privileges that are instead denied to Plain-Belly Sneetches. There are, however, Star-On and Star-Off machines that can make a Plain-Belly into a Star-Belly, and viceversa. Thanks to these machines, the causal relationship between race and privileges can be clearly measured, because stars can be placed on or removed from any belly, and multiple outcomes can be observed for an individual. Therefore, we could readily answer the counterfactual question, saying with certainty what would have happened to a Plain-Belly Sneetch had he or she been a Star-Belly Sneetch. In the real world, however, proving discrimination episodes is much harder, as we cannot manipulate race, gender, or sexual orientation of an individual. This limitation highlights the need to assess discrimination as a causal inference problem [6] from a database of past decisions, where causality can be inferred probabilistically.

Unfortunately, *the state of the art of data mining methods for discrimination discovery (surveyed in* Sect. 3*) does not properly address the causal question*, as it is mainly correlation based.

1.3 Correlation is not causation

It is well known that correlation between two variables does not necessarily imply that one causes the other. Consider a unique cause X of two effects, Y and Z: if we do not take into account X, we might derive wrong conclusions because of the observable correlation between Y and Z. In this situation, X is said to act as a *confounding factor* for the relationship between Y and Z.

Variants of the complex relation just discussed can arise even if, in the example, X is not the actual cause of either Y or Z, but it is only correlated to them, for instance, because of how the data were collected. Consider for instance a credit dataset where there exists high correlation between a variable representing low income and another variable representing loan denial and let us assume that this is due to an actual legitimate causal relationship in the sense that, legitimately, a loan is denied if the applicant has low income. Let us now assume that we also observe high correlation between low income and being female, which, for instance, can be due to the fact that the women represented in the specific dataset in analysis, tend to be underpaid. Given these settings, in the data we would also observe high correlation between the variable gender being female and the variable representing loan denial, due to the fact that we do not account for the presence of the variable low income. Following common terminologies, we will say that such situations are due to *spurious correlations*.

However, the picture is even more complicated: it could be the case, in fact, that being female is the actual cause of the low income and, hence, be the *indirect cause* of loan denial *through* low income. This would represent a causal relationship between the gender and the loan denial, that we would like to detect as discrimination. Disentangling these two different cases, i.e., female is only correlated with low income in a spurious way, or being female is the actual cause of low income, is at the same time important and challenging. This highlights the need for a principled causal approach to discrimination detection.

Another typical pitfall of correlation-based reasoning is expressed by what is known as Simpson's paradox[3] according to which, correlations observed in different groups might disappear when these heterogeneous groups are aggregated, leading to *false-positive* cases of discrimination discovery. One of the most famous false-positive examples due to Simpson's paradox occurred when in 1973 the University of California, Berkeley, was sued for discrimination against women who had applied for admission to graduate schools. In fact, by looking at the admissions of 1973, it first appeared that men applying were significantly more likely to be admitted than women. But later, by examining the individual departments carefully, it was discovered that none of them was significantly discriminating against women. On the contrary, most departments had exercised a small bias in favor of women. The apparent discrimination was due to the fact

[2] http://en.wikipedia.org/wiki/The_Sneetches_and_Other_Stories.

[3] http://en.wikipedia.org/wiki/Simpson's_paradox.

that women tended to apply to departments with lower rates of admission, while men tended to apply to departments with higher rates [1]. Later in Sect. 6.6 we will use the dataset from this episode to highlight the differences between correlation-based and causation-based methods.

Another very recent example is the scientific "debate" on PNAS (December 2015) where Volker and Steenbeek [43] reacted to a previous article of Van der Lee and Ellemers [42] which analyzed data about research grant in the Netherlands and claimed a gender bias. In their reaction, Volker and Steenbeek state that the overall gender effect borders on statistical significance and that the conclusion of Van der Lee and Ellemers could be a prime example of Simpson's paradox. This example again highlights the importance, timeliness, and hardness of discrimination detection and the need for principled causal approaches.

Our proposal and contributions. In this paper, we take a principled causal approach to the data mining problem of discrimination detection in databases. Following Suppes' *probabilistic causation theory* [13,41], we define a method to extract, from a dataset of historical decision records, the causal structures existing among the attributes in the data.

In particular, we define the *Suppes-Bayes causal network* (SBCN), i.e., a directed acyclic graph (DAG) where we have a node representing a Bernulli variable of the type $\langle attribute = value \rangle$ for each pair attribute value present in the database. In this DAG an arc (A, B) represents the existence of a causal relation between A and B (i.e., A causes B). Moreover, each arc is labeled with a score, representing the strength of the causal relation.

Our SBCN is a constrained Bayesian network reconstructed by means of maximum likelihood estimation (MLE) from the given database, where we force the conditional probability distributions induced by the reconstructed graph to obey Suppes' constraints, i.e., *temporal priority* and *probability rising*. Imposing Suppes' temporal priority and probability raising, we obtain what we call the *prima facie causes* graph [41], which might still contain spurious causes (false positives). In order to remove these spurious causes, we add a bias term to the likelihood score, favoring sparser causal networks: in practice, we sparsify the *prima facie causes* graph by extracting a minimal set of edges which best explain the data. This regularization is done by means of the Bayesian information criterion (BIC) [40].

The obtained SBCN provides a clear summary, amenable to visualization, of the probabilistic causal structures found in the data. Such structures can be used to reason about different types of discrimination. In particular, we show how using several random-walk-based methods, where the next step in the walk is chosen proportionally to the edge weights, we can address different anti-discrimination legal concepts. This step makes SBCN a very general tool for discrimination detection. Our experiments show that the measures of discrimination produced by our methods are very strong, almost binary, signals: our measures are very clearly separating the discrimination and the non-discrimination cases.

To the best of our knowledge, this is the first proposal of discrimination detection in databases grounded in probabilistic causal theory.

1.4 Roadmap

The rest of the paper is organized as follows. In the next section, we provide some basic definitions of discrimination, and then in Sect. 3, we discuss the state of the art in discrimination detection in databases. In Sect. 4, we formally introduce the SBCN and present the method for extracting such causal network from the input dataset. In Sect. 5, we show how to exploit the SBCN for different concepts of discrimination detection, by means of random-walk-based methods. Finally, Sect. 6 presents our experimental assessment and comparison with correlation-based methods on four real-world datasets.

2 Definitions of discrimination

Different technical definitions of discrimination are based on different legal principles. Differently from privacy legislation, anti-discrimination legislation is very diverse and includes different legal concepts, e.g., direct and indirect discrimination, group and individual discrimination and the so-called genuine occupational requirement. Here, we present the detailed definitions of different anti-discrimination legal concepts that will be addressed by our methods in Sect. 5.

2.1 Group discrimination

According to current legislation, discrimination occurs when a group is treated "less favorably" [20] than others, or when "a higher proportion of people not in the group is able to comply" [21] with a qualifying criterium. Thus, groups discrimination reflects the structural bias against groups. It is also known as inequality of outcomes or disparate impact. To quantify the degree of group discrimination, several discrimination measures have been defined over a fourfold contingency table [31], as shown in Fig. 1, where: the *protected* group is a social group which is suspected of being discriminated against; the *decision* is a binary attribute recording whether a benefit was granted (value "+") or not (value "−") to an individual; the *total population* denotes a context of possible discrimination, such as individuals from a specific city, job sector, income, or combination thereof.

Different outcomes between groups are measured in terms of the proportion of people in each group with a specific out-

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$p_1 = a/n_1$
$p_2 = c/n_2$
$p = m_1/n$
$RR = \frac{p_1}{p_2}$
$RD = p_1 - p_2$

Fig. 1 Discrimination contingency table

come. Figure 1 considers the proportions of benefits denied to the protected group (p_1), the unprotected group (p_2), and the overall population (p). Differences or rates of these proportions can model the legal principle of group underrepresentation of the protected group in positive outcomes or, equivalently, of overrepresentation in negative outcomes.

Once provided with a threshold α between "legal" and "illegal" degree of discrimination, we can isolate contexts of possible discrimination [39].

2.2 Individual discrimination

Individual discrimination occurs when an individual treated differently because of his/her sensitive features such as race, color, religion, nationality, sex, marital status, age, and pregnancy. In other words, individual discrimination occurs if individuals with similar abilities (qualifications) are treated differently [7,25]. Individual discrimination requires to measure the amount of discrimination for a specific individual, i.e., an entire record in the database. For an individual record r in original data table D, individual discrimination quantifies as follows [25]:

$$\text{diff(r)} = p_x - p_y,$$

where p_x is the proportions of benefits denied to the protected group in the $2k$ closest neighborhoods of r and p_y is the proportions of benefits denied to the unprotected group in the $2k$ closest neighborhoods of r. As observed by Dwork et al. [7] and others, removing group discrimination does not prevent discrimination at an individual level. This highlights the need for measuring and accessing discrimination at an individual level.

2.3 Favoritism

Favoritism refers to the case of an individual treated better than others for reasons not related to individual merit or business necessity: for instance, favoritism in the workplace might result in a person being promoted faster than others unfairly.

2.4 Indirect discrimination

The European Union Legislation [21] provides a broad definition of indirect discrimination as occurring "where an apparently neutral provision, criterion, or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons." In other words, the actual result of the apparently neutral provision is the same as an explicitly discriminatory one. A typical legal case study of indirect discrimination is concerned with *redlining*: e.g., denying a loan because of ZIP code, which in some areas is an attribute highly correlated to race. Therefore, even if the attribute race cannot be required at loan application time (thus would not be present in the data), still race discrimination is perpetrated.

2.5 Genuine requirement

The legal concept of genuine requirement refers to detecting that part of the discrimination which may be explained by other, legally grounded, attributes; for example, denying credit to women may be explainable by the fact that most of them have low salary or delay in returning previous credits. A typical example in the literature is the one of the "genuine occupational requirement," also called "business necessity" in [8,22].

3 Related work

Discrimination analysis is a multi-disciplinary problem, involving sociological causes, legal reasoning, economic models, and statistical techniques [5,36]. Some authors [11,17] study how to prevent data mining from becoming itself a source of discrimination. In this paper, instead we focus on the data mining problem of detecting discrimination in a dataset of historical decision records, and in the rest of this section, we present the most relevant literature.

Pedreschi et al., see [30,31] and [39], propose a technique based on extracting classification rules (inductive part) and ranking the rules according to some legally grounded measures of discrimination (deductive part). The result is a (possibly large) set of classification rules, providing local and overlapping niches of possible discrimination. This model only deals with group discrimination.

Luong et al. [25] exploit the idea of *situation-testing* [37] to detect individual discrimination. For each member of the protected group with a negative decision outcome, testers with similar characteristics (k-nearest neighbors) are considered. If there are significantly different decision outcomes between the testers of the protected group and the testers of the unprotected group, the negative decision can be ascribed to discrimination.

Zliobaite et al. [45] focus on the concept of *genuine requirement* to detect that part of discrimination which may be explained by other, legally grounded, attributes. In [7], Dwork et al. address the problem of fair classification that

achieves both group fairness, i.e., the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole, and individual fairness, i.e., similar individuals should be treated similarly.

The above approaches assume that the dataset under analysis contains attributes that denote protected groups (i.e., direct discrimination). This may not be the case when such attributes are not available, or not even collectible at a micro-data level as in the case of the loan applicant's race. In these cases, we talk about indirect discrimination discovery. Ruggieri et al. [29,38] adopt a form of rule inference to cope with the indirect discovery of discrimination. The correlation information is called background knowledge and is itself coded as an association rule.

Mancuhan and Clifton [26] propose Bayesian networks as a tool for discrimination discovery. Bayesian networks consider the dependence between all the attributes and use these dependencies in estimating the joint probability distribution without any strong assumption, since a Bayesian network graphically represents a factorization of the joint distribution in terms of conditional probabilities encoded in the edges. Although Bayesian networks are often used to represent causal relationships, this needs not be the case; in fact, a directed edge from two nodes of the network does not imply any causal relation between them. As an example, let us observe that the two graphs $A \rightarrow B \rightarrow C$ and $C \rightarrow B \rightarrow A$ impose exactly the same conditional independence requirements and, hence, any Bayesian network would not be able to disentangle the direction of any causal relationship among these events.

Our work departs from this literature as:

1. It is grounded in probabilistic causal theory instead of being based on correlation;
2. It proposes a holistic approach able to deal with different types of discrimination in a single unifying framework, while the methods in the state of the art usually deal with one and only one specific type of discrimination;
3. It is the first work to adopt graph theory and social network analysis concepts, such as random-walk-based centrality measures and community detection, for discrimination detection.

Our proposal has also lower computational cost than methods such as [30,31] and [39] which require computing a potentially exponential number of association/classification rules.

4 Suppes-Bayes causal network

Theories of causality are old and enjoy contributions from many fields. Some of the most prominent results are due to Judea Pearl [28], whose theories have been of great impact in the computational community. However, algorithms derived from this theory may result to be computationally intractable. For this reason, in this paper we follow a different approach based on the theory of probabilistic causation by Patrick Suppes [13] which we combine with state-of-the-art Bayesian learning approach, in order to build an effective method but still keeping its complexity tractable. More details about different approaches to evaluate causal claims and justification of our choice are presented in Appendix.

In order to study discrimination as a causal inference problem, we exploit the criteria defined in the theories of *probabilistic causation* [13]. In particular, we follow [41], where Suppes proposed the notion of *prima facie causation* that is at the core of probabilistic causation. Suppes' definition is based on two pillars: (i) any cause must happen before its effect (*temporal priority*) and (ii) it must raise the probability of observing the effect (*probability raising*).

Definition 1 (*Probabilistic causation* [41]) For any two events h and e, occurring, respectively, at times t_h and t_e, under the mild assumptions that $0 < P(h), P(e) < 1$, the event h is called a prima facie cause of the event e if it occurs before the effect and the cause raises the probability of the effect, i.e., $t_h < t_e$ and $P(e \mid h) > P(e \mid \neg h)$.

In the rest of this section, we introduce our method to construct, from a given relational table D, a type of causal Bayesian network constrained to satisfy the conditions dictated by Suppes' theory, which we dub *Suppes-Bayes causal network* (SBCN).

In the literature, many algorithms exist to carry out structural learning of general Bayesian networks and they usually fall into two families [19]. The first family, *constraint based learning*, explicitly tests for pairwise independence of variables conditioned on the power set of the rest of the variables in the network. These algorithms exploit structural conditions defined in various approaches to causality [13,15,18,27,28,44]. The second family, *score-based learning*, constructs a network which maximizes the likelihood of the observed data with some regularization constraints to avoid overfitting. Several hybrid approaches have also been recently proposed [2,24,34].

Our framework can be considered a hybrid approach exploiting *constrained maximum likelihood estimation* (MLE) as follows: (i) we first define all the possible causal relationship among the variables in D by considering only the oriented edges between events that are consistent with Suppes' notion of probabilistic causation and, subsequently, (ii) we perform the reconstruction of the SBCN by a score-based approach (using BIC), which considers only the valid edges.

We next present in details the whole learning process.

4.1 Suppes' constraints

We start with an input relational table D defined over a set A of h categorical attributes and s samples. In case continuous numerical attributes exist in D, we assume they have been discretized to become categorical. From D, we derive D', an $m \times s$ binary matrix representing m Bernoulli variables of the type $\langle attribute = value \rangle$, where an entry is 1 if we have an observation for the specific variable and 0 otherwise.

4.1.1 Temporal priority

The first constraint, temporal priority, cannot be simply checked in the data as we have no timing information for the events. In particular, in our context the events for which we want to reason about temporal priority are the Bernoulli variables $\langle attribute = value \rangle$.

The idea here is that, e.g., $income = low$ cannot be a cause of $gender = female$, because the time when the gender of an individual is determined is antecedent to that of when the income is determined. This intuition is implemented by simply letting the data analyst provide as input to our framework a partial temporal order $r : A \to \mathbb{N}$ for the h attributes, which is then inherited from the m Bernoulli variables[4].

Based on the input dataset D and the partial order r, we produce the first graph $G = (V, E)$ where we have a node for each of the Bernoulli variables, so $|V| = m$, and we have an arc $(u, v) \in E$ whenever $r(u) \leq r(v)$. This way we will immediately rule out causal relations that do not satisfy the temporal priority constraint.

4.1.2 Probability raising

Given the graph $G = (V, E)$ built as described above the next step requires pruning the arcs which do not satisfy the second constraint, probability raising, thus building $G' = (V, E')$, where $E' \subseteq E$. In particular, we remove from E each arc (u, v) such that $P(v \mid u) \leq P(v \mid \neg u)$. The graph G' so obtained is called *prima facie* graph.

We recall that the probability raising condition is equivalent to constraining for positive statistical dependence [24]: in the prima facie graph we model *all and only* the positive correlated relations among the nodes already partially ordered by temporal priority, consistent with Suppes' characterization of causality in terms of relevance.

[4] Note that our learning technique requires the input order r to be correct and complete in order to guarantee its convergence. Nevertheless, if this is not the case, it is still capable of providing valuable insights about the underlying causal model, although with the possibility of false-positive or false-negative causal claims.

4.2 Network simplification

Suppes' conditions are necessary but not sufficient to evaluate causation [34]: especially when the sample size is small, the model may have false positives (spurious causes), even after constraining for Suppes' temporal priority and probability raising criteria (which aim at removing false negatives). Consequently, although we expect all the statistically relevant causal relations to be modeled in G', we also expect some spurious ones in it.

In our proposal, in place of other structural conditions used in various approaches to causality (see, e.g., [13,27,44]), we perform a network simplification (i.e., we sparsify the network by removing arcs) with a score-based approach, specifically by relying on the Bayesian information criterion (BIC) as the regularized likelihood score [40].

We consider as inputs for this score the graph G' and the dataset D'. Given these, we select the set of arcs $E^* \subseteq E'$ that maximizes the score:

$$\text{score}_{\text{BIC}}(D', G') = LL(D'|G') - \frac{\log s}{2}\dim(G').$$

In the equation, G' denotes the graph, D' denotes the data, s denotes the number of samples, and $\dim(G')$ denotes the number of parameters in G'. Thus, the regularization term $-\dim(G')$ favors graphs with fewer arcs. The coefficient $\log s/2$ weighs the regularization term, such that the higher the weight, the more sparsity will be favored over "explaining" the data through maximum likelihood. Note that the likelihood is implicitly weighted by the number of data points, since each point contributes to the score.

Assume that there is one *true* (but unknown) probability distribution that generates the observed data, which is, eventually, uniformly randomly corrupted by false positives and negatives rates (in $[0, 1)$). Let us call *correct model*, the statistical model which best approximate this distribution. The use of BIC on G' results in removing the false positives and, asymptotically (as the sample size increases), converges to the correct model. In particular, BIC is attempting to select the candidate model corresponding to the highest Bayesian Posterior probability, which can be proved to be equivalent to the presented score and its $log(s)$ penalization factor.

We denote with $G^* = (V, E^*)$ the graph that we obtain after this step. We note that, as for general Bayesian network, G^* is a DAG by construction.

4.3 Confidence score

Using the reconstructed SBCN, we can represent the probabilistic relationships between any set of events (nodes). As an example, suppose to consider the nodes representing, respec-

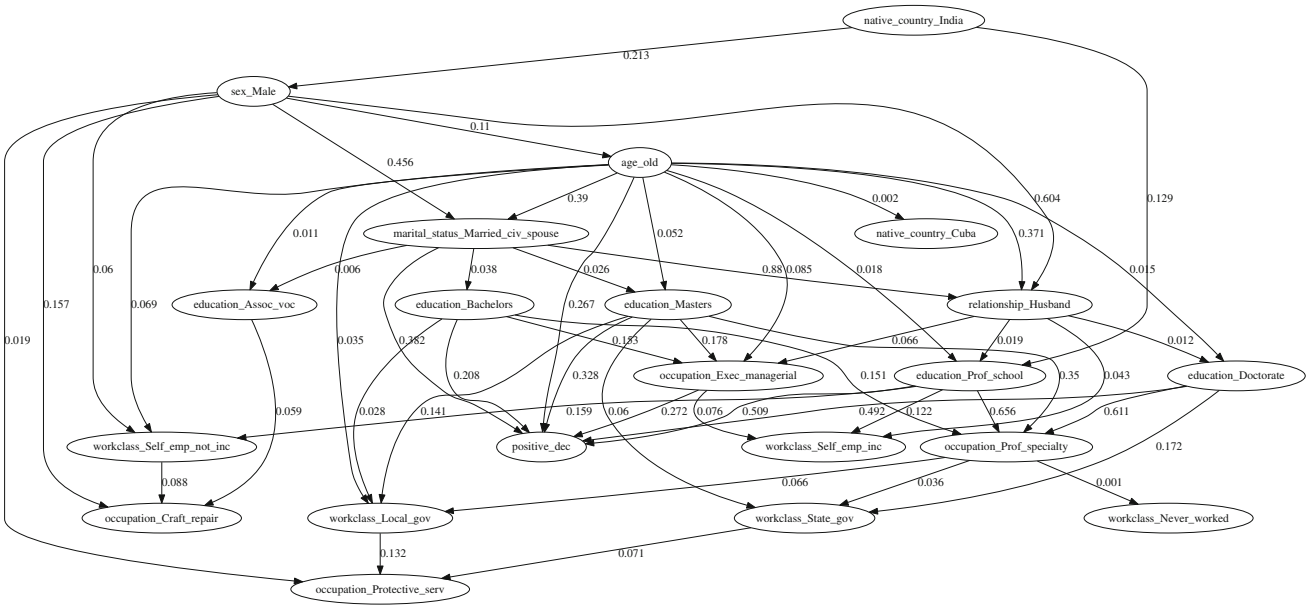

Fig. 2 One portion of the SBCN extracted from the Adult dataset. This subgraph corresponds to the C_2 community reported later in Table 3 (Sect. 6) extracted by a community detection algorithm

tively, $income = low$ and $gender = female$ being the only two direct causes (i.e., with arcs toward) of $loan = denial$. Given SBCN, we can estimate the conditional probabilities for each node in the graph, i.e., probability of $loan = denial$ given $income = low\ AND\ gender = female$ in the example, by computing the conditional probability of only the pair of nodes directly connected by an arc. For an overview of state-of-the-art methods for doing this, see [19]. However, we expect to be mostly dealing with full data, i.e., for every directly connected node in the SBCN, we expect to have several observations of any possible combination $attribute = value$. For this reason, we can simply estimate the node probabilities by counting the observations in the data. Moreover, we will exploit such conditional probabilities to define the confidence score of each arc in terms of their causal relationship.

In particular, for each arc $(v, u) \in E^*$ involving the causal relationship between two nodes $u, v \in V$, we define a confidence score $W(v, u) = P(u \mid v) - P(u \mid \neg v)$, which, intuitively, aims at estimating the observations where the cause v is followed by its effect u, that is, $P(u \mid v)$, and the ones where this is not observed, i.e., $P(u \mid \neg v)$, because of imperfect causal regularities. We also note that, by the constraints discussed above, we require $P(u \mid v) \gg P(u \mid \neg v)$ and, for this reason, each weight is positive and no larger than 1, i.e., $W : E^* \to (0, 1]$.

Combining all of the concepts discussed above, we conclude with the following definition.

Definition 2 *(Suppes-Bayes causal network)* Given an input dataset D' of m Bernoulli variables and s samples, and given a partial order r of the variables, the Suppes-Bayes causal network SBCN $= (V, E^*, W)$ subsumed by D' is a weighted DAG such that the following requirements hold:

- **[Suppes' constraints]** for each arc $(v, u) \in E^*$ involving the causal relationship between nodes $u, v \in V$, under the mild assumptions that $0 < P(u), P(v) < 1$:

$$r(v) \leq r(u) \quad \text{and} \quad P(u \mid v) > P(u \mid \neg v)\,.$$

- **[Simplification]** let E' be the set of arcs satisfying the Suppes' constraints as before; among all the subsets of E', the set of arcs E^* is the one whose corresponding graph maximizes BIC:

$$E^* = \arg\max_{E \subseteq E', G=(V,E)} (LL(D'|G) - \frac{\log s}{2} \dim(G))\,.$$

- **[Score]** $W(v, u) = P(u \mid v) - P(u \mid \neg v),\ \forall (v, u) \in E^*$

An example of a portion of a SBCN extracted from a real-world dataset is reported in Fig. 2.

Algorithm 1 summarizes the learning approach adopted for the inference of the SBCN . Given D' an input dataset over m Bernoulli variables and s samples, and r a partial order of the variables, Suppes' constraints are verified (Lines 4–9) to construct a DAG as described in Sect. 4.1.

The likelihood fit is performed by hill climbing (Lines 12–21), an iterative optimization technique that starts with an arbitrary solution to a problem (in our case an empty graph)

and then attempts to find a better solution by incrementally visiting the neighborhood of the current one. If the new candidate solution is better than the previous one it is considered in place of it. The procedure is repeated until the stopping criterion is matched.

The !*StoppingCriterion* occurs (Line 12) in two situations: (i) the procedure stops when we have performed a large enough number of iterations or (ii) it stops when none of the solutions in $G_{\text{neighbors}}$ is better than the current G_{fit}. Note that $G_{\text{neighbors}}$ denotes all the solutions that are derivable from G_{fit} by removing or adding at most one edge.

Algorithm 1 Learning the SBCN

1: Inputs: D' an input dataset of m Bernoulli variables and s samples, and r a partial order of the variables
2: Output: $SBCN(V, E^*, W)$ as in Definition 2
3: **[Suppes' constraints]**
4: **for all** pairs (v, u) among the m Bernoulli variables **do**
5: **if** $r(v) \leq r(u)$ **and** $P(u \mid v) > P(u \mid \neg v)$ **then**
6: add the arc (v, u) to $SBCN$.
7: **end if**
8: **end for**
9: **[Likelihood fit by hill climbing]**
10: Consider $G(V, E^*, W)_{fit} = \emptyset$.
11: **while** !*StoppingCriterion*() **do**
12: Let $G(V, E^*, W)_{neighbors}$ be the neighbor solutions of $G(V, E^*, W)_{fit}$.
13: Remove from $G(V, E^*, W)_{neighbors}$ any solution whose arcs are not included in *SBCN*.
14: Consider a random solution $G_{current}$ in $G(V, E^*, W)_{neighbors}$.
15: **if** $score_{BIC}(D', G_{current}) > score_{BIC}(D', G_{fit})$ **then**
16: $G_{fit} = G_{current}$.
17: $\forall$ arc (v, u) of G_{fit}, $W(v, u) = P(u \mid v) - P(u \mid \neg v)$.
18: **end if**
19: **end while**
20: $SBCN = G_{fit}$.
21: **return** *SBCN*.

4.3.1 Time and space complexity

The computation of the valid DAG according to Suppes' constraints (Lines 4–10) requires a pairwise calculation among the m Bernoulli variables. After that, the likelihood fit by hill climbing (Lines 11–21) is performed. Being an heuristic, the computational cost of hill climbing depends on the stopping criterion. However, constraining by Suppes' criteria tends to regularize the problem leading on average to a quick convergence to a good solution. The time complexity of Algorithm 1 is $O(sm)$ and the space required is $O(m^2)$, where m, however, is usually not too large, being the number of attribute-value pairs, and not the number of examples.

4.4 Expressivity of a SBCN

We conclude this section with a discussion on the causal relations that we model by a *SBCN*.

Let us assume that there is one true (but unknown) probability distribution that generates the observed data whose structure can be modeled by a DAG. Furthermore, let us consider the causal structure of such a DAG and let us also assume each node with more then one cause to have conjunctive parents: any observation of the child node is preceded by the occurrence of all its parents. As before we call correct model, the statistical model which best approximates the distribution. On these settings, we can prove the following theorem.

Theorem 1 *Let the sample size $s \to \infty$, the provided partial temporal order r be correct and complete and the data be uniformly randomly corrupted by false-positive and false-negative rates (in $[0, 1)$), then the $SBCN$ inferred from the data is the correct model.*

Proof (Sketch) Let us first consider the case where the observed data have no noise. On such an input, we observe that the prima facie graph has no false negatives: in fact $\forall[c \to e]$ modeling a genuine causal relation, $P(e \wedge c) = P(e)$, thus the probability raising constraint is satisfied, so it is the temporal priority given that we assumed r to be correct and complete.

Furthermore, it is known that the likelihood fit performed by BIC converges to a class of structures equivalent in terms of likelihood among which there is the correct model: all these topologies are the same unless the directionality of some edges. But, since we started with the prima facie graph which is already ordered by temporal priority, we can conclude that in this case the $SBCN$ coincides with the correct model.

To extend the proof to the case of data uniformly randomly corrupted by false positives and negatives rates (in $[0, 1)$), we note that the marginal and joint probabilities change monotonically as a consequence of the assumption that the noise is uniform. Thus, all inequalities used in the preceding proof still hold, which concludes the proof. □

In the more general case of causal topologies where any cause of a common effect is independent from any other cause (i.e., we relax the assumption of conjunctive parents), the $SBCN$ is not guaranteed to converge to the correct model but it coincides with a subset of it modeling all the edges representing statistically relevant causal relations (i.e., where the probability raising condition is verified).

4.5 Learning the temporal ordering of variables

Up to this point, we have described how to infer the structure of a Suppes-Bayes causal network assuming a given partial temporal order of the variables of the database: as an example, the variables `gender` and `DOB` are naturally defined earlier than `education`, which in turn comes earlier than `occupation`. Such a partial order is essential to capture Suppes' condition of temporal priority.

We next discuss how to automatically infer the partial temporal order, which allows learning the $SBCN$ also in the cases when it is not possible to define a priori the needed temporal order.

We start by observing that, while reconstructing a Bayesian Network is a known NP-hard problem, when an ordering among the variables in the network is given, finding the maximum likelihood network is not NP-hard anymore [3,4]. For this reason, we iterate our learning procedure by searching in the ordering space rather than the space of the directed acyclic graphs.

Intuitively, to do so we may iteratively consider all the possible partial orderings among the nodes. Given one, we may learn the structure of the SBCN normally, that is, by constraining for Suppes' probability raising and maximum likelihood estimation over the given order. The final structure would be the one at maximum likelihood among all the structures generated at the different possible orderings.

Obviously, given the intractability of evaluating of the solutions in the space of the partial orderings, we need once again to rely on some heuristic search strategy in order to perform this task. In particular, we will adopt the hill-climbing (HC) schema. HC is one of the simplest iterative techniques to solve optimization problems, which is based on the concept of neighborhood. It is a procedure that iteratively evaluates solutions in the search space and, for each candidate valid solution i, defines a neighborhood $N(i)$. Given $N(i)$, the solution j at the subsequent iteration is searched only in $N(i)$. Hence, the neighborhood is a function $N : S \rightarrow 2^S$ that assigns at each solution in the search space S a (non-empty) subset of S.

In our case, a solution is a partial order among the nodes, described as a rank from 0 to k where one or more nodes are associated with each position in the ranking and where all the lower ranked nodes are predecessors of higher ranked ones. Moreover, given any solution, we define its neighborhood as the set of partial orderings that can be reached from it with the following operations:

1. *Swap:* invert the order of 2 nodes;
2. *Increase:* subtract 1 to the rank of a node;
3. *Decrease:* add 1 to the rank of a node.

Given these premises, we now describe our learning procedure in Algorithm 2.

Algorithm 2 Learning the SBCN without manual temporal ordering

1: Being D' an input dataset of m Bernoulli variables and s samples.
2: Let $SBCN(V, E^*, W) = \emptyset$ represent a weighted DAG of the solutions.
3: **while** !$StoppingCriterion1()$ **do**
4: **[Initialization]**
5: **if** $SBCN(V, E^*, W) == \emptyset$ **then**
6: Set $r = r_{rand}$ a random partial order of the variables.
7: **else**
8: Set r randomly within $N(r)$, being the neighborhood of the current solution.
9: **end if**
10: Let $SBCN_{current}(V_c, E_c^*, W_c) = \emptyset$ represent a weighted DAG of the current solutions.
11: **[Suppes' constraints]**
12: **for all** the arcs (v, u) between each pair of the m Bernoulli variables **do**
13: **if** $r(v) \leq r(u)$ **and** $P(u \mid v) > P(u \mid \neg v)$ **then**
14: Set to the arc (v, u) its weight, being $W_c(v, u) = P(u \mid v) - P(u \mid \neg v)$.
15: Add the arc (v, u) to $SBCN_{current}$.
16: **end if**
17: **end for**
18: **[Simplification]**
19: Consider $G(V, E^*, W)_{fit} = \emptyset$.
20: **while** !$StoppingCriterion2()$ **do**
21: Let $G(V, E^*, W)_{neighbors}$ be the neighbor solutions of $G(V, E^*, W)_{fit}$.
22: Remove from $G(V, E^*, W)_{neighbors}$ any solution whose arcs are not included in $SBCN_{current}$.
23: Consider a random solution $G_{current}$ in $G(V, E^*, W)_{neighbors}$.
24: **if** $score_{BIC}(D', G_{current}) > score_{BIC}(D', G_{fit})$ **then**
25: $G_{fit} = G_{current}$.
26: Assign to the arcs of G_{fit} the related weights of $SBCN_{current}$.
27: **end if**
28: **end while**
29: $SBCN_{current} = G_{fit}$.
30: **[Update best solution]**
31: **if** $score_{BIC}(SBCN_{current}) > score_{BIC}(SBCN)$ **then**
32: $SBCN = G_{current}$.
33: **end if**
34: **end while**
35: **return** $SBCN$.

In the algorithm, !$StoppingCriterion1()$ and !$StoppingCriterion2()$ occur in two situations: (i) the procedure stops when we have performed a big enough number of iterations which is predefined or (ii) it stops when none of the solutions in the neighborhood are better than the current one.

Furthermore, we observed that, as for when the ordering of the nodes is provided as an input, Algorithm 2 first constrains the current solution for Suppes' conditions (Lines 12–17) and then performs the maximum likelihood estimation of the DAG by hill climbing (Lines 19–29). But this time, these steps are iterated along the neighbor orders by an outer hill-climbing procedure (Lines 5–10 and 31–33).

5 Discrimination discovery by random walks

In this section, we propose several random-walk-based methods over the reconstructed SBCN, to deal with different discrimination-detection tasks.

5.1 Group discrimination and favoritism

As defined in Sect. 2, the basic problem in the analysis of direct discrimination is precisely to quantify the degree of discrimination suffered by a given protected group (e.g., an ethnic group) with respect to a decision (e.g., loan denial). In contrast to discrimination, favoritism refers to the case of an individual treated better than others for reasons not related to individual merit or business necessity. In the following, we denote favoritism as positive discrimination in contrast to negative discrimination.

Given an SBCN we define a measure of group discrimination (either negative or positive) for each node $v \in V$. Recall that each node represents a pair $\langle attribute = value \rangle$, so it is essentially what we refer to as a group, e.g., $\langle gender = female \rangle$. Our task is to assign a score of discrimination $\mathrm{d}s^- : V \to [0, 1]$ to each node, so that the closer $\mathrm{d}s^-(v)$ is to 1 the more discriminated is the group represented by v.

We compute this score by means of a number n of random walks that start from v and reaches either the node representing the positive decision or the one representing the negative decision. In these random walks, the next step is chosen proportionally to the weights of the outgoing arcs. Suppose a random walk has reached a node u, and let $deg_{\mathrm{out}}(u)$ denote the set of nodes whish have an arc from u. Then, the arc (u, z) is chosen with probability

$$p(u, z) = \frac{W(u, z)}{\sum_{v \in deg_{\mathrm{out}}(u)} W(u, v)}.$$

When a random walk ends in a node with no outgoing arc before reaching either the negative or the positive decision, it is restarted from the source node v.

Definition 3 *(Group discrimination score)* Given an SBCN $= (V, E^*, W)$, let $\delta^- \in V$ and $\delta^+ \in V$ denote the nodes indicating the negative and positive decision, respectively. Given a node $v \in V$, and a number $n \in \mathbb{N}$ of random walks to be performed, we denote as $rw_{v \to \delta^-}$ the number of random walks started at node v that reach δ^- earlier than δ^+. The discrimination score for the group corresponding to node v is then defined as

$$\mathrm{d}s^-(v, \delta^-) = \frac{rw_{v \to \delta^-}}{n}.$$

This implicitly also defines a score of positive discrimination (or favoritism): $\mathrm{d}s^+(v, \delta^+) = 1 - \mathrm{d}s^-(v, \delta^-)$.

Taking advantage of the SBCN we also propose two additional measures capturing how far a node representing a group is from the positive and negative decision, respectively. This is done by computing the average number of steps that the random walks take to reach the two decisions: we denote these scores as $as^-(v)$ and $as^+(v)$.

5.2 Indirect discrimination

A typical legal case study of indirect discrimination is concerned with *redlining*, e.g., denying a loan because of ZIP code, which in some areas is an attribute highly correlated with race. Therefore, even if the attribute race cannot be required at loan application time (thus would not be present in the data), still race discrimination is perpetrated. Indirect discrimination discovery refers to the data mining task of discovering the attributes values that can act as a proxy to the protected groups and lead to discriminatory decisions indirectly [11,30,31]. In our setting, indirect discrimination can be detected by applying the same method described in Sect. 5.1.

5.3 Genuine requirement

In the state of the art of data mining methods for discrimination discovery, it is also known as *explainable discrimination* [12] and *conditional discrimination* [45].

The task here is to evaluate to which extent the discrimination apparent for a group is "explainable" on a legal ground. Let $v \in V$ be the node representing the group which is suspected of being discriminated, and $u_l \in V$ be a node whose causal relation with a negative or positive decision is legally grounded. As before, δ^- and δ^+ denote the negative and positive decision, respectively. Following the same random-walk process described in Sect. 5.1, we define the *fraction of explainable discrimination* for the group v:

$$\mathrm{fed}^-(v, \delta^-) = \frac{rw_{v \to u_l \to \delta^-}}{rw_{v \to \delta^-}},$$

i.e., the fraction of random walks passing trough u_l among the ones started in v and reaching δ^- earlier than δ^+. Similarly we define $\mathrm{fed}^+(v, \delta^+)$, i.e., the fraction of explainable positive discrimination.

5.4 Individual and subgroup discrimination

As defined in Sect. 2, individual discrimination requires measuring the amount of discrimination for a specific individual, i.e., an entire record in the database. Similarly, subgroup discrimination refers to discrimination against a subgroup described by a combination of multiple protected and non-protected attributes: personal data, demographics, social,

economic and cultural indicators, etc. For example, consider the case of gender discrimination in credit approval: although an analyst may observe that no discrimination occurs in general, it may turn out that older women obtain car loans only rarely.

Both problems can be handled by generalizing the technique introduced in Sect. 5.1 to deal with a set of starting nodes, instead of only one. Given an $SBCN = (V, E^*, W)$ let $v_1, \ldots, v_n$ be the nodes of interest. In order to define a discrimination score for $v_1, \ldots, v_n$, we perform a *personalized PageRank* [16] computation with respect to $v_1, \ldots, v_n$. In personalized PageRank, the probability of jumping to a node when abandoning the random walk is not uniform, but it is given by a vector of probabilities for each node. In our case, the vector will have the value $\frac{1}{n}$ for each of the nodes $v_1, \ldots, v_n \in V$ and zero for all the others. The output of personalized PageRank is a score $\text{ppr}(u|v_1, \ldots, v_n)$ of proximity/relevance to $\{v_1, \ldots, v_n\}$ for each other node u in the network. In particular, we are interested in the score of the nodes representing the negative and positive decision, i.e., $\text{ppr}(\delta^-|v_1, \ldots, v_n)$ and $\text{ppr}(\delta^+|v_1, \ldots, v_n)$, respectively.

Definition 4 *(Generalized discrimination score)* Given an $SBCN = (V, E^*, W)$, let $\delta^- \in V$ and $\delta^+ \in V$ denote the nodes indicating the negative and positive decision, respectively. Given a set of nodes $v_1, \ldots, v_n \in V$, we define the generalized (negative) discrimination score for the subgroup or the individual represented by $\{v_1, \ldots, v_n\}$ as

$$gds^-(v_1, \ldots, v_n, \delta^-, \delta^+) = \frac{\text{ppr}(\delta^-|v_1, \ldots, v_n)}{\text{ppr}(\delta^-|v_1, \ldots, v_n) + \text{ppr}(\delta^+|v_1, \ldots, v_n)}.$$

This implicitly also defines a generalized score of positive discrimination: $gds^+(v_1, \ldots, v_n, \delta^-, \delta^+) = 1 - gds^-(v_1, \ldots, v_n, \delta^-, \delta^+)$.

6 Experimental evaluation

This section reports the experimental evaluation of our approach on four datasets, *Adult*, *German credit* and *Census-income* from the UCI Repository of machine learning databases[5], and *Berkeley Admissions Data* from [10]. These are well-known real-life datasets typically used in discrimination-detection literature.

- Adult consists of 48,842 tuples and 10 attributes, where each tuple corresponds to an individual and it is described by personal attributes such as age, race, sex, relationship, education, and employment. Following the literature, in order to define the decision attribute we use the income levels, ≤50K (negative decision) or >50K (positive decision). We use four levels in the partial order for temporal priority: age, race, sex, and native country are defined in the first level; education, marital status, and relationship are defined in the second level; occupation and work class are defined in the third level, and the decision attribute (derived from income) is the last level.
- German credit consists of 1000 tuples with 21 attributes on bank account holders applying for credit. The decision attribute is based on repayment history, i.e., whether the customer is labeled with good or bad credit risk. Also for this dataset, the partial order for temporal priority has four orders. Personal attributes such as gender, age, and foreign worker are defined in the first level. Personal attributes such as employment status and job status are defined in the second level. Personal properties such as savings status and credit history are defined in the third level, and finally, the decision attribute is the last level.
- Census-income consists of 299,285 tuples and 40 attributes, where each tuple corresponds to an individual and it is described by demographic and employment attributes such as age, sex, relationship, education, and employment. Similar to Adult dataset, the decision attribute is the income levels and we define four levels in the partial order for temporal priority.

Building the SBCN just takes a handful of seconds in German credit and Adult, and few minutes in Census-income on a commodity laptop. The main characteristics of the extracted SBCN are reported in Table 1. As discussed in Introduction, we also use the dataset from the famous 1973 episode at University of California at Berkeley, in order to highlight the differences between correlation-based and causation-based methods.

- Berkeley Admissions Data consist of 4,486 tuples and three attributes, where each tuple corresponds to an individual and it is described by the gender of applicants and the department that they apply for it. For this dataset, the partial order for temporal priority has three orders. Gender is defined in the first level, department in the second level, and finally, the decision attribute in the last level. Table 2 is a three-way table that presents admissions data at the University of California, Berkeley, in 1973 according to the variables department (A, B, C, D, E), gender (male, female), and outcome (admitted, denied). The table is adapted from data in the text by Freedman et al. [10].

[5] http://archive.ics.uci.edu/ml.

Table 1 SBCN main characteristics

Dataset	$\|V\|$	$\|A\|$	*avgDeg*	*maxInDeg*	*maxOutDeg*
Adult	92	230	2.5	7	19
German credit	73	102	1.39	3	7
Census-income	386	1426	3.69	8	54

Table 2 Berkeley admission data

Male		Female		Department
Admitted	Denied	Admitted	Denied	
512	313	89	19	A
313	207	17	8	B
120	205	202	391	C
138	279	131	244	D
53	138	94	299	E
22	351	24	317	F

6.1 Community detection on the SBCN

Given that our SBCN is a directed graph with edge weight, as a first characterization we try to partition it using a random-walk-based community detection algorithm, called *Walktrap* and proposed in [32], whose unique parameter is the maximum number of steps in a random walk (we set it to 8), and which automatically identifies the right number of communities. The idea is that short random walks tend to stay in the same community (densely connected area of the graph). Using this algorithm over the reconstructed SBCN from Adult dataset, we obtain 5 communities: two larger ones and three smaller ones (reported in Table 3). Interestingly, the two larger communities seem built around the negative (C_1) and the positive (C_2) decisions.

Figure 2 in Sect. 4 shows the subgraph of the SBCN corresponding to C_2 (that we can call, the favoritism cluster): we note that such cluster also contains nodes such as sex_Male, age_old, relationship_Husband. The other large community C_1 can be considered the discrimination cluster: beside the negative decision it contains other nodes representing disadvantaged groups such as sex_Female, age_young, race_Black, marital_status_Never_married. This good separability of the SBCN in the two main clusters of discrimination and favoritism highlights the goodness of the causal structure captured by the SBCN.

6.2 Group discrimination and favoritism

We next focus on assessing the discrimination score $\mathrm{d}s^-$ we defined in Sect. 5.1, as well as the average number of steps that the random walks take to reach the negative and positive decisions, denote $as^-(v)$ and $as^+(v)$ respectively.

Tables 4, 5, and 6 report the top-5 and bottom-5 nodes w.r.t. the discrimination score $\mathrm{d}s^-$, for datasets Adult, German and Census-income, respectively. The first and most important observation is that our discrimination score provides a very clear signal, with some disadvantaged groups having very high discrimination score (equal to 1 or very close), and similarly clear signals of favoritism, with groups having $\mathrm{d}s^-(v) = 0$, or equivalently $\mathrm{d}s^+(v) = 1$. This is more clear in the Adult dataset, where the positive and negative decisions are artificially derived from the income attribute. In the German credit dataset, which is more realistic as the decision attribute is truly about credit, both discrimination and favoritism are less palpable. This is also due to the fact that German credit contains less proper causal relations, as reflected in the higher sparsity of the SBCN. A consequence of this sparsity is also that the random walks generally need more steps to reach one of the two decisions. In Census-income dataset, we observe favoritism with respect to married and asian_pacific individuals.

6.3 Genuine requirement

We next focus on genuine requirement (or explainable discrimination). Table 7 reports some examples of fraction of explainable discrimination (both positive and negative) on the Adult dataset. We can see how some fractions of discrimination against protected groups can be "explained" by intermediate nodes such as having a low education profile, or a simple job. In case these intermediate nodes are considered legally grounded, then one cannot easily file a discrimination claim.

Similarly, we can observe that the favoritism toward groups such as married men, is explainable, to a large extent, by higher education and good working position, such as managerial or executive roles.

6.4 Subgroup and individual discrimination

We next turn our attention to subgroup and individual discrimination discovery. Here the problem is to assign a score of discrimination not to a single node (a group), but to multiple nodes (representing the attributes of an individual or a subgroup of citizens). In Sect. 5.4, we have introduced based on the PageRank of the positive and negative decision, $\mathrm{ppr}(\delta^+)$ and $\mathrm{ppr}(\delta^-)$, respectively, personalized on

Table 3 Communities found in the SBCN extracted from the Adult dataset by *Walktrap* [32]

C_1
negative_dec, wc:Private, ed:Some_college, ed:Assoc_acdm, **ms:Never_married**, ms:Divorced, **ms:Widowed**, ms:Married_AF_spouse, oc:Sales, oc:Other_service, oc:Priv_house_serv, re:Own_child, re:Not_in_family, re:Wife, **re:Unmarried**, re:Other_relative, **ra:Black**, oc:Armed_Forces, oc:Handlers_cleaners, oc:Tech_support, oc:Transport_moving, ed:7th_8th, ed:10th, ed:12th, ms:Separated, ed:HS_grad,ed:11th, nc:Outlying_US_Guam_USVI_etc, nc:Haiti, **ag:young**, **sx:Female**, ra:Amer_Indian_Eskimo, nc:Trinadad_Tobago, nc:Jamaica, oc:Machine_op_inspct, ms:Married_spouse_absent, oc:Adm_clerical,
C_2
positive_dec, oc:Prof_specialty, wc:Self_emp_not_inc, ms:Married_civ_spouse, oc:Craft_repair,oc:Protective_serv, **re:Husband**, ed:Prof_school, wc:Self_emp_inc, **ag:old** , wc:Local_gov, **oc:Exec_managerial**, ed:Bachelors, ed:Assoc_voc, ed:Masters, wc:Never_worked, wc:State_gov, ed:Doctorate, **sx:Male**, nc:India, nc:Cuba
C_3
oc:Farming_fishing, wc:Without_pay, nc:Mexico, nc:Canada, nc:Italy, nc:Guatemala, nc:El_Salvador, ra:White, nc:Poland, ed:1st_4th, ed:9th,ed:Preschool, ed:5th_6th
C_4
nc:Iran, nc:Puerto_Rico, nc:Dominican_Republic, nc:Columbia, nc:Peru, nc:Nicaragua, ra:Other
C_5
nc:Philippines, nc:Cambodia, nc:China, nc:South, nc:Japan, nc:Taiwan, nc:Hong, nc:Laos, nc:Thailand, nc:Vietnam, ra:Asian_Pac_Islander

In the table, the attributes are shortened as in parenthesis: age (ag), education (ed), marital_status (ms), native_country (nc), occupation (oc), race(ra), relationship (re), sex (sx), workclass (wc)

Table 4 Top-5 and bottom-5 groups by discrimination score $ds^-(v)$ in Adult dataset

	$ds^-(v)$	$as^-(v)$	$as^+(v)$
relationship_Unmarried	1	1.164	–
marital_status_Never_married	0.996	1.21	2.14
age_Young	0.995	2.407	3.857
race_Black	0.994	2.46	4.4
sex_Female	0.98	2.60	3.76
relationship_Husband	0	–	2
marital_status_Married_civ_spouse	0	–	2.06
sex_Male	0	–	3.002
native_country_India	0.002	4.0	3.25
age_Old	0.018	2.062	2.14

the nodes of interest. Figure 3 presents a scatter plot of $\text{ppr}(\delta^+)$ versus $\text{ppr}(\delta^-)$ for each individual in the German credit dataset. We can observe the perfect separation between individuals corresponding to a high personalized PageRank with respect to the positive decision, and those associated with a high personalized PageRank relative to the negative decision.

Such good separation is also reflected in the *generalized discrimination score* (Definition 4) that we obtain by combining $\text{ppr}(\delta^+)$ versus $\text{ppr}(\delta^-)$.

In Fig. 4, we report the distribution of the generalized discrimination score gds^- for the population of the German credit dataset: we can make a note of the clear separation between the two subgroups of the population.

Table 5 Top-5 and bottom-4 groups by discrimination score $\mathrm{d}s^-(v)$ in German credit. We report only the bottom-4, because there are only four nodes in which $\mathrm{d}s^+(v) > \mathrm{d}s^-(v)$

	$\mathrm{d}s^-(v)$	$as^-(v)$	$as^+(v)$
residence_since_le_1d6	1	6.0	–
residence_since_gt_2d8	1	2.23	–
residence_since_from_1d6_le_2d2	1	6.0	–
age_gt_52d6	0.86	3.68	4.0
personal_status_male_single	0.791	5.15	5.0
job_unskilled_resident	0	–	2.39
personal_status_male_mar_or_wid	0.12	8.0	4.4
age_le_30d2	0.186	7.0	3.34
personal_status_female_	0.294	6.48	4.4
div_or_sep_or_mar			

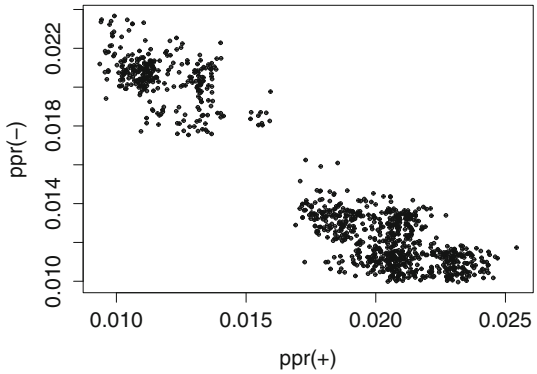

Fig. 3 Scatter plot of $\mathrm{ppr}(\delta^+)$ versus $\mathrm{ppr}(\delta^-)$ for each individual in the German credit dataset

In the Adult dataset (Fig. 5), we do not observe the same neat separation in two subgroups as in the German credit dataset, also due to the much larger number of points. Nevertheless, as expected, $\mathrm{ppr}(\delta^+)$ and $\mathrm{ppr}(\delta^-)$ still exhibit anti-correlation. In Fig. 5, we also use colors to show two different groups: red dots are for age_Young and blue dots are for age_Old individuals. As expected, we can see that the red dots are distributed more in the area of higher $\mathrm{ppr}(\delta^-)$.

The plots in Fig. 6 have a threshold $t \in [0, 1]$ on the X-axis, and the fraction of tuples having $gds^-() \geq t$ on the Y-axis, and they show this for different subgroups. The first plot, from the Adult dataset, shows the group female, young, and young female. As we can see the individuals that are both young and female have a higher generalized discrimination score.

Table 6 Top-5 and bottom-5 groups by discrimination score $\mathrm{d}s^-(v)$ in Census-income dataset

	$\mathrm{d}s^-(v)$	$as^-(v)$	$as^+(v)$
MIGSAME_Not_in_universe_under_1_year_old	0.71	4.09	8.82
WKSWORK_94_5_inf	0.625	3.0	6.76
AWKSTAT_Not_in_labor_force	0.59	2.0	6.16
VETYN_0_5_20_5	0.58	1.01	5.17
MARSUPWT_3188_455_4277_98	0.55	5.0	9.25
AHGA_Doctorate_degreePhD_EdD	0	–	3.07
AMARITL_Married_A_F_spouse_present	0	–	4.49
AMJOCC_Sales	0	–	2.0
ARACE_Asian_or_Pacific_Islander	0	–	6.47
VETYN_20_5_32_5	0	–	5.89

Table 7 Fraction of explainable discrimination for some exemplar pair of nodes in the Adult dataset

Source node	Intermediate	$\mathrm{fed}^-(v)$
race_Amer_Indian_Eskimo	education_HS_grad	0.481
sex_Female	occupation_Other_service	0.310
age_Young	occupation_Other_service	0.193
relationship_Unmarried	education_HS_grad	0.107
race_Black	education_11th	0.083
Source node	**Intermediate**	$\mathrm{fed}^+(v)$
relationship_Husband	occupation_Exec_managerial	0.806
sex_Male	occupation_Exec_managerial	0.587
native_country_Iran	education_Bachelors	0.480
native_country_India	education_Prof_school	0.415
age_Old	occupation_Exec_managerial	0.39

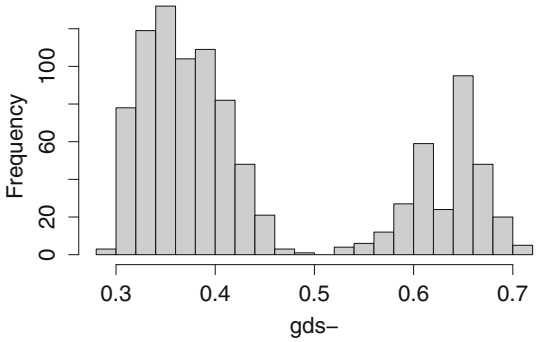

Fig. 4 Individual discrimination: histogram representing the distribution of the values of the generalized discrimination score gds^- for the population of the German credit dataset

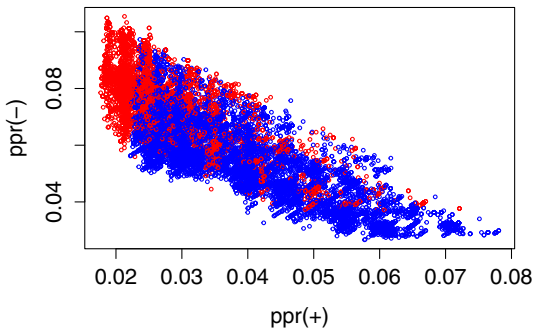

Fig. 5 Individual discrimination: scatter plot of $\text{ppr}(\delta^+)$ versus $\text{ppr}(\delta^-)$ for each individual in the Adult dataset. Red dots are for age_Young and blue dots are for age_Old

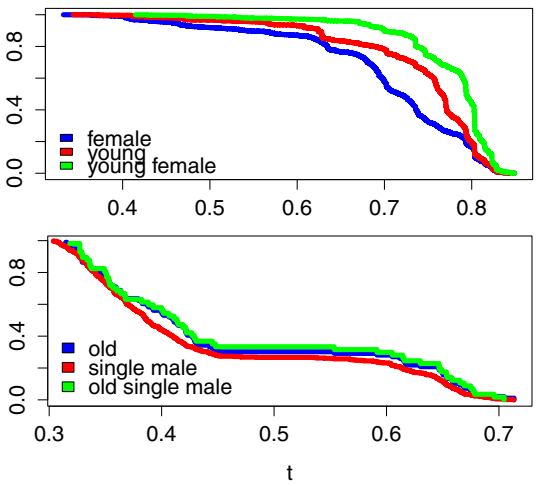

Fig. 6 Subgroup discrimination: plots reporting a threshold $t \in [0, 1]$ on the X-axis and the fraction of tuples having $gds^-() \geq t$ on the Y-axis. The top plot is from Adult, while the bottom is from German credit

Similarly, the second plot shows the groups old, single male, and old single male from the German credit dataset. Here we can observe much lower rates of discrimination with only 1/5 of the corresponding populations having $gds^-() \geq 0.5$, while in the previous plot it was more than 85%.

Table 8 Top-5 and bottom-5 groups by discrimination score $\text{d}s^-(v)$ in Adult dataset, where SBCN is learned by Algorithm 2

	$\text{d}s^-(v)$	$as^-(v)$	$as^+(v)$
relationship_Unmarried	0.993	1.69	7.2
marital_status_Never_married	0.972	2.12	5.90
race_Black	0.968	2.89	6.57
sex_Female	0.941	3.01	4.83
age_Young	0.937	2.01	6.12
sex_Male	0.034	3.56	1.14
native_country_India	0.07	5.2	3.98
relationship_Husband	0.089	5.0	2.05
marital_status_Married_civ_spouse	0.11	2.885	2.06
age_Old	0.16	2.83	1.73

6.5 Learning SBCN without manual temporal ordering

Table 8 reports the top-5 and bottom-5 nodes w.r.t. the discrimination score $\text{d}s^-$, for dataset Adult, using the Algorithm 2 presented in Sect. 4.5. We can observe that although the values of $\text{d}s^-(v)$, $as^-(v)$ and $as^+(v)$ are slightly different from the ones in Table 4, the top-5 nodes and bottom-5 nodes w.r.t. the discrimination score $\text{d}s^-$ are the same.

Table 9 reports some examples of fraction of explainable discrimination (both positive and negative) on the Adult dataset, where here SBCN generated by the Algorithm 2, presented in Sect. 4.5. We can observe that although the values of $\text{fed}^-(v)$ are slightly different from the ones in Table 7, the same kind of relationship is observed between every pair of nodes. The above results highlight the fact that we can learn SBCN without accessing manual temporal ordering while the values of discrimination measures are very similar to the case where temporal ordering is given in advance.

Finally, we also computed the value of individual discrimination score for Adult dataset, where SBCN has been learned by Algorithm 2. Figure 7 presents a scatter plot of $\text{ppr}(\delta^+)$ versus $\text{ppr}(\delta^-)$ for each individual in the Adultcredit dataset. Similar to Fig. 5, here we also observe the anti-correlation between $\text{ppr}(\delta^+)$ and $\text{ppr}(\delta^-)$. Comparing the two figures shows that the distribution of the discrimination score using the both approaches is very similar.

6.6 Comparison with prior methods

In this section, we discuss examples in which our causation-based method draws different conclusions from the correlation-based methods presented in [30,31] and [39] using the same datasets and the same protected groups [6].

The first example involves the foreign_worker group from German Credit dataset, whose contingency table is

[6] We could not compare with [26] due to repeatability issues.

Table 9 Fraction of explainable discrimination for some exemplar pair of nodes in the Adult dataset. , where SBCN is learned by Algorithm 2

Source node	Intermediate	$\text{fed}^-(v)$
race_Amer_Indian_Eskimo	education_HS_grad	0.368
sex_Female	occupation_Other_service	0.230
age_Young	occupation_Other_service	0.123
relationship_Unmarried	education_HS_grad	0.107
race_Black	education_11th	0.05
Source node	**Intermediate**	$\text{fed}^+(v)$
relationship_Husband	occupation_Exec_managerial	0.783
sex_Male	occupation_Exec_managerial	0.773
native_country_Iran	education_Bachelors	0.409
native_country_India	education_Prof_school	0.37
age_Old	occupation_Exec_managerial	0.336

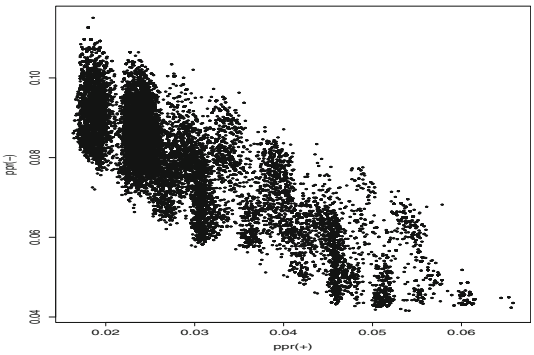

Fig. 7 Individual discrimination: scatter plot of $\text{ppr}(\delta^+)$ versus $\text{ppr}(\delta^-)$ for each individual in the Adult dataset, where SBCN is learned by Algorithm 2

	decision		
	-	+	
foreign_worker=yes	298	667	968
foreign_worker=no	2	30	32
	300	700	1000

$p_1 = 298/968 = 0.307$
$p_2 = 2/32 = 0.0625$
$RD = p_1 - p_2 = 0.244$

Fig. 8 Contingency table for foreign_worker in the German credit dataset

	decision		
	-	+	
race=black	4119	566	4685
race≠black	33036	11121	44157
	37155	11687	48842

$p_1 = 4119/4685 = 0.879$
$p_2 = 33036/44157 = 0.748$
$RD = p_1 - p_2 = 0.13$

Fig. 9 Contingency table for race_black in the Adult dataset

reported in Fig. 8. Following the approaches of [30,31], and [39], the foreign_worker group results strongly discriminated. In fact, Fig. 8 shows an RD value (*risk difference*) of 0.244 which is considered a strong signal: in fact $RD > 0$ is already considered discrimination [39].

However, we can observe that the foreign_worker group is not very significant, as it contains 963 tuples out of 1000 total. In fact, our causal approach does not detect any discrimination with respect to foreign_worker which appears as a disconnected node in the SBCN.

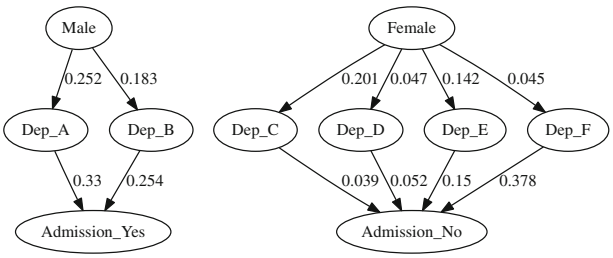

Fig. 10 The SBCN constructed from Berkeley Admission Data dataset

The second example is in the opposite direction. Consider the race_black group from Adult dataset whose contingency table is shown in Fig. 9. Our causality-based approach detects a very strong signal of discrimination ($\text{ds}^-() = 0.994$), while the approaches of [30,31], and [39] fail to discover discrimination against black minority when the value of minimum support threshold used for extracting classification rules is more than 10%. On the other hand, when such minimum support threshold is kept lower, the number of extracted rules might be overwhelming.

Finally, we turn our attention to the famous example of false-positive discrimination case happened at Berkeley in 1973 that we discussed in Sect. 1. Figure 10 presents the SBCN extracted by our approach from Berkeley Admission Data. Interestingly, we observe that there is no direct edge between node sex_Female and Admission_No. And sex_Female is connected to node Admission_No through nodes of Dep_C, Dep_D, Dep_E, and Dep_F, which are exactly the departments that have lower admission rate. By running our random-walk-based methods over SBCN we obtain the value of 1 for the score of explainable discrimination confirming that apparent discrimination in this dataset is due the fact that women tended to apply to departments with lower rates of admission.

	decision		
	-	+	
gender=female	1278	557	1835
gender=male	1493	1158	2651
	2771	1715	4486

$p_1 = 1278/1835 = 0.696$
$p_2 = 1493/2651 = 0.563$
$RD = p_1 - p_2 = 0.133$

Fig. 11 Contingency table for female in the Berkeley Admission Data dataset

Similarly, we observe that there is no direct edge between node sex_Male and Admission_Yes. And sex_Male is connected to node Admission_Yes through nodes of Dep_A, and Dep_B, which are exactly the departments that have higher admission rate. By running our random-walk-based methods over SBCN we obtain the value of 1 for the score of explainable discrimination confirming that apparent favoritism toward men is due to the fact that men tended to apply to departments with higher rates of admission.

However, following the approaches of [30,31], and [39], the contingency table shown in Fig. 11 can be extracted from Berkeley Admission Data. As shown in Fig. 11, the value of *RD* suggests a signal of discrimination versus women.

This highlights once more the pitfalls of correlation-based approaches to discrimination detection and the need for a principled causal approach, as the one we propose in this paper.

7 Conclusions and future work

Discrimination discovery from databases is a fundamental task in understanding past and current trends of discrimination, in judicial dispute resolution in legal trials, in the validation of micro-data before they are publicly released. While discrimination is a causal phenomenon, and any discrimination claim requires proving a causal relationship, the bulk of the literature on data mining methods for discrimination detection is based on correlation reasoning.

In this paper, we propose a new discrimination discovery approach that is able to deal with different types of discrimination in a single unifying framework. We define a method to extract a graph representing the causal structures found in the database, and then we propose several random-walk-based methods over the causal structures, addressing a range of different discrimination problems. Our experimental assessment confirmed the great flexibility of our proposal in tackling different aspects of the discrimination detection task, and doing so with very clean signals, clearly separating discrimination cases.

To the best of our knowledge, this is the first proposal of discrimination detection in databases grounded in probabilistic causal theory: as such there are several research paths that are worth further investigation.

In various fields such as epidemiology, social sciences, psychology, and statistics, it is a common practice to perform the so-called *observational studies* to draw inferences of causal effects such as for instance the possible effect (e.g., survival) of a treatment (e.g., a drug) on subjects. In this case, we rephrase the task of causal inference in terms of counterfactual evidence, aims at assessing potential causal behaviors between a factor (e.g., being female) and an outcome (e.g., income) given a set of confounding factors (the other covariates in the model) which provides the context for the causal relation. However, it is not easy to conduct such intervention studies in the social settings in which discrimination occurs naturally. Thus introducing counterfactual causality into our framework is not straightforward and we leave further investigations regarding this for future work.

In our framework, we assume that the nodes of interest are known. This is a reasonable assumption for the group discrimination as the protected (sensitive) groups are usually predefined by legislation. However, this might not be the case for the subgroup discrimination. Automatically exploring the whole space of possible discrimination patterns is a research challenge that we aim to tackle in the future work. One possible solution is to start from individual discrimination scores: the attribute values of the individual records with high discrimination score are potential candidates for measuring subgroup discrimination.

Acknowledgements The research leading to these results has received funding from the European Unions Horizon 2020 Innovation Action Program under grant agreement No 653449 TYPES project, the Catalonia Trade and Investment Agency (Agncia per la competitivitat de l'empresa, ACCI) and CMU grant No 15-00314-SUB-000.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

8 Appendix: supplementary materials

Theories of causality are old and central to many fields: philosophy, logic, biology, inductive and statistical inference, and, most recently, invading computer and data sciences—distinct versions of theories enjoying varying degrees of popularity within various fields. The differences among these theories are subtle and are often not readily appreciated by the practitioners, leading to acrimonious disputes . In order to situate our work properly within these naturally confusing contexts, we provide here a brief review of some of the influential theories of causality. In particular, we start with the foundational work of Hume (e.g., regularity) and Lewis (e.g., counterfactual), which represents the origin of the more recent developments in machine learning, most prominently by Judea Pearl—currently enjoying a viral popularity among the engineers and computer scientists reenergizing the study of causal inference and related algorithms. Of related interest to computer scientists is the question of complexity of inference: for instance, Pearl's approach usually leads to computational intractability. However, the theories of prob-

abilistic causation due to Patrick Suppes, despite various known limitations and pitfalls, are expressible in probabilistic computational tree logic with efficient model checkers that allow devising efficient learning algorithms, as discussed and proved in this paper to be effective. Furthermore, since there are various paradoxes (e.g., Goodman and Simpsons) plaguing the theories, the quest for a single universal theory of causality will likely remain elusive. Nonetheless, with the advent of data science generating massive amount of measurements from natural experiments, there is now a possibility of a consensus theory that allows reasonably efficient algorithms and demonstrates empirically its power to minimize false and missing discoveries. Based on a detailed empirical analysis of Kleinberg [33] and her coauthors, we suspect that such a theory will build upon the work of Suppes, focusing on prima facie causality, modulo various regularization techniques to reduce spurious causalities. Beyond what we describe below, a more detailed presentation of the extensive (and steadily growing) literature of causality theories is beyond the scope of this work. We refer the interested reader to and citations within.

8.1 Hume's regularity theory

The modern study of causation begins with the Scottish philosopher David Hume (1711–1776). According to Hume, a theory of causation could be defined axiomatically, using the following ingredients: *temporal priority*, implying that causes are invariably followed by their effects [14], augmented by various constraints, such as contiguity and constant conjunction[7]. Theories of this kind, that try to analyze causation in terms of invariable patterns of succession, have been referred to as *regularity theories* of causation.

Nonetheless, as described earlier, the notion of causation has spawned far too many variants and has been a source of acerbic debates. All these theories present well-known limitations and confusion, but have led to a small number of modern versions of commonly accepted (at least among the philosophers) frameworks. Thus, in the next sections we will provide a review of the main state-of-the-art theories of causation that have attempted to formulate a *sound and complete* theory of causation.

8.2 Lewis's counterfactuals

The most complete known counterfactual theory of causation is due to David Lewis [23] and exploits a possible world semantics to state truth conditions for counterfactuals in terms of *similarity* among possible worlds: one possible world is closer to actuality than another, if it is more similar to the actual world.

Following this idea, Lewis defined two important constraints on the resulting similarity relation: (i) similarity induces an ordering of worlds in terms of closeness to the actual world and (ii) the actual world is the closest possible world to actuality. Then, the evaluation of the counterfactual *"if c were the case, e would be the case"* is true just in case it is closer to actuality to make the first term true along with the second—as opposed to making it true without. Therefore, in terms of counterfactuals Lewis defines the following notion of causality: given c and e, whether e occurs or not depends on whether c occurs or not, and e causally depends on c if and only if, if c were not to occur e would not occur. Thus, the idea of cause is conceptually linked to the idea of *something that makes a difference,* and this concept in turn is naturally described in terms of counterfactuals. Lewis also characterized causation in terms of temporal direction by stating that the direction of causation is the direction of causal dependence and that, typically, events causally depend on earlier events but not on later ones.

8.3 Manipulability theories of causation

We now briefly discuss the notion of *intervention* as propounded by Judea Pearl [28]; in general, interventionist versions of manipulability theories can be seen as counterfactual theories. For a detailed discussion on this and manipulability theories of causation, refer to [44].

Pearl characterizes his notion of intervention in terms of a primitive notion of causal mechanism. According to him, the world is organized in the form of stable mechanisms (i.e., physical laws) which are autonomous. Therefore, he states that we can change one of them, without changing all the others. Thus an intervention may imply that: *if we manipulate c and nothing happens, then c cannot be cause of e, but if a manipulation of c leads to a change in e, then we know that c is a cause of e, although there might be other causes as well.*

In other words, when among many events a causal relationship between some e and its parents (i.e., directed causes, say $c_1, \ldots, c_n$) is present, the interventions will disrupt completely the relationships between e and $c_1, \ldots, c_n$ such that the value of e is determined by the intervention only. Thus, intervention is a surgical operation in the sense that no other causal relationship in the system is changed by it. Hence, Pearl's assumption is that the other variables that change in values under this intervention will do so *only because they are effects of* e. Pearl's theory has been very influential among the computational causality theorists and has generated state-of-the-art algorithms for causal network inference [19].

[7] Some of these notions have been modernized with the introduction of the machinery from statistical inference, logic and model theory; but they have stayed more or less true to Hume's program.

8.4 Issues of interventionist causation

Next, we point the reader to some problems that can arise in practice, when applying intervention in the context of causal inference. For a deeper discussion, we refer to [44].

Circularity. An intervention on an event e leaves intact *all* the other *causal* mechanisms besides the ones involving c as a cause. Because of this, Pearl's intervention could lead to circularity problems, i.e., it seems that the causal mechanisms need to be known in advance in order to validate (or refute) them.

Possible and impossible interventions. Causal claims are described in terms of counterfactuals of what would happen when applying intervention to a given causal relationship. Moreover, the notion of intervention is connected with the possibility of a *human action* to intervene in a system. In some contexts, however, it may be *impossible* to evaluate what would happen by performing a *surgical* intervention. Thus, it should be clear that, regardless of the possible criticisms to Pearl's framework, there are situations where, at least relative to the current human capabilities, it is very complicated, if not impossible, to perform intervention.

8.5 Suppes' prima facie cause

Patrick Suppes proposed the notion of a *prima facie cause* that represents the core of *probabilistic causation* and also provides the algorithmic foundations of our analysis.

Definition 5 (*Probabilistic causation,* [41]) For any two events c and e, occurring, respectively, at times t_c and t_e, under the mild assumptions that $0 < P(c), P(e) < 1$, the event c is called a *prima facie cause* of e if it occurs *before* and *raises the probability* of e, i.e.,

$$t_c < t_e \quad \text{and} \quad P(e \mid c) > P(e \mid \ c)\,. \tag{1}$$

From now on, the first condition will be referred to as *temporal priority*, whereas the second as *probability raising*. This notion of causation has some advantages over the simplest version of a regularity theory of causation, e.g., it deals with various issues usually associated with imperfect regularities.

Unfortunately, however, prima facie causality is still *not sufficient* in capturing a causation relationship *in its full generality*. For instance, the problem of spurious regularities still remains, additionally requiring that prima facie causes be refined further into two classes: *genuine* and *spurious*. In the latter case, as discussed, we may observe a prima facie cause to be so labeled only because of spurious correlations. Also, as discussed extensively in the literature (see [13]), one may encounter certain situations, in which Suppes' characterization fails to provide a *necessary* condition. In the next paragraph, we will briefly discuss an attempt to make Suppes' conditions sufficient for any causal claims.

8.6 Reichenbach's screening-off

In [35], Reichenbach discussed the notion of *screening-off* to describe a particular type of probabilistic relationship. Consider, e.g., events a, c and e, and assume to observe $P(e \mid a \wedge c) = P(e \mid c)$, then we say that c is *screening a off* from e. When $P(e \wedge c) > 0$, this is equivalent to stating that $P(a \wedge e \mid c) = P(a \mid c) \cdot P(e \mid c)$ – i.e., a and e happen to be probabilistically independent, when conditioned upon c. The preceding situation could occur in two cases.

In the first case, c is a genuine cause of e while a is a genuine cause of c as well, and the correlations between a and e are only just manifestations of these known causal connections. For example, unprotected sex (a) appears to cause AIDS (e) *only* because of sexually transmitted HIV infection (c). Then, we would expect that among those who have already been infected with HIV, the probability of contacting AIDS would be the same regardless of whether one is engaged in unprotected sex or not. Here c is a *proximate* cause of e and an *intermediate* cause leading from a to e, i.e., an instance of *causal transitivity*. In the second case, c is a common cause of both a and e, that is exactly a situation of spurious correlation.

Building upon this idea, Reichenbach formulated the so-called *common cause principle* (CCP) to detect situations leading to "screening-off," and so identify when a spurious correlation can be explained in terms of a common cause. Unfortunately, there are situations where such a principle leads to computationally intractable criteria. Since these issues are not germane to the context of this work, we will not discuss them further, other than pointing the interested readers to appropriate literature [13]. Nevertheless, the idea of screening-off has significantly influenced some of the most widely used recent theories of causation and has become central to the topic.

8.7 Issues of probabilistic causation

Now, we describe some thorny issues in the theory of probabilistic causation. We also briefly point out some unresolved problems, proposed plans of attack, and ensuing criticisms. For a deeper discussion, see [13].

Pearl's criticism. In [28], Pearl argues that the notion that causes "raise the probabilities" of their effects *cannot be expressed in the language of probability theory*. In particular, according to Pearl, the inequality $P(e \mid c) > P(e \mid \ \bar{c})$ fails to capture the intuition behind probability raising, which

must be *manipulative* or *counterfactual*. Because of this limit, Pearl argues that it is not possible to rigorously describe the intuitions behind the probability raising theory and, for this reason, the only way to properly assess a causal claim is exclusively by *intervention*.

Note that, as discussed before, it would be impossible to execute randomized experiments involving discrimination, thus making Pearl's criticism largely irrelevant in our context.

References

1. Bickel, P.J., Hammel, E.A., O'Connell, J.W.: Sex bias in graduate admissions: data from Berkeley. Science **187**(4175), 398–404 (1975)
2. Brenner, E., Sontag, D.: Sparsityboost: A new scoring function for learning bayesian network structure. arXiv preprint arXiv:1309.6820 (2013)
3. Buntine, W.: Theory refinement on bayesian networks. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, pp. 52–60. Morgan Kaufmann Publishers Inc. (1991)
4. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
5. Custers, B., Calders, T., Schermer, B., Zarsky, T.: Discrimination and Privacy in the Information Society: Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 3. Springer, Berlin (2012)
6. Dabady, M., Blank, R.M., Citro, C.F., et al.: Measuring Racial Discrimination. National Academies Press, Washington (2004)
7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226. ACM (2012)
8. Ellis, E., Watson, P.: EU Anti-discrimination Law. Oxford University Press, Oxford (2012)
9. Foster, S.R.: Causation in antidiscrimination law: beyond intent versus impact. Houst. Law Rev. **41**, 1469 (2004)
10. Freedman, D., Pisani, R., Purves, R.: Statistics, 3rd edn. Norton, New york (1998)
11. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. Knowl. Data Eng. **25**(7), 1445–1459 (2013)
12. Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., Giannotti, F.: Discrimination-and privacy-aware patterns. Data Min. Knowl. Discov. 29(6), 1733–1782 (2015)
13. Hitchcock, C.: Probabilistic causation. In Zalta, E.N. (ed.), The Stanford Encyclopedia of Philosophy, winter 2012 edn. (2012)
14. Hume, D., Hendel, C.W.: An inquiry concerning human understanding, vol. 49, Bobbs-Merrill Indianapolis (1955)
15. Illari, P.M., Russo, F., Williamson, J.: Causality in the Sciences. Oxford University Press, Oxford (2011)
16. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the 12th International Conference on World Wide Web, pp. 271–279. ACM (2003)
17. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
18. Kleinberg, S.: Causality, Probability, and Time. Cambridge University Press, Cambridge (2012)
19. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press, Cambridge (2009)
20. Legislation, A.: Equal opportunity act–Victoria state, (b) anti-discrimination act–Queensland state (2008)
21. E. U. Legislation. European union legislation, (a) race equality directive, 2000/43/ec, 2000; (b) employment equality directive, 2000/78/ec, 2000; (c) equal treatment of persons, European Parliament legislative resolution (2009)
22. U. F. Legislation. Us federal legislation, (a) equal credit opportunity act, (b) fair housing act, (c) intentional employment discrimination, (d) equal pay act, (e) pregnancy discrimination act, (f) civil right act. (2010)
23. Lewis, D.: Causation. J. Philos **70**, 56–567 (1973)
24. Loohuis, L.O., Caravagna, G., Graudenzi, A., Ramazzotti, D., Mauri, G., Antoniotti, M., Mishra, B.: Inferring tree causal models of cancer progression with probability raising. PLoS ONE **10**(9), e108358 (2014)
25. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 502–510. ACM (2011)
26. Mancuhan, K., Clifton, C.: Combating discrimination using bayesian networks. Artif. Intell. Law **22**(2), 211–238 (2014)
27. Menzies, P.: Counterfactual theories of causation. In E. N. Zalta, editor, The Stanford Encyclopedia of Philosophy. The Stanford Encyclopedia of Philosophy, spring 2014 edition, 2014
28. Pearl, J.: Causality: models, reasoning, and inference. Econom. Theory **19**, 675–685 (2003)
29. Pedreschi, D., Ruggieri, S., Turini, F.: Integrating induction and deduction for finding evidence of discrimination. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 157–166. ACM (2009)
30. Pedreschi, D., Ruggieri, S., Turini,F.: Measuring discrimination in socially-sensitive decision records. In: SDM, pp. 581–592. SIAM (2009)
31. Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568. ACM (2008)
32. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005, pp. 284–293. Springer (2005)
33. Ramazzotti, D.: A model of selective advantage for the efficient inference of cancer clonal evolution. arXiv preprint arXiv:1602.07614 (2016)
34. Ramazzotti, D., Caravagna, G., Loohuis, L.O., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., Mishra, B.: Capri: efficient inference of cancer progression models from cross-sectional data. Bioinformatics **31**, 3016–3026 (2015)
35. Reichenbach, H., Reichenbach, M.: The Direction of Time, vol. 65. University of California Press, Berkeley (1991)
36. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. Knowl. Eng. Rev. **29**(05), 582–638 (2014)
37. Rorive, I.: Proving discrimination cases: the role of situation testing, Centre for Equal Rights and MPG (2009)
38. Ruggieri, S., Hajian, S., Kamiran, F., Zhang, X.: Anti-discrimination analysis using privacy attack strategies. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp 694–710. Springer (2014)
39. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Trans. Knowl. Discov. Data (TKDD) **4**(2), 9 (2010)
40. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
41. Suppes, P.: A Probabilistic Theory of Causality. North-Holland Publishing Company, Amsterdam (1970)

42. van der Lee, R., Ellemers, N.: Gender contributes to personal research funding success in the Netherlands. Proc. Natl. Acad. Sci. **112**(40), 12349–12353 (2015)
43. Volker, B., Steenbeek, W.: No evidence that gender contributes to personal research funding success in the Netherlands: a reaction to van der Lee and Ellemers. In: Proceedings of the National Academy of Sciences, pp. 201519046 (2015)
44. Woodward, J.: Causation and manipulability. In Zalta, E. N. (ed.) The Stanford Encyclopedia of Philosophy, winter 2013 edition (2013)
45. Zliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Data Mining (ICDM), 2011 IEEE 11th International Conference on, pp. 992–1001. IEEE (2011)