Survey of Fair Clustering

Hanna Kondratiuk[1]

[1] Department of Computer Science – University of Bonn
Bonn, Germany

Abstract. *Fair clustering is usually defined as clustering with each subgroup having approximately same representation in every cluster. This concept is important as given the blind application of the algorithm the biases of the data can be exposed. However, the biggest challenge of the topic is the ambiguity of the fair clustering definition and the notations the definition is based on. In the session "Fair Clustering" the reviewed papers [Chierichetti et al. 2017], [Bera et al. 2019], [Huang et al. 2019] are featuring different algorithms providing solutions to fair variants of classical clustering problems, such as k-means and k-median problems. This survey paper focuses on the major ideas of the session "Fair Clustering" from the NeurIPS 2019 conference. We present main contributions of each paper and discuss the difficulties of defining the notion of fairness and fair clustering problem in general. Moreover, we conduct a systematic literature review of [Chierichetti et al. 2017] in order to provide a solid foundation on the prior use of fair clustering. Then we elaborate on the comparison of the three papers and discuss the ethical moment of the notion of fair clustering as the conclusion of our work.*

1. Introduction

The notion of fairness is an emerging concept in the field of the machine learning, that was extensively studied over the past years. The reason for that is that given the blind application of the learning algorithms, the biases of the data are exposed [Chierichetti et al. 2017]. For instance, natural language processing, a critical ingredient of common AI systems like Amazon's Alexa and Apple's Siri, among others, has been found to show gender biases [Bolukbasi et al. 2016].

In order to address the concerns with bias and discrimination in decision systems that rely on statistical inference and learning, the general notion of *fairness* was introduced [Zemel et al. 2013]. In the problem of the clustering, the *fairness* was revisited by [Chierichetti et al. 2017], featuring usage of the introduced notions of *balance* and *fairlets*. Two papers, namely [Bera et al. 2019] and [Huang et al. 2019] are based on the introduced notion of the *fair clustering*, expanding the applicability of the method and proposing their own approaches of solving the fair clustering problem.

However, as stated in [Kleinberg et al. 2017] and [Corbett-Davies et al. 2017] some fairness conditions cannot be simultaneously achieved, which makes it extremely challenging to formally define the concept of *fairness*. Indeed, in the course of the survey we notice that the problem of fair clustering, as well as notions of *fairness* and *balance*, can be defined differently. The problem of fair clustering outside of the context is ill-defined, as it only requires for each subgroup to have approximately same representation in every cluster. But how can the approximately same representation be quantified? One of the ways to address this question is using the *disparate impact* doctrine.

The disparate impact doctrine does not allow to use a protected attribute (such as gender) in decision-making explicitly. Moreover, the decision should not be disproportionately different for points from different protected classes. The decision made by the algorithm, that does not explicitly use the protected attribute could still be disproportionately different for points from different protected classes, because the used unprotected features (such as height) can be closely correlated to the protected features (such as gender).

On the other hand, fair clustering notation can be based on different doctrines. For example, it can be based on the disparate impact doctrine or on doctrine that violates disparate impact. Disparate impact is a law in United States, however it allows for violation if this violation is justified. In this case, the challenge is finding the trade-off between following the doctrine and minimizing the clustering objective. For example, if the doctrine is violated such that the outcome of a fair clustering is widely different from the unfair version, the points would be no longer assigned to its nearest cluster center. Then *cost of fairness*, which is the ratio of the objective values of the fair clustering over the unfair clustering, would be high.

Consequently, one might ask: how can we define the problem of fair clustering the way, that allows for wide range of clustering objectives, would not be controversial to the disparate impact doctrine and still minimizes the cost of fairness? Moreover, how a point belonging to multiple protected groups should be treated - should it affect the final decision of the fair algorithm, if the point belongs to not only one, but to multiple protected attributes? Attempting to address those questions, we conduct a comparison of the papers [Chierichetti et al. 2017], [Bera et al. 2019] and [Huang et al. 2019] with each paper proposing the own connotation of *fairness* definition and consequently the definition of *fair clustering*.

The paper is organised as follows. Firstly, we conduct the systematic literature review of [Chierichetti et al. 2017] and draw conclusions from it. Secondly, we briefly present each paper of the session with the main focus on the derived results, introducing the necessary definitions of each paper on the way. Then, we provide the comparison of each paper with respect to the such criteria as introduced notions, efficiency, running time, applicable clustering algorithms and fundamental doctrines. Moreover, we discuss advantages and disadvantages of each approach. As the conclusion, we summarise our thoughts and discuss the ethical perspective of the *fair clustering*.

2. Related Work

Using the backward snowball method, we are taking a look at the references of the paper of [Chierichetti et al. 2017], by using the inclusion criteria of the words “fair”, “fairness” or “discrimination” in the context of the machine learning. By doing so we are attempting to find the relevant information for the study, the origin of the notion fairness as well as to dive into the history of the fair clustering or fair machine learning algorithms in general. The idea of fair classifiers dates back to such early works as [Dwork et al. 2012], [Zemel et al. 2013]. None of the papers are using the idea of fair clustering, apart from [Zemel et al. 2013] in context of prototypes acting like clusters.

The paper of [Zemel et al. 2013] is referenced in the paper [Edwards and Storkey 2016] in context of “the authors learn a representation of the

data that is a probability distribution over clusters — a form of 'fair clustering' — where learning the cluster of a data-point tells one nothing about the sensitive variable S" and the notion of fairness was defined, however, none of the papers had concentrated directly on the fair clustering approach. Therefore the papers of [Zemel et al. 2013] and [Edwards and Storkey 2016] were the closest to the idea of fair clustering approach.

According to the systematic literature review, the goal of the paper [Chierichetti et al. 2017] is finding the previous artifacts at the area and not to replicate them. In the context of the unsupervised learning, the fairness concept was claimed to be the first applied to the clustering by [Chierichetti et al. 2017]. In the paper, it was indeed mentioned that "in this work we initiate the study of fair clustering algorithms" and the systematic literature review implies the same, as none of the papers found by search in Google Scholar or using the snowball method implied otherwise.

3. Approaches to Fair Clustering

All of the session's papers are applicable only to the task of clustering, that is finding structure within data based on similarity or distance in an unsupervised setting. The following subsections discuss the works of [Chierichetti et al. 2017], [Bera et al. 2019] and [Huang et al. 2019] respectively.

3.1. Fair Algorithms Through Fairlets

The paper [Chierichetti et al. 2017] initiates the research direction in the field of clustering. Moreover, it introduces the concept of *fairlets* - minimal sets that satisfy fair representation while approximately preserving the clustering objective.

This paper is based on the doctrine called disparate impact. In essence, the doctrine's idea for the resulting decision is not to be made disproportionately different for every class, even if the protected attribute (such as gender) was not directly used in the decision making process. The reason why it is not sufficient enough to do the blind clustering is that the data can be not only biased, but also that unprotected features can be closely correlated with the protected ones. For instance, the example used in [Chierichetti et al. 2017] mentions height as an unprotected feature and gender as a protected one. In this example, people with greater height are more likely to be accepted to the leading position than people with lower height, despite the algorithm used being gender-blind.

The main contributions of [Chierichetti et al. 2017] are the novel formulation of the fair clustering problem using *fairlets* and the introduction of *balance* and *cost* of fair clustering. Here *balance* encapsulates a specific notion of fairness, where a clustering with a monochromatic cluster (i.e., fully unbalanced) is considered unfair. The *cost* of fair clustering is the measure of the loss in the quality of the clustering when all of the protected classes are required to have approximately the same representations in the clusters returned. By using the *fairlets* the authors show that for any fair clustering problem it is enough to find good *fairlet decomposition* and then use existing algorithm. Here *fairlet decomposition* is one of the ways to partition the input. The choice of the partition directly impacts the approximation guarantees of the final clustering algorithm.

Moreover, in [Chierichetti et al. 2017] the computational cost for proposed algorithms for k-means objectives of finding good fairlets was proven to be NP-hard.

In the experimental part, the approximation approach opposing to the NP-hardness was described and evaluated. The approximations for k-center and k-means theoretically obtain 2- and 5- approximation respectively, yet perform way better in practice.

However, it is impossible to reformulate the clustering problem with original, unfair objective, such that the resulting clustering would be fair without loss of the quality of the original clustering itself. Hence, the *price of fairness* is introduced as the ratio of the cost of the original, unfair clustering to the cost of fair clustering. Then, unsurprisingly, the *balance* comes with a corresponding increase in *cost*, and the fair solutions are costlier than other unfair counterparts. In all the scenarios, the overall *cost* of clustering converges to the *cost* of the fairlet decomposition.

It is worth noting, that the fair clustering problem was formulated only under k-center and k-mean objectives. In the paper it is assumed that the maximum number of the protected classes is one, and the notion of balance is introduced with respect with having only two classed: red and blue points. The future work direction was featuring the extension of the protected class to the non-binary values, that is the direction of the work of [Rösner and Schmidt 2018] and further-mentioned [Bera et al. 2019].

3.2. Fair Algorithms for Clustering

[Bera et al. 2019] expand the work of [Chierichetti et al. 2017] in the regard that it generalizes the definition of the fairness in the context of clustering. With the new fairness notion, one individual can lie in multiple protected groups, or in other words, the presented algorithm allows for the use of non-disjoint classes. Another major contribution of the work is that proposed clustering algorithm works on any l_p-norm objective, which allows for the use of a wider range of algorithms. The comparison of the possible objectives for the clustering algorithms can be observed in Table 1. Any solution for many clustering objectives can be transformed into a fair version with only a slight loss in quality, while in [Chierichetti et al. 2017] only the optimal algorithm for the k-center objective was provided.

The presented fairness notion consists of the three main parameters. That is *restricted dominance*, *minority protection* and *maximum number of the protected groups*. The *restricted dominance* asserts that the fraction of people from the group i in any cluster is at most α_i. The *minority protection* asserts that the fraction of people from the group i in any cluster is at least β_i.The last parameter Δ is the *maximum number of protected groups*, to which a person can belong. The case of $\Delta = 1$ represents the case of disjoint groups.

The main algorithm of the paper is a simple two-step procedure: firstly solve the original clustering using some unfair algorithm, secondly solve a fair reassignment problem on the centers S acquired in the previous step to get the assignment ϕ. Then the algorithm returns (S,ϕ) as the fair solution. However, it has a notable disadvantage due to having the additive violation of of $(4\Delta + 3)$. The additive violation stands for the violations of the *restricted dominance* and *minority protection* constraints. It means, that up to the number of $(4\Delta + 3)$, the upper bound of *restricted dominance* and lower bound of *minority protection* can be violated. Practically, it means that there can be $2(4\Delta + 3)$ points violating the fairness constraints, or, in other words, belonging to the other group than they should according to the expanded $\alpha_i, \beta_i, \Delta$ fairness notion. However, practically

$(4\Delta + 3)$ shows itself negligible when the clusters are large. The empirical results run on five datasets from the UCI repository show that additive violation $(4\Delta + 3)$ almost never exceeds 3.

The *cost of fairness* is defined in the same way as in [Chierichetti et al. 2017] - as the ratio of the objective values of the fair clustering over the original, unfair clustering. The *balance* is, however, defined in a generalized way: it includes the values of representation of group i in the dataset (that is in practice recommended to be used to set the parameters α_i, β_i) and the representation of group i in the cluster f. Here, to reduce the degrees of freedom, authors introduce δ that can be used to set up both parameters α_i, β_i. With $\delta = 0.2$ corresponding to the common 80% rule of the disparate impact doctrine, we can practically set $\delta = 1 -$ *disparate impact rule*. Here, we can observe that introduced fairness notion is flexible, compliant to the disparate impact doctrine with the violation of the doctrine by $(4\Delta + 3)$.

Interestingly enough, for the case of overlapping protected groups, enforcing fairness with respect to one sensitive attribute can lead to unfairness with respect to the other. With respect to that, it is incredibly important to consider the case of overlapping groups when $\Delta > 1$.

3.3. Coresets for Clustering with Fairness Constraints

The paper of [Huang et al. 2019] is building upon any applicable clustering algorithm provided before, namely the method of coresets described can be used with the algorithms of the above-mentioned papers of [Bera et al. 2019] and [Chierichetti et al. 2017] as can be seen in Table 2.

Even though not being mentioned in the future work section nor for [Bera et al. 2019], and [Chierichetti et al. 2017], the scalability of the algorithms was an ongoing research direction with a high demand for the big volume datasets.

The main contribution of [Huang et al. 2019] providing a scalable algorithm for fair clustering, allowing for multiple, non-disjoint types. It is being done using coresets - small weighted point sets, such that the cost of the fair clustering objective computed over the subset is approximately the same as computed over the whole dataset. Applying coresets indeed accelerates the running time of computing fair clustering objective as it operates on far less number of data points, ensuring that the resulting objective difference is small. A lot of previous work was focused on scalable algortihms, yet at the time of publication of the paper there was no existing scalable algorithm with multiple types. Namely, the only result featuring coresets for fair clustering was an efficient streaming algorithm by [Schmidt et al. 2018] for constructing the k-means clustering. However, [Schmidt et al. 2018] restrict to one protected attribute, the derived coreset size includes log n factor and there is no coreset construction provided for other than k-means clustering.

The notion of fairness is defined as in [Bera et al. 2019] as the most general notion of fairness. The notion of *balance* is not used directly in the algorithm evaluation, as the coresets can be applied to any base algorithm. The *cost of fairness* is also not mentioned distinctly, as the idea of the paper is to provide the ϵ-coreset for already known dataset, fair clustering objective and fair clustering algorithm. Hence, the fairness, balance and cost of fairness is defined as in the base algorithm. The comparison in the paper is conducted

Table 1. Applicable algorithms and datasets. Notably, large scale dataset Census 1990 with 2.5 million records.

Paper	Applicable algorithms	Datasets
[Chierichetti et al. 2017]	k-center objective	Bank, Census, Diabetes
[Bera et al. 2019]	any l_p-norm objective	Bank, Census, Census 1990 (subsample), Credit-card, Diabetes
[Huang et al. 2019]	k-center objective, l_p-norm objective	Adult, Athlete, Bank, Census 1990, Diabetes

between the fair clustering objectives with the provided fairness constraints with ϵ-coreset and the whole dataset. It is valuable to notice, that the empirical error is the same measure as ϵ in the definition of the ϵ-coreset.

In [Huang et al. 2019] the authors construct the ϵ-coresets for both k-median and k-means. The coreset size is independent of n, hence with n being large, it brings about a great reduction of the coreset size comparing to the prior work of [Schmidt et al. 2018]. For k-median and k-means, the coresets sizes are $O(k^2\epsilon^{(-d)})$ and $O(k^3\epsilon^{(-d-1)})$, which improves the result of [Schmidt et al. 2018] for k-means and generalizes it to multiple, non-disjoint types, while for k-median it provides the first known coreset. Moreover, applying the coreset construction is not only restricted to the k-means objective: it might as well be used with any known fair clustering algorithm suffering only the $(1+\epsilon)$-factor in the approximation ratio. Applying it into the existing algorithm of [Bera et al. 2019], the method directly achieves scalable fair clustering algorithms. However, coreset size depends exponentially on the Euclidean dimension d, which is addressed at the related work section.

3.4. Paper Comparison

Firstly, [Chierichetti et al. 2017] initiate the studies of fairlets in the context of clustering. The exact algorithm of finding the fairlets is NP-hard, so approximation is used. The algorithms are given only for k-means and k-median objectives, with the 2- and 5-approximation, respectively. It is possible to use the algorithms only on the disjoint data types. For the possibility of the usage of non-disjoint data types and multiple protected classes, readers can refer to Table 4.

Based on this work, [Bera et al. 2019] introduce the fair clustering algorithms, that is addressing the problems of [Chierichetti et al. 2017]. Namely, it generalizes the notion of *balance*, allows the use of not only k-center objective, but also of any l_p-norm objective. Moreover, the algorithm enables the use of non-disjoint groups. Having a superior running time compared to [Chierichetti et al. 2017] the fair clustering solution however theoretically violates the restricted dominance and minority protection property by $(4\Delta + 3)$ as can be observed in Table 3, but in the experiments almost never being greater than 3. With the increasing number of centers for k-center objectives, the superiority of the running time of [Bera et al. 2019] becomes more and more noticeable. However, it was used only on small-scale datasets and the subsampled data of large scale dataset Census 1990 showing a new direction of the development of the algorithm - it's scalability.

Motivated to make the algorithms scalable for the large scale datasets,

Table 2. The connection between the papers

Paper	Key topic	Connection
[Chierichetti et al. 2017]	fairlets	initial study
[Bera et al. 2019]	tunable fairness notion: Δ, α, β	based on [Chierichetti et al. 2017], also works with non-disjoint types, l_p-norm objectives
[Huang et al. 2019]	coresets	introduces scalability, can be used with any algorithm provided before, such as i.e. [Bera et al. 2019] or [Chierichetti et al. 2017]

Table 3. Violation of the DI and approximation guarantees

Paper	Doctrine	Approximation
[Chierichetti et al. 2017]	Disparate Impact (DI)	2-, 5- for k-means and k-median
[Bera et al. 2019]	DI* violated by $(4\Delta + 3)$	$(\rho + 2)$ for unfair ρ-approximation algorithm
[Huang et al. 2019]	based on reference algo	$(1 + \epsilon)$ factor for non k-objective, see Tables 5,6 for k-objective approximation guarantees

[Huang et al. 2019] introduce the construction of coresets as a way to address the problem. Plugging the coresets is indeed resulting in faster algorithms for several cases. Namely, coreset construction can be used based on [Chierichetti et al. 2017] or on [Bera et al. 2019] algorithms. Hence, the [Huang et al. 2019] spawn wide variety of opportunities when it comes to the choice of the reference method. More precisely, if we need the exact solution with binary class and only one protected feature from [Chierichetti et al. 2017], or 2-approximation to it we can choose any of those methods described in [Chierichetti et al. 2017]. In a real-world scenario however, it is advantageous to use [Bera et al. 2019] with $(4\Delta + 3)$ additive violation, as it allows for non-disjoint groups and works on any l_p-norm objective. As we are looking at the problem from the perspective of large scale datasets, the violation which is empirically almost never greater than 3 can be a reasonable trade-off. So, using the algorithm of [Bera et al. 2019] in a black-box fashion, we can construct a coreset giving an approximate solution to the problem with the superior runtime. The running time comparison, as well as approximation ratios for k-median and k-means are summarized in Tables 5 and 6. In the brackets we can see that the violation of the disparate impact doctrine is $(4\Delta + 4)$, unlike in [Bera et al. 2019] with $(4\Delta + 3)$, which derivation is not justified in [Huang et al. 2019]. $T_1(n)$ and $T_2(n)$ denote the running time of an $O(1)$-approximate algorithm for k-median and k-means respectively.

4. Discussion

Disparate treatment is intentional discrimination opposed to disparate impact being unintentional in a way blind. For example, testing a particular skill of only certain minority applicants is disparate treatment, which his considered to be unlawful. From

Table 4. Number of protected classes, possibility of no-disjoint groups

Paper	Non-disjoint classes possible	Number of protected classes
[Chierichetti et al. 2017]	no	one
[Bera et al. 2019]	yes	multiple
[Huang et al. 2019]	yes	multiple

Table 5. k-median approximation ratio, additive violation and runtime

Paper	(Approx ratio, DI violation)	Runtime
[Chierichetti et al. 2017]	($O(1)$,0)	$\Omega(n^2)$
[Bera et al. 2019]	$(O(1), 4\Delta + 4)$	$\Omega(n^2)$
[Huang et al. 2019]	based on reference algorithm	
base algorithm:		
[Chierichetti et al. 2017]	$(\widetilde{O}(d \log n), 0)$	$O(dlk^2 \log (lk) + T_1(lk^2))$
[Bera et al. 2019]	$(O(1), 4\Delta + 4)$	$\Omega(l^{2\Delta}k^4)$

that perspective, it also would be possible to argue, that is disparate treatment is not taking place, then the decision made, for instance at the workplace and so on, would be fair by default.

However, this is not true. The disparate impact as another and more subtle type of discrimination, has received a lot of attention from the scientific community. It led to the interesting discussion about the use of disparate impact as the base of the ML algorithms, namely in clustering algorithms discussed in the "Fair clustering" session. As stated at the MLSS 2020 presentation, *fairness trough unawareness fails*, summarizing the reasoning behind the fair clustering algorithms. It is not enough for the algorithm to be colour-blind like in the example of [Chierichetti et al. 2017], as the protected attributes have to be indeed taken into consideration explicitly. Statistical fairness, on the other hand, can not be a proof of fairness, as statistical algorithms is exposing the biases that are already present in the data collected.

The work of fairness is mostly touching the notion of fairness only since 2010, having an explosive increase since [Chierichetti et al. 2017] paper in the clustering field. As before the fairness algorithms were concentrated on the fair classification as stated in the systematic literature review section. However, the pioneering works on fairness outside of the ML context are dating back to the late sixties [Becker 1957].

The discussion about the fairness and the approximation boundaries also take

Table 6. k-means approximation ratio, additive violation and runtime

Paper	(Approx ratio, DI violation)	Runtime
[Chierichetti et al. 2017]		
[Bera et al. 2019]	$(O(1), 4\Delta + 4)$	$\Omega(n^2)$
[Huang et al. 2019]	based on reference algorithm	
base algorithm:		
[Chierichetti et al. 2017]		
[Bera et al. 2019]	$(O(1), 4\Delta + 4)$	$\Omega(l^{2\Delta}k^6)$

another turn when being applied to the social sciences fields. Indeed, it would not be a problem to have an additive violation of the minority protection and restricted dominance properties, given that the algorithm does not have a crucial impact over our lives. However, as fascinating and frightening as it is, ML affecting more fields than we can imagine - from giving the allowance for a loan or assigning a person to a certain group that would further affect the person's life. Before doing that, we have to be absolutely sure, that the algorithm's disadvantages cannot be exposed when being applied to the real-world scenario. The topic of fairness, being really sensitive field, requires some additional input not only from the perspective of machine learning, but also from the inter-disciplinary field, including social studies.

References

Becker, G. (1957). *The Economics of Discrimination*. Chicago University Press, Chicago, first edition.

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. (2019). Fair algorithms for clustering. In *Advances in Neural Information Processing Systems*, pages 4954–4965.

Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? *CoRR*, abs/1607.06520.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806. ACM.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*.

Edwards, H. and Storkey, A. J. (2016). Censoring representations with an adversary. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Huang, L., Jiang, S., and Vishnoi, N. (2019). Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems*, pages 7589–7600.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Rösner, C. and Schmidt, M. (2018). Privacy preserving clustering with constraints. In Chatzigiannakis, I., Kaklamanis, C., Marx, D., and Sannella, D., editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018,*

July 9-13, 2018, Prague, Czech Republic, volume 107 of *LIPIcs*, pages 96:1–96:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Schmidt, M., Schwiegelshohn, C., and Sohler, C. (2018). Fair coresets and streaming algorithms for fair k-means clustering. *CoRR*, abs/1812.10854.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.