Probabilistic Fair Clustering

Seyed A. Esmaeili[1,3], Brian Brubach[2,4], Leonidas Tsepenekas[1,3], and John P. Dickerson [1,3]

[1]Department of Computer Science, University of Maryland, College Park
[2]Department of Computer Science, Wellesley College
[3]{esmaeili,ltsepene,john}@cs.umd.edu
[4]bb100@wellesley.edu

Abstract

In clustering problems, a central decision-maker is given a complete metric graph over vertices and must provide a clustering of vertices that minimizes some objective function. In *fair* clustering problems, vertices are endowed with a *color* (e.g., membership in a group), and the requirements of a valid clustering might also include the representation of colors in the solution. Prior work in fair clustering assumes complete knowledge of group membership. In this paper, we generalize this by assuming imperfect knowledge of group membership through probabilistic assignments, and present algorithms in this more general setting with approximation ratio guarantees. We also address the problem of "metric membership", where group membership has a notion of order and distance. Experiments are conducted using our proposed algorithms as well as baselines to validate our approach, and also surface nuanced concerns when group membership is not known deterministically.

1 Introduction

Machine-learning-based decisioning systems are increasingly used in high-stakes situations, many of which directly or indirectly impact society. Examples abound of automated decisioning systems resulting in, arguably, morally repugnant outcomes: hiring algorithms may encode the biases of human reviewers' training data Bogen and Rieke [2018], advertising systems may discriminate based on race and inferred gender in harmful ways Sweeney [2013], recidivism risk assessment software may bias its risk assessment improperly by race Angwin et al. [2016], and healthcare resource allocation systems may be biased against a specific race Ledford [2019]. A myriad of examples such as these and others motivate the growing body of research into defining, measuring, and (partially) mitigating concerns of fairness and bias in machine learning. Different metrics of algorithmic fairness have been proposed, drawing on prior legal rulings and philosophical concepts; Mehrabi et al. [2019] give a recent overview of sources of bias and fairness as presently defined by the machine learning community.

The earliest work in this space focused on fairness in *supervised* learning Luong et al. [2011], Hardt et al. [2016] as well as *online* learning Joseph et al. [2016]; more recently, the literature has begun expanding into fairness in *unsupervised* learning Chierichetti et al. [2017]. In this work, we address a novel model of fairness in clustering—a fundamental unsupervised learning problem. Here, we are given a complete metric graph where each vertex also has a color associated with it, and we are concerned with finding a clustering that takes both the metric graph and vertex colors into account. Most of the work in this area (e.g., Ahmadian et al. [2019a], Bercea et al. [2018], Chierichetti et al. [2017]) has defined a fair clustering to be one that minimizes the cost function subject to the constraint that each cluster satisfies a lower and an upper bound on the percentage of each color it contains—a form of approximate *demographic parity* or its closely-related cousin, the $p\%$-rule Biddle [2006].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

We relax the assumption that a vertex's color assignment is known deterministically; rather, for each vertex, we assume only knowledge of a distribution over colors.

Our proposed model addresses many real-world use cases. Ahmadian et al. [2019a] discuss clustering news articles such that no political viewpoint—assumed to be known deterministically—dominates any cluster. Here, the color membership attribute—i.e., the political viewpoint espoused by a news article—would not be provided directly but could be predicted with some probability of error using other available features. Awasthi et al. [2019] discuss the case of supervised learning when class labels are not known with certainty (e.g., due to noisy crowdsourcing or the use of a predictive model). Our model addresses such motivating applications in the unsupervised learning setting, by defining a fair cluster to be one where the color proportions satisfy the upper and lower bound constraints *in expectation*. Hence, it captures standard deterministic fair clustering as a special case.

Outline & Contributions. We begin (§2) with an overview of related research in general clustering, fairness in general machine learning, as well as recent work addressing fairness in unsupervised learning. Next (§3), we define two novel models of clustering when only probabilistic membership is available: the first assumes that colors are unordered, and the second embeds colors into a metric space, thus endowing them with a notion of order and distance. This latter setting addresses use cases where, e.g., we may want to cluster according to membership in classes such as age or income, whose values are naturally ordered. Following this (§4), we present two approximation algorithms with theoretical guarantees in the settings above. We also briefly address the (easier but often realistic) "large cluster" setting, where it is assumed that the optimal solution does not contain pathologically small clusters. Finally (§5), we verify our proposed approaches on four real-world datasets. We note that all proofs are put in the appendix due to the page limit.

2 Related Work

Classical forms of the metric clustering problems k-center, k-median, and k-means are well-studied within the context of unsupervised learning and operations research. While all of these problems are NP-hard, there is a long line of work on approximating them and heuristics are commonly used in many practical applications. This vast area is surveyed by Aggarwal and Reddy [2013] and we focus on approximation algorithms here. For k-center, there are multiple approaches to achieve a 2-approximation and this is the best possible unless $P = NP$ Hochbaum and Shmoys [1985], Gonzalez [1985a], Hochbaum and Shmoys [1986]. Research for the best approximations to k-median and k-means is ongoing. For k-median there is a $(2.675 + \epsilon)$-approximation with a running time of $n^{O((1/\epsilon)\log(1/\epsilon))}$ Byrka et al. [2014], and for k-means a 6.357-approximation is the best known Ahmadian et al. [2019b].

The study of approximation algorithms that achieve demographic fairness for metric clustering was initiated by Chierichetti et al. [2017]. They considered a variant of k-center and k-median wherein each point is assigned one of two colors and the color of each point is known. Followup work extended the problem setting to the k-means objective, multiple colors, and the possibility of a point being assigned multiple colors (i.e. modeling intersecting demographic groups such as gender and race combined) Bercea et al. [2018], Bera et al. [2019], Backurs et al. [2019], Huang et al. [2019]. Other work considers the one-sided problem of preventing over-representation of any one group in each cluster rather than strictly enforcing that clusters maintain proportionality of all groups Ahmadian et al. [2019a].

In all of the aforementioned cases, the colors (demographic groups) assigned to the points are known a priori. By contrast, we consider a generalization where points are assigned a distribution on colors. We note that this generalizes settings where each point is assigned a single deterministic color. Moreover, our setting is distinct from the setting where points are assigned multiple colors in that we assume each point has a single true color. In the area of supervised learning, the work of Awasthi et al. [2019] addressed a similar model of uncertain group membership. Other recent work explores unobserved protected classes from the perspective of assessment Kallus et al. [2019]. However, no prior work has addressed this model of uncertainty for metric clustering problems in unsupervised learning.

3 Preliminaries and Problem Definition

Let $\mathcal{C}$ be the set of points in a metric space with distance function $d : \mathcal{C} \times \mathcal{C} \to \mathbf{R}_{\geq 0}$. The distance between a point v and a set S is defined as $d(v, S) = \min_{j \in S} d(v, j)$. In a k-clustering an objective function $L^k(\mathcal{C})$ is given, a set $S \subseteq \mathcal{C}$ of at most k points must be chosen as the set of centers, and each point in $\mathcal{C}$ must get assigned to a center in S through an assignment function $\phi : \mathcal{C} \to S$, forming a k-partition of the original set: $\mathcal{C}_1, \ldots, \mathcal{C}_k$. The optimal solution is defined as a set of centers and an assignment function that minimizes the objective $L^k(\mathcal{C})$. The well known k-center, k-median, and k-means can all be stated as the following problem:

$$\min_{S:|S|\leq k,\phi} L_p^k(\mathcal{C}) = \min_{S:|S|\leq k,\phi} \Big(\sum_{v \in \mathcal{C}} d^p(v, \phi(v)) \Big)^{1/p} \tag{1}$$

where p equals $\infty, 1$, and 2 for the case of the k-center, k-median, and k-means, respectively. For such problems the optimal assignment for a point v is the nearest point in the chosen set of centers S. However, in the presence of additional constraints such as imposing a lower bound on the cluster size Aggarwal et al. [2010] or an upper bound Khuller and Sussmann [2000], Cygan et al. [2012], An et al. [2015] this property no longer holds. This is also true for fair clustering.

To formulate the fair clustering problem, a set of colors $\mathcal{H} = \{h_1, \ldots, h_\ell, \ldots, h_m\}$ is introduced and each point v is mapped to a color through a given function $\chi : \mathcal{C} \to \mathcal{H}$. Previous work in fair clustering Chierichetti et al. [2017], Ahmadian et al. [2019a], Bercea et al. [2018], Bera et al. [2019] adds to (1) the following proportional representation constraint, i.e.:

$$\forall i \in S, \forall\, h_\ell \in \mathcal{H} : l_{h_\ell} |\, \mathcal{C}_i \,| \leq |\, \mathcal{C}_{i,h_\ell} \,| \leq u_{h_\ell} |\, \mathcal{C}_i \,| \tag{2}$$

where $\mathcal{C}_{i,h_\ell}$ is the set of points in cluster i having color h_ℓ. The bounds $l_{h_\ell}, u_{h_\ell} \in (0, 1)$ are given lower and upper bounds on the desired proportion of a given color in each cluster, respectively.

In this work we generalize the problem by assuming that the color of each point is not known deterministically but rather probabilistically. We also address the case where the colors lie in a 1-dimensional Euclidean metric space.

3.1 Probabilistic Fair Clustering

In *probabilistic fair clustering*, we generalize the problem by assuming that the color of each point is not known deterministically but rather probabilistically. That is, each point v has a given value $p_v^{h_\ell}$ for each $h_\ell \in \mathcal{H}$, representing the probability that point v has color h_ℓ, with $\sum_{h_\ell \in \mathcal{H}} p_v^{h_\ell} = 1$.

The constraints are then modified to have the expected color of each cluster fall within the given lower and upper bounds. This leads to the following optimization problem:

$$\min_{S:|S|\leq k,\phi} L_p^k(\mathcal{C}) \tag{3a}$$

$$\text{s.t. } \forall i \in S, \forall\, h_\ell \in \mathcal{H} : l_{h_\ell} |\phi^{-1}(i)| \leq \sum_{v \in \phi^{-1}(i)} p_v^{h_\ell} \leq u_{h_\ell} |\phi^{-1}(i)| \tag{3b}$$

Following Bera et al. [2019], we define a γ violating solution to be one for which for all $i \in S$:

$$l_{h_\ell} |\phi^{-1}(i)| - \gamma \leq \sum_{v \in \phi^{-1}(i)} p_v^{h_\ell} \leq u_{h_\ell} |\phi^{-1}(i)| + \gamma \tag{4}$$

This effectively captures the amount γ, by which a solution violates the fairness constraints.

3.2 Metric Membership Fair Clustering

Representing a point's (individual's) membership using colors may be sufficient for binary or other unordered categorical variables. However, this may leave information "on the table" when a category is, for example, income or age, since colors do not have an inherent sense of order or distance.

For this type of attribute, the membership can be characterized by a 1-dimensional Euclidean space. Without loss of generality, we can represent the set of all possible memberships as the set of all

consecutive integers from 0 to some $R > 0$, where R is the maximum value that can be encountered. Hence, let $\mathcal{H}_R = \{0, 1, \ldots, R\}$. Each point v has associated with it a value $r_v \in \mathcal{H}_R$. In this problem we require the average total value of each cluster to be within a given interval. Hence:

$$\min_{S:|S|\leq k,\phi} L_p^k(\mathcal{C}) \tag{5a}$$

$$\text{s.t. } \forall i \in S : l|\phi^{-1}(i)| \leq \sum_{v\in\phi^{-1}(i)} r_v \leq u|\phi^{-1}(i)| \tag{5b}$$

where l and u are respectively upper and lower bounds imposed on each cluster. We do not have a subscript h_ℓ for either l or u because we essentially only have one color (the metric membership value).

Similar to section 3.1, we define a γ violating solution to be one for which $\forall i \in S$:

$$l|\phi^{-1}(i)| - \gamma \leq \sum_{v\in\phi^{-1}(i)} r_v \leq u|\phi^{-1}(i)| + \gamma \tag{6}$$

If we consider the case of income, then the objective of (5) can be used to force each cluster to have an average income around the global average preventing the possibility of having low or high income individuals from being over or under represented in any given cluster.

4 Approximation Algorithms and Theoretical Guarantees

We essentially have two algorithms although they involve similar steps. One algorithm is for the two-color and metric membership case which is discussed in section (4.1) and the other algorithm us for the multiple-color case under a large cluster assumption which is discussed in section (4.2).

4.1 Algorithms for the Two Color and Metric Membership Case

Our algorithm follows the two step method of Bera et al. [2019], although we differ in the LP rounding scheme. Let $\mathrm{PFC}(k, p)$ denote the probabilistic fair clustering problem. The color-blind clustering problem, where we drop the fairness constraints, is denoted by $\mathrm{Cluster}(k, p)$. Further, define the fair assignment problem FA-$\mathrm{PFC}(S, p)$ as the problem where we are given a fixed set of centers S and the objective is to find an assignment ϕ minimizing $L_p^k(\mathcal{C})$ and satisfying the fairness constraints 3b for probabilistic fair clustering or 5b for metric-membership. We prove the following (similar to theorem 2 in Bera et al. [2019]):

Theorem 4.1. *Given an α-approximation algorithm for $\mathrm{Cluster}(k, p)$ and a γ-violating algorithm for FA-$\mathrm{PFC}(S, p)$, a solution with approximation ratio $\alpha + 2$ and constraint violation at most γ can be achieved for $\mathrm{PFC}(k, p)$.*

An identical theorem and proof follows for the metric membership problem as well.

4.1.1 Step 1, Color-Blind Approximation Algorithm:

At this step an ordinary (color-blind) α-approximation algorithm is used to find the cluster centers. For example, the Gonzalez algorithm Gonzalez [1985b] can be used for the k-center problem or the algorithm of Byrka et al. [2014] can be used for the k-median. This step results in a set S of cluster centers. Since this step does not take fairness into account, the resulting solution does not necessarily satisfy constraints 3b for probabilistic fair clustering and 5b for metric-membership.

4.1.2 Step 2, Fair Assignment Problem:

In this step, a linear program (LP) is set up to satisfy the fairness constraints. The variables of the LP are x_{ij} denoting the assignment of point j to center i in S. Specifically, the LP is:

$$\min \sum_{j\in\mathcal{C}, i\in S} d^p(i, j)x_{ij} \tag{7a}$$

$$\text{s.t. } \forall i \in S \text{ and } \forall\, h_\ell \in \mathcal{H} : \tag{7b}$$

$$l_{h_\ell} \sum_{j \in \mathcal{C}} x_{ij} \leq \sum_{j \in \mathcal{C}} p_j^{h_\ell} x_{ij} \leq u_{h_\ell} \sum_{j \in \mathcal{C}} x_{ij} \tag{7c}$$

$$\forall j \in \mathcal{C} : \sum_{i \in S} x_{ij} = 1, \quad 0 \leq x_{ij} \leq 1 \tag{7d}$$

Since the LP above is a relaxation of FA-PFC(S, p), we have $\text{OPT}^{\text{LP}}_{\text{FA-PFC}} \leq \text{OPT}_{\text{FA-PFC}}$. We note that for k-center there is no objective, instead we have the following additional constraint: $x_{ij} = 0$ if $d(i, j) > w$ where w is a guess of the optimal radius. Also, for k-center the optimal value is always the distance between two points. Hence, through a binary search over the polynomially-sized set of distance choices we can WLOG obtain the minimum satisfying distance. Further, for the metric membership case $p_j^{h_\ell}, l_{h_\ell}$ and u_j in 7c are replaced by r_j, l and u, respectively.

What remains is to round the fractional assignments x_{ij} resulting from solving the LP.

4.1.3 Rounding for the Two Color and Metric Membership Case

First we note the connection between the metric membership problem and the two color case of probabilistic fair clustering. Effectively the set $\mathcal{H}_R = \{0, 1, \ldots, R\}$ is the unnormalized version of the set of probabilities $\{0, \frac{1}{R}, \frac{2}{R}, \ldots, 1\}$. Our rounding method is based on calculating a minimum-

Algorithm 1 Form Flow Network Edges for Culster C_i

$\vec{A}_i$ are the points $j \in \phi^{-1}(i)$ in non-increasing order of p_j
initialize array $\vec{a}$ of size $|C_i|$ to zeros, and set $s = 1$
put the assignment x_{ij} for each point j in $\vec{A}_i$ in $\vec{z}_i$ according the vertex order in $\vec{A}_i$
for $q = 1$ **to** $|C_i|$ **do**
  $\vec{a}(q) = \vec{a}(q) + x_{i\vec{A}_i(s)}$, and add edge $(\vec{A}_i(s), q)$
  $\vec{z}_i(s) = 0$
  $s = s + 1$ {Move to the next vertex}
  repeat
    valueToAdd $= min(1 - \vec{a}(q), \vec{z}_i(s))$
    $\vec{a}(q) = \vec{a}(q) +$ valueToAdd, and add edge $(\vec{A}_i(s), q)$
    $\vec{z}_i(s) = \vec{z}_i(s) -$ valueToAdd
    if $\vec{z}_i(s) = 0$ **then**
      $s = s + 1$
    end if
  until $\vec{a}(q) = 1$ or $s > |\vec{A}_i|$ {until we have accumulated 1 or ran out of vertices}
end for

cost flow in a carefully constructed graph. For each $i \in S$, a set C_i with $|C_i| = \left\lceil \sum_{j \in \mathcal{C}} x_{ij} \right\rceil$ vertices is created. Moreover, the set of vertices assigned to cluster i, i.e. $\phi^{-1}(i) = \{j \in \mathcal{C} \,|x_{ij} > 0\}$ are sorted in a non-increasing order according to the associated value r_j and placed into the array $\vec{A}_i$. A vertex in C_i (except possibly the last) is connected to as many vertices in $\vec{A}_i$ by their sorting order until it accumulates an assignment value of 1. A vertex in $\vec{A}_i$ may be connected to more than one vertex in C_i if that causes the first vertex in C_i to accumulate an assignment value of 1 with some assignment still remaining in the $\vec{A}_i$ vertex. In this case the second vertex in C_i would take only what remains of the assignment. See Algorithm 1 for full details. Appendix C demonstrates an example.

We denote the set of edges that connect all points in $\mathcal{C}$ to points in C_i by $E_{\mathcal{C},C_i}$. Also, let $V_{\text{flow}} = \mathcal{C} \cup (\cup_{i \in S} C_i) \cup S \cup \{t\}$ and $E_{\text{flow}} = E_{\mathcal{C},C_i} \cup E_{C_i,S} \cup E_{S,t}$, where $E_{C_i,S}$ has an edge from every vertex in C_i to the corresponding center $i \in S$. Finally $E_{S,t}$ has an edge from every vertex i in S to the sink t if $\sum_{j \in \mathcal{C}} x_{ij} > \left\lfloor \sum_{j \in \mathcal{C}} x_{ij} \right\rfloor$. The demands, capacities, and costs of the network are:

- **Demands**: Each $v \in \mathcal{C}$ has demand $d_v = -1$ (a supply of 1), $d_u = 0$ for each $u \in C_i$, $d_i = \left\lfloor \sum_{j \in \mathcal{C}} x_{ij} \right\rfloor$ for each $i \in S$. Finally t has demand $d_t = |\mathcal{C}| - \sum_{i \in S} d_i$.

- **Capacities**: All edge capacities are set to 1.
- **Costs**: All edges have cost 0, expect the edges in $E_{\mathcal{C},C_i}$ where $\forall (v,u) \in E_{\mathcal{C},C_i}, d(v,u) = d(v,i)$ for the k-median and $d(v,u) = d^2(v,i)$. For the k-center, either setting suffices.

See Figure 1 for an example. It is clear that the entire demand is $|\mathcal{C}|$ and that this is the maximum possible flow. The LP solution attains that flow. Further, since the demands, capacities and distances are integers, an optimal integral minimum-cost flow can be found in polynomial time. If $\bar{x}_{ij}$ is the integer assignment that resulted from the flow computation, then violations are as follows:

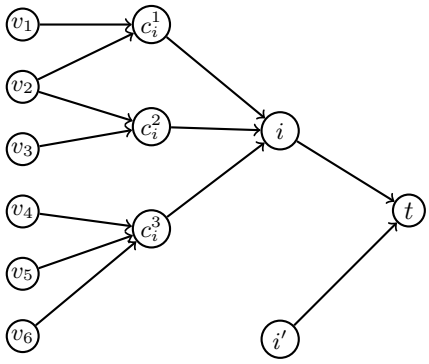

Figure 1: Network flow construction.

Theorem 4.2. *The number of vertices assigned to a cluster (cluster size) is violated by at most 1, i.e.* $|\sum_{j\in\mathcal{C}} \bar{x}_{ij} - \sum_{j\in\mathcal{C}} x_{ij}| \leq 1$. *Further for metric membership, the violation in the average value is at most R, i.e.* $|\sum_{j\in\mathcal{C}} \bar{x}_{ij} r_j - \sum_{j\in\mathcal{C}} x_{ij} r_j| \leq R$. *It follows that for the probabilistic case, the violation in the expected value is at most 1.*

Our rounding scheme results in a violation for the two color probabilistic case that is at most 1, whereas for metric-membership it is R. The violation of R for the metric membership case suggests that the rounding is too loose, therefore we show a lower bound of at least $\frac{R}{2}$ for any rounding scheme applied to the resulting solution. This also makes our rounding asymptotically optimal.

Theorem 4.3. *Any rounding scheme applied to the resulting solution has a fairness constraint violation of at least* $\frac{R}{2}$ *in the worst case.*

4.2 Algorithms for the Multiple Color Case Under a Large Cluster Assumption:

First, we point out that for the multi-color case, the algorithm is based on the assumption that the cluster size is large enough. Specifically:

Assumption 4.1. *Each cluster in the optimal solution should have size at least* $L = \Omega(n^r)$ *where* $r \in (0,1)$.

We firmly believe that the above is justified in real datasets. Nonetheless, the ability to manipulate the parameter r, gives us enough flexibility to capture all occurring real-life scenarios.

Theorem 4.4. *If Assumption 4.1 holds, then independent sampling results in the amount of color for each clusters to be concentrated around its expected value with high probability.*

Given Theorem 4.4 our solution essentially forms a reduction from the problem of probabilistic fair clustering $\mathrm{PFC}(k,p)$ to the problem of deterministic fair clustering with lower bounded cluster sizes which we denote by $\mathrm{DFC}_{\mathrm{LB}}(k,p,L)$ (the color assignments are known deterministically and each cluster is constrained to have size at least L). Our algorithm (2) involves three steps. In the first

Algorithm 2 Algorithm for Large Cluster $\mathrm{PFC}(k,p)$

Input: $\mathcal{C}, d, k, p, L, \{(l_{h_\ell}, u_{h_\ell})\}_{h_\ell \in \mathcal{H}}$
Relax the upper and lower by ϵ: $\forall\, h_\ell \in \mathcal{H}$, $l_{h_\ell} \leftarrow l_{h_\ell}(1-\epsilon)$ and $u_{h_\ell} \leftarrow u_{h_\ell}(1+\epsilon)$
For each point $v \in \mathcal{C}$ sample its color independently according to $p_v^{h_\ell}$
Solve the deterministic fair clustering problem with lower bounded clusters $\mathrm{DFC}_{\mathrm{LB}}(k,p,L)$ over the generated instance and return the solution.

step, the upper and lower bounds are relaxed since -although we have high concentration guarantees around the expectation- in the worst case the expected value may not be realizable (could not be an integer). Moreover the upper and lower bounds could coincide with the expected value causing violations of the bounds with high probability. See appendix B for more details.

After that, the color assignments are sampled independently. The following deterministic fair clustering problem is solved for resulting set of points:

(8a)

$$\text{s.t. } \forall i \in S : (1-\delta) l_{h_\ell} |\, \mathcal{C}_i \,| \leq |\, \mathcal{C}_{i,h_\ell} \,| \leq (1+\delta) u_{h_\ell} |\, \mathcal{C}_i \,| \tag{8b}$$

$$\forall i \in S : |\, \mathcal{C}_i \,| \geq L \tag{8c}$$

The difference between the original deterministic fair clustering problem and the above is that the bounds are relaxed by ϵ and a lower bound L is required on the cluster size. This is done in order to guarantee that the resulting solution satisfies the relaxed upper and lower bounds in expectation, because small size clusters do not have a Chernoff bound and therefore nothing ensures that they are valid solutions to the original $\mathrm{PFC}(k, p)$ problem.

The algorithm for solving deterministic fair clustering with lower bounded cluster sizes $\mathrm{DFC}_{\mathrm{LB}}$ is identical to the algorithm for solving the original deterministic fair clustering Bera et al. [2019], Bercea et al. [2018] problem with the difference being that the setup LP will have a bound on the cluster size. That is we include the following constraint $\forall i \in S : \sum_{ij} x_{ij} \geq L$. See appendix E for further details. In theorem A.2 we show that it has an approximation ratio of $\alpha + 2$ like the ordinary (deterministic) fair clustering case, where again α is the approximation ratio of the color blind algorithm.

Theorem 4.5. *Given an instance of the probabilistic fair clustering problem* $\mathrm{PFC}(k, p)$*, with high probability algorithm 2 results in a solution with violation at most* ϵ *and approximation ratio* $(\alpha + 2)$.

5 Experiments

We now evaluate the performance of our algorithms over a collection of real-world datasets. We give experiments in the two (unordered) color case (§5.2), metric membership (i.e., ordered color) case (§5.3), as well as under the large cluster assumption (§5.4). We include experiments for the k-means case here, and the (qualitatively similar) k-center and k-median experiments to Appendix F.

5.1 Experimental Framework

Hardware & Software. We used only commodity hardware through the experiments: Python 3.6 on a MacBook Pro with 2.3GHz Intel Core i5 processor and 8GB 2133MHz LPDDR3 memory. A state-of-the-art commercial optimization toolkit, `CPLEX` Manual [2016], was used for solving all LPs. `NetworkX` Hagberg et al. [2013] was used to solve minimum cost flow problems, and `Scikit-learn` Pedregosa et al. [2011] for standard machine learning tasks such as training SVMs, pre-processing, and performing traditional k-means clustering.

Color-Blind Clustering. The color-blind clustering algorithms we use are as follows.

- Gonzalez [1985b] gives a 2-approximation for k-center.
- We use `Scikit-learn`'s k-means++ module.
- A 5-approximation algorithm due to Arya et al. [2004] modified with D-sampling Arthur and Vassilvitskii [2006] according to Bera et al. [2019].

Generic-Experimental Setup and Measurements. For a chosen dataset, a given color h_ℓ would have a proportion $f_{h_\ell} = \frac{|v \in \mathcal{C} \,|\chi(v) = h_\ell|}{|\, \mathcal{C} \,|}$. Following Bera et al. [2019], the lower bound is set to $l_{h_\ell} = (1-\delta) r_{h_\ell}$ and the upper bound is to $u_{h_\ell} = \frac{f_{h_\ell}}{(1-\delta)}$. For metric membership, we similarly have $f = \frac{\sum_{v \in \mathcal{C}} r_v}{|\, \mathcal{C} \,|}$ as the proportion, $l = (1-\delta) f$ and $u = \frac{f}{1-\delta}$ as the lower and upper bound, respectively. We set $\delta = 0.2$, as Bera et al. [2019] did, unless stated otherwise.

For each experiment, we measure the price of fairness $\mathrm{POF} = \frac{\text{Fair Solution Cost}}{\text{Color-Blind Cost}}$. We also measure the maximum additive violation γ as it appears in inequalities 4 and 6.

5.2 Two Color Case

Here we test our algorithm for the case of two colors with probabilistic assignment. We use the **Bank** dataset Moro et al. [2014] which has 4,521 data points. We choose marital status, a categorical variable, as our fairness (color) attribute. To fit the binary color case, we merge single and divorced into one category. Similar to the supervised learning work due to Awasthi et al. [2019], we make

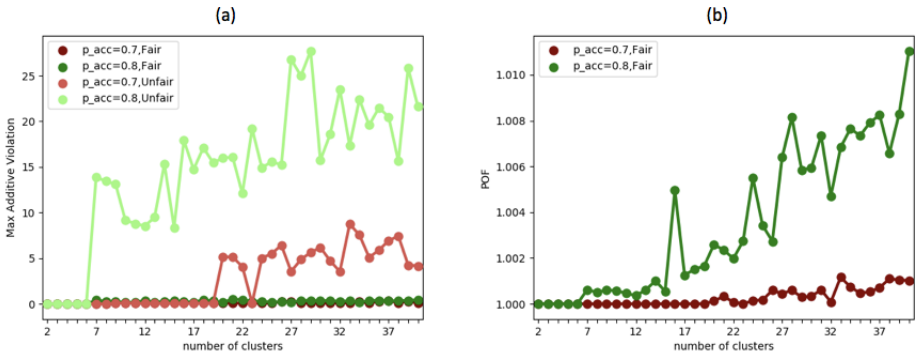

Figure 2: For $p_{\text{acc}} = 0.7$ & $p_{\text{acc}} = 0.8$, showing (a): #clusters vs. maximum additive violation; (b): #clusters vs. POF .

Bank's deterministic color assignments probabilistic by independently perturbing them for each point with probability p_{noise}. Specifically, if v originally had color c_v, then now it has color c_v with probability $1 - p_{\text{noise}}$ instead. To make the results more interpretable, we define $p_{\text{acc}} = 1 - p_{\text{noise}}$. Clearly, $p_{\text{acc}} = 1$ corresponds to the deterministic case, and $p_{\text{acc}} = \frac{1}{2}$ corresponds to completely random assignments.

In Fig. 2(a), we see that the violations of the color-blind solution can be as large as 25 while our algorithm is within the theoretical guarantee that is less than 1. In Fig. 2(b), we see that in spite of the large violation, fairness is achieved at a low relative efficiency loss, not exceeding 2% (POF ≤ 1.02). In appendix F.1 we show further experiments which explore the relationship between p_{acc} and POF. We also show that the simple solution of assigning the most probable color does not work.

5.3 Metric Membership

Here we test our algorithm for the metric membership problem. We use two additional well-known datasets: **Adult** Kohavi [1996], with age being the fairness attribute, and **CreditCard** Yeh and Lien [2009], with credit being the fairness attribute. We apply a pre-processing step where for each point we subtract the minimum value of the fairness attribute over the entire set. This has the affect of reducing the maximum fairness attribute value, thus reducing the maximum violation of $\frac{1}{2}R$, but still keeping the values non-negative.

Fig. 3 (a) shows POF with respect to the number of clusters. For the **Adult** dataset, POF is at most less than 5%, whereas for the **CreditCard** dataset it is as high at 25%. While the POF, intuitively, rises with the number of clusters allowed, it is substantially higher with the **CreditCard** dataset. This may be explained because of the correlation that exists between credit and other features represented in the metric space.

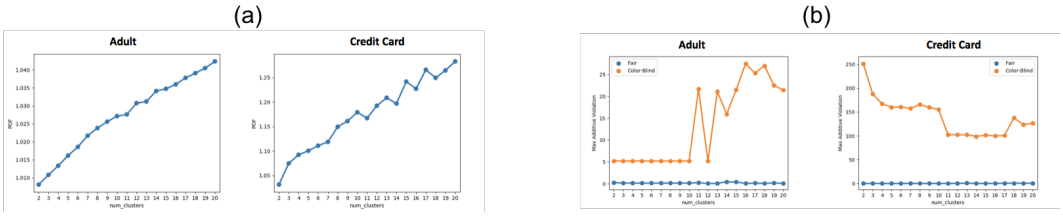

Figure 3: **Adult** and **CreditCard** dataset results: (a) #Clusters (x-axis) vs. POF (y-axis), and (b):#Clusters (x-axis) vs. normalized maximum additive violation (y-axis).

In Fig. 3 (b), we compare the number of clusters against the normalized maximum additive violation. The normalized maximum additive violation is the same maximum additive violation γ from inequality 6—but normalized by R. We see that the normalized maximum additive violation is indeed less than 1 as theoretically guaranteed by our algorithm, where for the color-blind solution it is as high as 250.

5.4 The Large Cluster Assumption

Here we test our algorithm for the case of probabilistically assigned multiple colors under Assumption 4.1, which addresses cases where the optimal clustering does not include pathologically small clusters. We use the **Census1990** Meek et al. [2002] dataset. We note that **Census1990** is large, with over 2.4 million points. We use age groups (attribute `dAge` in the dataset) as our fairness attribute, which yields 7 age groups (colors).[1] We then sample 100,000 data points and use them to train an SVM classifier[2] to predict the age group memberships. The classifier achieves an accuracy of around 68%. We use the classifier to predict the memberships of another 100,000 points not included in the training set, and sample from that to form the probabilistic assignment of colors.

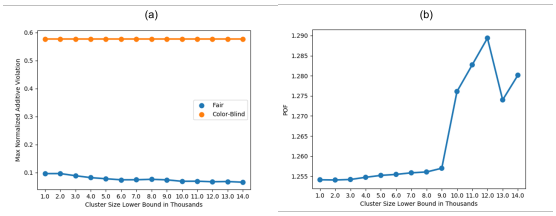

Figure 4: Plot showing the performance of our independent sampling algorithm over the **Census1990** dataset for $k = 5$ clusters with varying values on the cluster size lower bound: (a) maximum violation normalized by the cluster size, (b) the price of fairness.

Fig. 4 shows the output of our large cluster algorithm over 100,000 points and $k = 5$ clusters with varying lower bound assumptions. Since the clusters here are large, we normalize the additive violations by the cluster size. We see that our algorithm results in normalized violation that decrease as the lower bound on the cluster size increases—eventually dropping below 20%. The POF is high relative to our previous experiments, but still less than 50%.

6 Conclusions & Future Research

Prior research in fair clustering assumes deterministic knowledge of group membership. We generalized prior work by assuming probabilistic knowledge of group membership. In this new model, we presented novel clustering algorithms in this more general setting with approximation ratio guarantees. We also addressed the problem of "metric membership," where different groups have a notion of order and distance—this addresses real-world use cases where parity must be ensured over, e.g., age or income. We also conducted experiments on slate of datasets. The algorithms we propose come with strong theoretical guarantees; on real-world data, we showed that those guarantees are easily met. Future research directions involve the assignment of *multiple colors* (e.g., race as well as self-reported gender) to vertices, in addition to the removal of assumptions such as the large cluster assumption.

[1] Group 0 is extremely rare, to the point that it violates the "large cluster" assumption for most experiments; therefore, we merged it with Group 1, its nearest age group.

[2] We followed standard procedures and ended up with a standard RBF-based SVM; the accuracy of this SVM is somewhat orthogonal to the message of this paper, and rather serves to illustrate a real-world, noisy labeler.

Broader Impact

Guaranteeing that the color proportions are maintained in each cluster satisfies group (demographic) fairness in clustering. In real-world scenarios, however, group membership may not be known with certainty but rather probabilistically (e.g., learned by way of a machine learning model). Our paper addresses fair clustering in such a scenario and therefore both generalizes that particular (and well-known) problem statement and widens the scope of the application. In settings where a group-fairness-aware clustering algorithm is appropriate to deploy, we believe our work could increase the robustness of those systems. That said, we do note (at least) two broader points of discussion that arise when placing potential applications of our work in the greater context of society:

- We address a specific definition of fairness. While the formalization we address is a common one that draws directly on legal doctrine such as the notion of disparate impact, as expressed by Feldman et al. [2015] and others, we note that the Fairness, Accountability, Transparancy, and Ethics (FATE) in machine learning community has identified *many* such definitions Verma and Rubin [2018]. Yet, there is a growing body of work exploring the gaps between the FATE-style definitions of fairness and those desired in industry (see, e.g., recent work due to Holstein et al. [2019] that interviews developers about their wants and needs in this space), and there is growing evidence that stakeholders may not even comprehend those definitions in the first place Saha et al. [2020]. Indeed, "deciding on a definition of fairness" is an inherently morally-laden, application-specific decision, and we acknowledge that making a prescriptive statement about whether or not our model is *appropriate* for a particular use case is the purview of both technicians, such as ourselves, and policymakers and/or other stakeholders.
- Our work is motivated by the assumption that, in many real-world settings, group membership may not be known deterministically. If group membership is being estimated by a machine-learning-based model, then it is likely that this estimator itself could incorporate bias into the membership estimate; thus, our final clustering could also reflect that bias. As an example, take a bank in the United States; here, it may not be legal for a bank to store information on sensitive attributes—a fact made known recently by the "Apple Card" fiasco of late 2019 Knight [2019]. Thus, to audit algorithms for bias, it may be the case that either the bank or a third-party service infers sensitive attributes from past data, which likely introduces bias into the group membership estimate itself. (See recent work due to Chen et al. [2019] for an in-depth discussion from the point of view of an industry-academic team.)

We have tried to present this work without making normative statements about, e.g., the definition of fairness used; still, we emphasize the importance of open dialog with stakeholders in any system, and acknowledge that our proposed approach serves as one part of a larger application ecosystem.

Acknowledgments

Dickerson and Esmaeili were supported in part by NSF CAREER Award IIS-1846237, DARPA GARD Award #HR112020007, DARPA SI3-CMD Award #S4761, DoD WHS Award #HQ003420F0035, NIH R01 Award NLM-013039-01, and a Google Faculty Research Award. We thank Keegan Hines for discussion about "fairness under unawareness" in practical settings, and for pointers to related literature.

References

Charu C Aggarwal and Chandan K Reddy. Data clustering: Algorithms and applications. 2013.

Gagan Aggarwal, Rina Panigrahy, Tomás Feder, Dilys Thomas, Krishnaram Kenthapadi, Samir Khuller, and An Zhu. Achieving anonymity via clustering. *ACM Transactions on Algorithms (TALG)*, 6(3):49, 2010.

Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. *arXiv preprint arXiv:1905.12753*, 2019a.

Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. *SIAM Journal on Computing*, (0):FOCS17–97, 2019b.

Hyung-Chan An, Aditya Bhaskara, Chandra Chekuri, Shalmoli Gupta, Vivek Madan, and Ola Svensson. Centrality of trees for capacitated k-center. *Mathematical Programming*, 154(1-2): 29–53, 2015.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23: 2016, 2016.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing*, 33(3):544–562, 2004.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Effectiveness of equalized odds for fair classification under imperfect group information. *arXiv preprint arXiv:1906.03284*, 2019.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413, 2019.

Suman K Bera, Deeparnab Chakrabarty, and Maryam Negahbani. Fair algorithms for clustering. *arXiv preprint arXiv:1901.02393*, 2019.

Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*, 2018.

Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 737–756. SIAM, 2014.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 339–348, 2019.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

Marek Cygan, MohammadTaghi Hajiaghayi, and Samir Khuller. Lp rounding for k-centers with non-uniform hard capacities. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 273–282. IEEE, 2012.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, 2015.

Rajiv Gandhi, Samir Khuller, Srinivasan Parthasarathy, and Aravind Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM (JACM)*, 53(3):324–360, 2006.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 1985a. ISSN 0304-3975.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985b.

Aric Hagberg, Dan Schult, Pieter Swart, D Conway, L Séguin-Charbonneau, C Ellison, B Edwards, and J Torrents. Networkx. high productivity software for complex networks. *Webová strá nka https://networkx. lanl. gov/wiki*, 2013.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, May 1985. ISSN 0364-765X.

Dorit S. Hochbaum and David B. Shmoys. A unified approach to approximation algorithms for bottleneck problems. *J. ACM*, May 1986. ISSN 0004-5411.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2019.

Lingxiao Huang, Shaofeng Jiang, and Nisheeth Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems 32*, pages 7587–7598. Curran Associates, Inc., 2019.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 325–333, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.

Samir Khuller and Yoram J Sussmann. The capacitated k-center problem. *SIAM Journal on Discrete Mathematics*, 13(3):403–418, 2000.

Will Knight. The Apple Card didn't 'see' gender—and that's the problem. *Wired*, 2019.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

Heidi Ledford. Millions of black people affected by racial bias in health-care algorithms. *Nature*, 574(7780):608, 2019.

Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 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*, KDD '11, page 502–510, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020488. URL `https://doi.org/10.1145/2020408.2020488`.

IBM CPLEX User's Manual. Version 12 release 7. *IBM ILOG CPLEX Optimization*, 2016.

Christopher Meek, Bo Thiesson, and David Heckerman. The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research*, 2(Feb):397–418, 2002.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L Mazurek, and Michael Carl Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning (ICML)*, 2020.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009.