Towards Fairer Centroids in k-means Clustering

Stanley Simoes[1], Deepak P[1], Muiris MacCarthaigh[2]

[1]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast
[2]School of History, Anthropology, Philosophy and Politics, Queen's University Belfast
ssimoes01@qub.ac.uk, deepaksp@acm.org, M.MacCarthaigh@qub.ac.uk

Abstract

There has been much recent interest in developing fair clustering algorithms that seek to do justice to the representation of groups defined along sensitive attributes such as *race* and *sex*. Within the centroid clustering paradigm, these algorithms are seen to generate clusterings where different groups are disadvantaged within different clusters with respect to their representativity, *i.e.*, distance to centroid. In view of this deficiency, we propose a novel notion of *cluster-level centroid fairness* that targets the representativity unfairness borne by groups within each cluster, along with a metric to quantify the same. Towards operationalising this notion, we draw on ideas from political philosophy aligned with consideration for the worst-off group to develop *Fair-Centroid*; a new clustering method that focusses on enhancing the representativity of the worst-off group within each cluster. Our method uses an iterative optimisation paradigm wherein an initial cluster assignment is refined by reassigning objects to clusters such that the worst-off group in each cluster is benefitted. We compare our notion with a related fairness notion and show through extensive empirical evaluations on real-world datasets that our method significantly enhances cluster-level centroid fairness at low impact on cluster coherence.

Introduction

Fairness in clustering has seen much scholarly activity in recent times (Chhabra, Masalkovaitė, and Mohapatra 2021). Most of these endeavours, starting with Chierichetti et al. (2017), target proportional representation of sensitive groups – such as those defined on *race* and *sex* – within each cluster; this is often referred to as *group fairness* (Dwork et al. 2012). Such representational fairness of groups may be seen as the application of the notion of proportional representation, *aka* statistical parity (Besse et al. 2022), within clustering.

Within the *centroid clustering* paradigm pioneered by classical algorithms such as k-means, only ensuring proportional representation of groups within each cluster may be insufficient depending on how subsequent decisions are made. Where decisions are to be made solely on cluster membership, this notion of group fairness would be appropriate. In centroid clustering however, each cluster is additionally characterised by a *centroid*, making it distinct from other clustering paradigms. Here, a data object's proximity to its cluster centroid is key in determining the clustering quality; an object that is close to its centroid would be better represented within the clustering than another object that is farther away. This additional characteristic opens up a new avenue of unfairness in terms of how well a cluster's centroid represents the objects in that cluster. Such a consideration is critical in decision-making scenarios where centroids are central such as facility location (*e.g.*, polling sites (Chen et al. 2022; Brady and McNulty 2011)) and summarisation.

Towards this, P and Abraham (2020) seek to deepen the uniformity of every object's distance-to-centroid, dubbed *representativity fairness*, regardless of sensitive group membership; this falls within the scope of *individual fairness* (Dwork et al. 2012). Abbasi, Bhaskara, and Venkatasubramanian (2021) and Ghadiri, Samadi, and Vempala (2021) extend this notion to sensitive groups – they consider the mean representativity across objects within each group, and target equity along such aggregates. For example, in a social profile clustering scenario, the mean distance-to-centroid of female profiles should be as close as possible to that of males. This notion, called *fair k-means*, targets overall group fairness, *i.e.*, groups be treated fairly in the clustering regardless of cluster membership. Against this backdrop, we note two issues with fair k-means' approach to representativity aggregations across groups: (i) representativities are not comparable across clusters, and (ii) poor representativity in one cluster could negatively impact other clusters. We examine these using illustrative examples[1].

First, representativity is quantified as an object's distance to its cluster centroid. This construction does not yield well to cross-cluster comparisons when clusters are of different sizes as Figure 1 illustrates. The *purple* object here has better representativity in the large cluster than the *red* one, and vice versa. When conditioned on cluster size, the same difference in representativity is almost insignificant in the large cluster but very consequential in the small one, causing the *red* group to have a better overall representativity than the *purple* group. In sharp contrast to such intuitive judgement, the simple aggregation of representativity misleadingly puts both groups on an equal footing.

[1]The data objects in Figures 1–3 are two dimensional, shown on the xy-plane. Each grey region depicts a cluster.

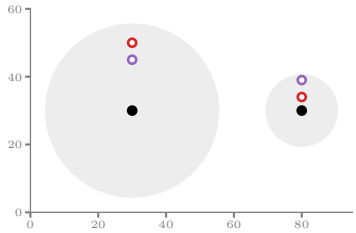


Figure 1: Representativity across differently sized clusters. The *red* and *purple* circles are two objects of distinct groups, and the distance between these two objects is the same in both clusters. Solid black circles depict centroids.

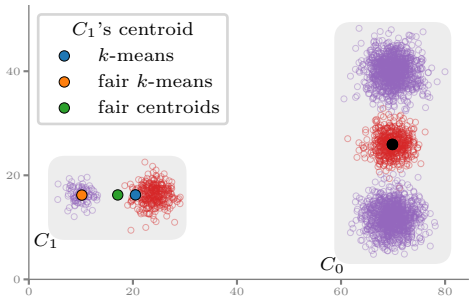
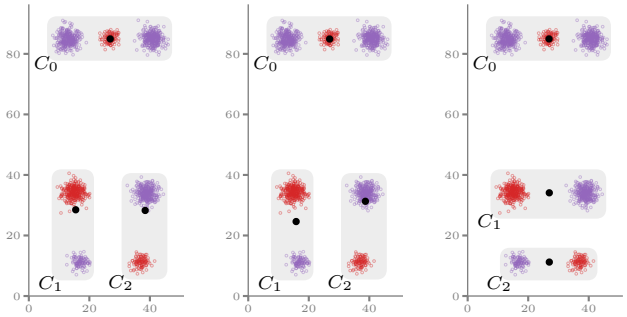


Figure 2: Poor representativity in one cluster (C_0) affecting another's (C_1's) centroid. The two groups are shown in *red* and *purple*. C_0's centroid, in black, is at the best position.

Second, simply aggregating representativities across clusters could cause a group's poor representativity in one cluster to negatively impact another cluster. In Figure 2, *purple*'s poor representativity in C_0 is compensated by placing C_1's centroid (in orange) closer to *purple*, thus improving its representativity but worsening *red*'s representativity in C_1. Yet, *purple* happens to be the most disadvantaged group across the clustering. If C_1 is looked at in isolation, this centroid heavily disadvantages the group that is not the worst-off overall (*red*). This is especially unjustifiable when the two clusters are distant and possibly unrelated, thus begging the question – *is it fair and justified to disadvantage a group in one cluster just because it is advantaged in another cluster?* Instead, a fair centroid for C_1 (in green) would be one that considers each cluster independently, appreciating that clusters could have different worst-off groups. Indeed, having different worst-off groups in different regions in the data is especially prevalent in geographical data where racially segregated regions is common (*e.g.*, Abbasi et al. (2023)). Furthermore, poor representativity in one cluster could also affect cluster membership, as Figure 3 shows. Notice here that fair k-means compensates for *purple*'s poor representativity in C_0 by choosing C_2's centroid close to *purple* (Figure 3b), even closer than k-means (Figure 3a), making fair k-means less fair than k-means for C_2. Apparently, fair k-means' paradigm of cross-cluster aggregation of representativities allows such compensatory effects to occur, thus acting as a veneer to conceal or exacerbate (as is the case in C_2) potentially deep levels of intra-cluster unfairness. A fairer



(a) k-means (b) fair k-means (c) fair centroids

Figure 3: Poor representativity in one cluster (C_0) affecting cluster membership far away. Solid black circles depict centroids. C_0's centroid is unchanged across the three methods.

clustering shown in Figure 3c would target fairness in individual clusters, preventing the unfairness experienced in one cluster from affecting other clusters.

Our Contributions We introduce *cluster-level centroid fairness* (CCF) – a novel formulation of group fairness extending the notion of representativity fairness along sensitive groups in the data. Our notion rectifies the identified cross-cluster representativity aggregation issues of fair k-means, thus generating fairer centroids. In the interest of empirically comparing this notion with fair k-means, we operationalise the former through a novel fair clustering method – *Fair-Centroid*, and illustrate through extensive empirical evaluations that our method is able to achieve high degrees of fairness on appropriate evaluation metrics.

Related Work

The group fairness notion targetting proportional representation of sensitive groups within clusters was pioneered by Chierichetti et al. (2017). Since then, this stream has diversified into considering paradigms such as spectral (Wang et al. 2023) and hierarchical clustering (Knittel et al. 2023). Variants of group fairness notions such as capped representation (Ahmadian et al. 2019) and fair labelled clustering (Esmaeili et al. 2022) have been explored. Orthogonal to these, Hotegni, Mahabadi, and Vakilian (2023) focus on the representation of groups in the set of centroids.

Along the facet of centroid proximity, P and Abraham (2020) optimise for the uniformity of representativity across all objects, while Abbasi, Bhaskara, and Venkatasubramanian (2021) and Ghadiri, Samadi, and Vempala (2021) target equitable overall group representativity. Better approximation algorithms (Makarychev and Vakilian 2021; Goyal and Jaiswal 2023) and generalisations (Chlamtáč, Makarychev, and Vakilian 2022; Gorantla et al. 2023) for the latter have since been designed. Buet-Golfouse and Utyagulov (2022) look at this problem within the framework of fair generalised low-rank models, and Chhabra, Singla, and Mohapatra (2022) proposed a dataset augmentation approach.

Positioning Our Work Our notion of cluster-level centroid fairness is a new conceptualisation of fairness unex-

plored in previous work. Our use of cluster-level group representativity fairness quantification as an intermediate level between individual and dataset-level group fairness makes it distinct from previous work on both fairness streams.

Background

We briefly outline the formulation of the popular k-means clustering problem (MacQueen 1967). Let X be a set of relational data objects to be partitioned into k clusters denoted by $\mathcal{C}$. Lloyd's heuristic for k-means (Lloyd 1982) uses an EM-style framework to optimise for the objective J_U

$$J_U(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{x \in C} \textit{dist}(x, C) \tag{1}$$

where $\textit{dist}(x, C)$ is the squared Euclidean distance of object x to the centroid μ_C of cluster C given by

$$\textit{dist}(x, C) = ||x - \mu_C||^2 \tag{2}$$

Thus, $\textit{dist}(x, C)$ quantifies the representativity of x with respect to C. Notice that the objective relates to a given cluster assignment $\mathcal{C}$; the EM-style optimisation starts with a given cluster assignment, and iteratively refines the cluster assignment and centroids to minimise the objective in Equation 1.

As a generic clustering problem, k-means ignores sensitive group membership – it optimises for the sum of representativities across all objects. This may be regarded as a Benthamite utilitarian objective that seeks *the greatest good for the greatest number* (Bentham 1996). Abbasi, Bhaskara, and Venkatasubramanian (2021) and Ghadiri, Samadi, and Vempala (2021) note that this can result in centroids representing groups differently, often favouring one over others, thus having implications when centroid fairness is central. To mitigate the disparity in representativities of groups, they independently propose the fair k-means notion. Let each object in X belong to one or more of several groups S (*e.g.*, *white*, *asian*, *etc.*) defined across a sensitive attribute $\mathcal{S}$ (*e.g.*, *race*). The fair k-means objective optimises for the overall (dataset-level) worst-off group by minimising

$$\max_{S \in \mathcal{S}} \frac{1}{|S|} \sum_{C \in \mathcal{C}} \sum_{x \in C \cap S} \textit{dist}(x, C) \tag{3}$$

Note that this objective's averaging property allows a group's poor representativity in one cluster to be offset by good representativities in other clusters (as seen for the *red* group in Figures 2 and 3b). Furthermore, the same group may not be the worst-off across all clusters, as this objective assumes. Thus, optimising for the overall worst-off may not augur well for clusters whose worst-off is not the worst-off overall. Towards this, we focus on mitigating group unfairness at the cluster-level rather than at the dataset-level.

Cluster-level Centroid Fairness

Our notion of cluster-level centroid fairness (CCF) targets to minimise the disparity in the representativities of groups within each cluster. We assume a single sensitive attribute $\mathcal{S}$. Consider a cluster C within a cluster assignment $\mathcal{C}$. The representativity of group S within C, denoted by $g(S, C)$, is

$$g(S, C) = \frac{1}{|C \cap S|} \sum_{x \in C \cap S} \textit{dist}(x, C) \tag{4}$$

Thus for each cluster C, we obtain a *representativity vector* with one component per group. *We say that a cluster is fair if its representativity vector is uniform, and a clustering is fair if all clusters in it are fair.* In view of this novel formulation of cluster-level group representativity fairness, we develop a metric towards quantifying adherence to the notion. Accordingly, we capture the representativity disparity between the groups within cluster C simply by taking the variance in their representativities, given by

$$\operatorname*{Var}_{S \in \mathcal{S}} \frac{1}{|C \cap S|} \sum_{x \in C \cap S} \textit{dist}(x, C) \tag{5}$$

Our choice of variance for quantifying group fairness is similar to P and Abraham (2020)'s quantification of individual fairness which is based on egalitarianism, our focus being at the group-level (*e.g.*, gender equality, racial equality). Here, the ideal value 0 is achieved when every group in the cluster has the same representativity. Now, when aggregating across clusters of different sizes, this formulation can allow variances in large clusters to conceal variances in small ones. Towards allowing a fair basis of aggregation, we standardise the representativities of objects in each cluster separately regardless of group membership, *i.e.*, the objects' representativities within a cluster would be transformed to have zero mean and unit variance. With σ referring to this transformation, we thus compute the *cluster disparity* $\delta(C)$ as

$$\delta(C) = \operatorname*{Var}_{S \in \mathcal{S}} \frac{1}{|C \cap S|} \sum_{x \in C \cap S} \sigma\big(\textit{dist}(x, C)\big) \tag{6}$$

To arrive at a single measure for the clustering $\mathcal{C}$, we aggregate this across clusters as

$$\overline{\delta}(\mathcal{C}) = \frac{1}{k} \sum_{C \in \mathcal{C}} \delta(C) \tag{7}$$

where k=$|\mathcal{C}|$ is the number of clusters. The *average cluster disparity* $\overline{\delta}(\mathcal{C})$, being an average of variances, evaluates to a non-negative value, with a lower value indicating a smaller disparity between the groups' representativities in individual clusters, and consequently a fairer clustering. This metric thus quantifies our notion of CCF. Observe that $\overline{\delta}(\mathcal{C})$ does not capture utility in any way, and thus provides no indication on the quality of the clustering.

Considerations Note that being underpinned by the concept of representativity fairness which is inherently centroid dependent, CCF is designed to work with centroid-based clustering algorithms; non-centroid paradigms are thus beyond our scope. Additionally, given a plethora of fairness notions, the one to be applied in a specific context needs to be deliberated and situated within the nuances of the scenario. For example, CCF is highly appealing for facility location in geographically segregated regions where different regions may have different worst-off groups.

Fair-Centroid

In the interest of comparing CCF with other fairness and utilitarian notions, we operationalise CCF through Fair-Centroid, a novel fair clustering method. We describe our objective function followed by the optimisation framework.

Objective Function

Recall that CCF targets to enhance the uniformity of the representativity vector for each cluster. This, as Abbasi, Bhaskara, and Venkatasubramanian (2021) note, can be trivially achieved by having a poor representativity for all sensitive groups thus making the resulting clustering of poor utility. Instead, we take cue from contemporary theories in political philosophy and focus on mitigating the representativity loss (*i.e.*, Equation 4) experienced by *the worst-off group within each cluster*. In contrast to weighted formulations for groups within clusters, this formulation espouses the ethos across several popular philosophical theories including *concern for the most vulnerable* within the famed *difference principle* (Freeman 2018) of distributive justice due to John Rawls[2], and pervades eastern ideals (*e.g.*, Gandhian thought (P and Abraham 2020)). Similar minimax formulations have also been explored for fairness in classification (Martinez, Bertran, and Sapiro 2020), dimensionality reduction (Samadi et al. 2018), and clustering (Equation 3). Thus, our goal here is to generate a coherent clustering (the singular focus of algorithms such as k-means) where additionally, the *representativity of the worst-off group within each cluster* is improved as much as possible. Given that the utilitarian consideration of *cluster coherence* (that classical k-means also targets to optimise) would be in apparent tension with the *cluster-level group representativity fairness* consideration, we look to deepen the latter at as little detriment to the former as possible.

Given our intent of improving the representativity of the worst-off group within each cluster, we model our fairness objective J_F, which targets fairer centroids, as the aggregate of representativities of the worst-off group in each cluster

$$J_F(\mathcal{C}) = \sum_{C \in \mathcal{C}} \max_{S \in \mathcal{S}} g(S, C) \tag{8}$$

This *fair centroid* objective captures the fairness ethos espoused by CCF, albeit using significantly different methodology. We also optimise for the utilitarian k-means objective J_U (Equation 1). Our overall objective function J is thus

$$J(\mathcal{C}) = \frac{1-\lambda}{n} J_U(\mathcal{C}) + \frac{\lambda}{k} J_F(\mathcal{C}) \tag{9}$$

where $n{=}|X|$ is the number of objects to be clustered and $0{\leq}\lambda{\leq}1$ is a hyperparameter that controls the trade-off between utility and fairness. To prevent one objective from dominating, we rescale; J_U being a summation over all objects in the dataset is divided by n, and since J_F being a summation over all clusters is divided by k.

Optimisation Framework

The parameters to be estimated in our objective function (Equation 9) are the cluster assignment $\mathcal{C}$ and the set of centroids $\{\mu_C\}$. Towards this, we follow the same EM-style iterative procedure as in Lloyd's heuristic for k-means (Lloyd 1982) that alternates between (i) estimating the cluster assignment keeping the set of centroids fixed (E-step), and (ii) estimating the set of centroids keeping the cluster assignment stationary (M-step). We now describe the E and M-steps within our optimisation framework.

[2]https://en.wikipedia.org/wiki/John_Rawls

Algorithm 1 EstimateAssignment

```
for all x ∈ X do
    for all C' ∈ C do
        Obtain C' using Equation 10
        if J(C') < J(C) then
            C ← C'
```

E-step: Estimating the Cluster Assignment Given the set of centroids $\{\mu_C\}$, we need to assign objects to clusters that minimise our objective function (Equation 9). Towards this, given the complexity of the objective function, we perturb the existing cluster assignment by making a single pass through all objects and greedily reassigning them to clusters such that the value of the objective function decreases. Thus, if an object x is reassigned from cluster C to cluster C', the new cluster assignment $\mathcal{C}'$ is

$$\mathcal{C}' = \mathcal{C} \setminus \{C, C'\} \cup \{C \setminus \{x\}, C' \cup \{x\}\} \tag{10}$$

Algorithm 1 outlines our greedy approach. Within each E-step, this entails trying out $\mathcal{O}(nk)$ cluster reassignments, $k{-}1$ per object. Note that the change between $J(\mathcal{C}')$ and $J(\mathcal{C})$ where $\mathcal{C}$ and $\mathcal{C}'$ differ in the membership of a single object can be efficiently computed without a full dataset-wide estimation, similar in spirit to what is outlined by Abraham, P, and Sundaram (2020, §4.2.1)[3].

M-step: Estimating the Cluster Centroids Our goal here is to estimate the set of centroids $\{\mu_C\}$ that minimises the objective function in Equation 9 while keeping the cluster assignment $\mathcal{C}$ fixed. We use the gradient descent framework where the intent is to move along the negative gradient. Since the $\max$ operator in Equation 8 is not differentiable, we follow P and Abraham (2020) in using a weighted LogSumExp as a differentiable approximation (Buet-Golfouse and Utyagulov 2022)

$$\max_{y \in Y} f(y) \approx \frac{1}{\phi} \log_e \sum_{y \in Y} \exp\big(\phi \times f(y)\big) \tag{11}$$

where $\phi{\in}\mathbb{R}^+$ is a large enough positive constant that amplifies the significance of the largest number. Substituting this in Equation 8 gives us a differentiable approximation for J_F

$$J_{F\text{-}approx}(\mathcal{C}) = \sum_{C \in \mathcal{C}} \frac{1}{\phi} \log_e \sum_{S \in \mathcal{S}} \exp\big(\phi \times g(S, C)\big) \tag{12}$$

The derivative of J with respect to μ_C is[3]

$$\frac{\partial}{\partial \mu_C} J = \frac{1-\lambda}{n} \frac{\partial}{\partial \mu_C} J_U + \frac{\lambda}{k} \frac{\partial}{\partial \mu_C} J_{F\text{-}approx} \tag{13}$$

where

$$\frac{\partial}{\partial \mu_C} J_U = -2 \times \sum_{x \in C} (x - \mu_C) \tag{14}$$

$$\frac{\partial}{\partial \mu_C} J_{F\text{-}approx} = \frac{\sum\limits_{S \in \mathcal{S}} \Big(w(S, C) \times \frac{-2}{|C \cap S|} \times \sum\limits_{x \in C \cap S} (x - \mu_C) \Big)}{\sum\limits_{S \in \mathcal{S}} w(S, C)} \tag{15}$$

[3]For more details, see the extended version at https://doi.org/10.48550/arXiv.2212.14467

Algorithm 2 EstimateCentroids

repeat
  for all $C \in \mathcal{C}$ **do**
    Compute $\frac{\partial}{\partial \mu_C} J$ using Equations 13–16
    Perform an update for μ_C using Equation 17
until convergence

$$w(S, C) = \exp\big(\phi \times g(S, C)\big) \tag{16}$$

Much like the E-step, regularities in the construction of Equation 13 allow for efficient incremental gradient computation. Equations 13–16 are used to iteratively update the set of cluster centroids $\{\mu_C\}$ within the gradient descent framework as

$$\mu_C \leftarrow \mu_C - \eta \frac{\partial}{\partial \mu_C} J \tag{17}$$

where $\eta \in \mathbb{R}^+$ is the learning rate. The M-step is outlined in Algorithm 2. Convergence is declared once the Frobenius norm of the difference in centroids in successive iterations is within a specified relative tolerance hyperparameter $\tau \in \mathbb{R}^+$, similar to the k-means implementation[4] in scikit-learn (Pedregosa et al. 2011), a popular Python machine learning library. This is complemented by another stopping condition where we break the iterative process if the Frobenius norm has not decreased in the last 10 iterations. To summarise, within each M-step, the cluster centroids are iteratively updated until convergence using the update in Equation 17.

Stopping Condition Our EM procedure stops when at least one of the following holds in successive EM-steps: (i) the cluster assignment is unchanged, or (ii) the Frobenius norm between the centroids in the two steps is within the specified tolerance hyperparameter τ. Additionally, following Ghadiri, Samadi, and Vempala (2021), we halt if the above stopping conditions are not met within 200 iterations.

Empirical Evaluation

Since CCF is a new notion with no associated method, our focus is more on evaluating CCF against utilitarian and fairness notions rather than rigorously evaluating the performance[3] of methods that operationalise these notions. We start by describing the experimental setup, followed by results and analyses[5]. Our implementations[6] are in Python 3, and all experiments were run on the Kelvin2 cluster[7] (2GHz AMD processor and 5GB RAM).

Experimental Setup

Datasets Our datasets (Table 1) are based on the publicly available Adult (Becker and Kohavi 1996) and CreditCard (Yeh 2016) datasets. Both contain people data and include sensitive information such as *race* and *sex* (which are protected from being the basis of discrimination by laws, *e.g.*, UK's Equality Act 2010[8]), making them popular for benchmarking in the algorithmic fairness community (Chhabra, Masalkovaitė, and Mohapatra 2021; Fabris, Silvello, and Susto 2022; Le Quy et al. 2022). We consider `sex` and `race` as the two sensitive attributes for Adult, and `SEX` as the sensitive attribute for CreditCard[3].

dataset/ sensitive attribute	sensitive groups	non-sensitive attributes	objects
Adult/`sex`	2	26	46033
Adult/`race`	5		
CreditCard/`SEX`	2	77	30000

Table 1: Datasets used

Baselines CCF being a novel fairness formulation that has been hitherto unexplored in literature, there are no suitable state-of-the-art baseline clustering methods to compare against. Abbasi, Bhaskara, and Venkatasubramanian (2021) and Ghadiri, Samadi, and Vempala (2021), being based on group representativity fairness, are related but optimised for a fairness objective different from ours. In the interest of comparing CCF with existing utilitarian and fairness notions, we benchmark our Fair-Centroid method against methods that operationalise these notions: Lloyd's heuristic for k-means and Fair-Lloyd (Ghadiri, Samadi, and Vempala 2021) for fair k-means. Since the available implementation of Fair-Lloyd[9] only handles binary sensitive attributes, we use an in-house implementation[3] with gradient descent instead of line search for centroid computation. We do not compare with Abbasi, Bhaskara, and Venkatasubramanian (2021), theirs being more suited for the facility location problem rather than k-means clustering.

Parameter Configuration In all experiments, unless specified, we set λ=0.5 to give equal importance to the utilitarian and fairness objectives in our method. For trends on k, we take values for k in the range 3 to 12 with a step size of 1. For gradient computation of the objective functions, we set ϕ=10^3 and η=10^{-3}. For convergence in Fair-Centroid's M-step, we set τ=10^{-4} similar to scikit-learn's k-means. All numbers reported are averaged over 100 runs with initial centroids estimated using k-means++ (Arthur and Vassilvitskii 2007).

Results

Fairness vs Utility It is widely accepted that increase in fairness almost always causes decrease in utility; it would be of interest to look at the CCF gains obtained due to Fair-Centroid and the corresponding loss in utility. Here, we quantify CCF with our average cluster disparity metric (Equation 7), and utility with the k-means objective[10]. Figures 4 and 5 show trends for the trade-off across different

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[5] For all figures in this section, the y-axes differ for the datasets, and lower values on the y-axes are better.

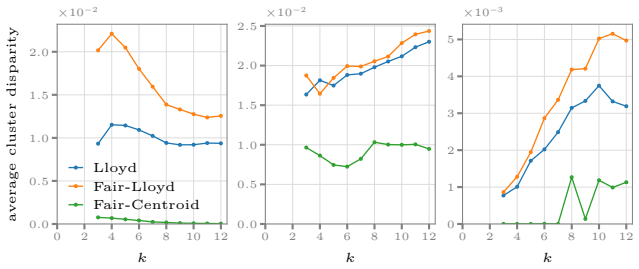
[6] https://github.com/stanleyts/faircentroid

[7] https://www.ni-hpc.ac.uk/Kelvin2/

[8] https://www.gov.uk/discrimination-your-rights

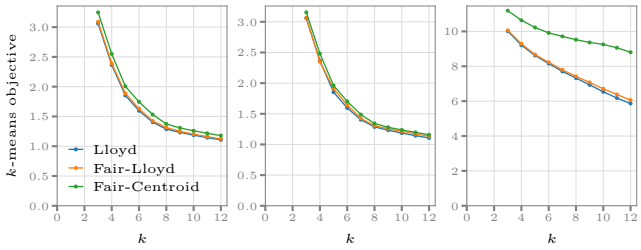
[9] https://github.com/samirasamadi/SociallyFairKMeans

[10] We rescale the k-means objective, *i.e.*, divide Equation 1 by n.



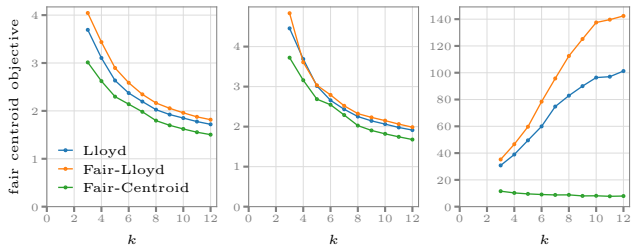
(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

Figure 4: Average cluster disparity $(\overline{\delta})$ across k



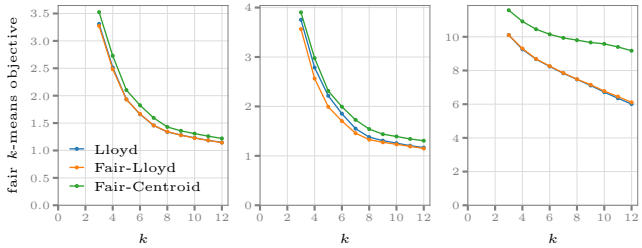
(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

Figure 5: k-means objective (J_U/n) across k



(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

Figure 6: Fair centroid objective (J_F/k) across k



(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

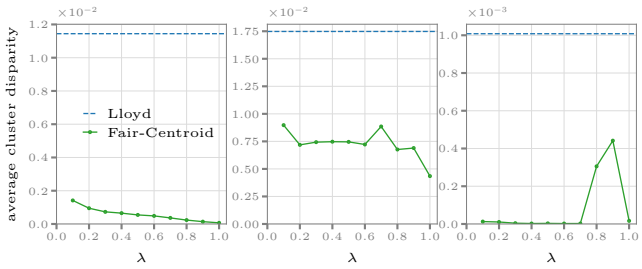
Figure 7: Fair k-means objective across k

values of k. Fair-Centroid consistently and significantly reduces the unfairness over Lloyd (Figure 4) at the cost of a comparably small increase in the utilitarian objective (Figure 5). Fair-Lloyd does not perform as well on this fairness metric, sometimes even worse than Lloyd. Thus, optimising for fair k-means, as Fair-Lloyd targets, provides an impression of overall fairness but conceals unfairness at a deeper level of individual clusters. In contrast, Fair-Centroid beats the baselines on our cluster-level group fairness metric, *i.e.*, average cluster disparity. From Figure 4 we infer that fair k-means, while being conceptually similar to CCF, advances representativity fairness in a way that is empirically antithetical to the consideration of the worst-off group in each cluster. Simply put, Fair-Lloyd (which optimises for fair k-means) nudges the clustering away from Lloyd to configurations in a different direction than Fair-Centroid. Figure 4 thus provides empirical evidence in favour of our motivation that fair k-means may not be considerate to the worst-off group in individual clusters.

Fair Centroid Objective vs Fair *k*-means Towards comparing the two notions of group representativity fairness, we look at how Fair-Centroid compares with the baselines on the two fairness objectives: (i) our fair centroid objective[11] (Equation 8), and (ii) fair k-means objective (Equation 3). Note that the two focus on worst-off groups at different levels – cluster-level vs overall. Figure 6 shows that Fair-Centroid improves the representativity of the worst-off groups over Lloyd, indicating that our method is moving in the right direction. Fair-Lloyd does not perform as well; this is expected as it is not designed for cluster-level group fairness. On the other hand, Fair-Centroid causes the overall worst-off group to have a worse representativity than Lloyd, as seen in Figure 7. While at first glance this may seem to be a deficiency of our method, any improvement in this group's representativity would potentially result in another group that is the worst-off in some cluster being further disadvantaged, as Figures 2 and 3 highlight. This would be unacceptable in cases where more than one group are historically disadvantaged (*e.g.*, Black and American Indians in case of *race*) and benefitting one would result in disadvantaging the other. Evidently, optimising for one fairness objective does not necessarily entail optimising for the other, *i.e.*, reducing overall unfairness could increase cluster-level unfairness, and vice versa.

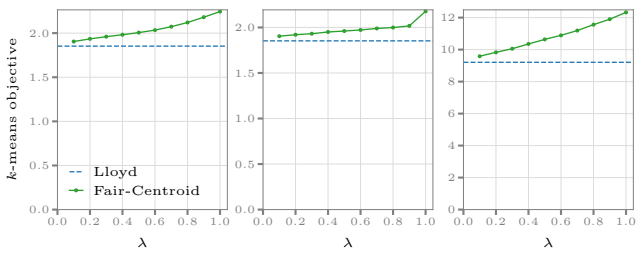
Controlling for Fairness We study how λ – the trade-off between the fairness and utilitarian objectives J_F and J_U in Equation 9 – affects Fair-Centroid; we quantify fairness with average cluster disparity (Equation 7) and utility with the k-means objective[10]. We set k=5 for Adult and k=4 for CreditCard, obtained using the elbow method (Thorndike 1953) on the k-means objective trend for Lloyd (Figure 5). We experiment with values for λ in the range 0.1 to 1 with a step size of 0.1 (λ=0 corresponds to Lloyd's heuristic). Figure 8 shows that even for small λ, optimising for the fair centroid objective in addition to the utilitarian objective (*i.e.*, Equation 9) substantially improves CCF over Lloyd's heuristic. The corresponding loss in utility is comparably small as can be seen in Figure 9 thus indicating the effectiveness of Fair-Centroid in balancing the tradeoff between utility and CCF, making it useful in scenarios where fairness is to be ensured at little impact to utility.

[11] We rescale our objective, *i.e.*, divide Equation 8 by k.



(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

Figure 8: Average cluster disparity ($\overline{\delta}$) across λ



(a) Adult/sex (b) Adult/race (c) CreditCard/SEX

Figure 9: k-means objective (J_U/n) across λ

Disparity Trends To qualitatively analyse Fair-Centroid's behaviour, we look at how the sensitive groups' representativities (Equation 4) vary within clusters over iterations. As an illustration, Figure 10 shows the trends in the clusters on a run on the Adult/race dataset. Notice in the line plots that over iterations: (i) the disparities in the representativities of the best-off and worst-off groups decrease, and (ii) the representativities of the worst-off groups improve, as we set out to achieve. Also observe from the bar plot that both the worst-off groups and the quanta of representativities differ across clusters; these traits are captured by our CCF notion but not by fair k-means. Also note in the bar plot that the worst-off sensitive groups in the clusters change over iterations. While we observe these trends to generally hold across datasets and k, Figure 10 showcase all key points made in this paper.

Computational Comparison Fair-Centroid was found to need more iterations to converge than Lloyd and Fair-Lloyd due to its E-step yielding suboptimal cluster assignments; Lloyd and Fair-Lloyd use a closed-form expression to optimally assign objects to clusters, thus requiring fewer iterations. In terms of runtime, Lloyd was the fastest, followed by Fair-Lloyd which takes longer due to the progressive refinement nature of its M-step, and Fair-Centroid taking the longest as both its E and M-steps use progressive refinement. To give the reader a sense of the runtime, Fair-Centroid with k=5 took ≈80 minutes on the Adult/sex dataset which has ≈46k data objects and 2 sensitive groups[3].

Conclusion and Future Work

We introduced cluster-level centroid fairness (CCF); a new formulation of fairness for centroid clustering with respect

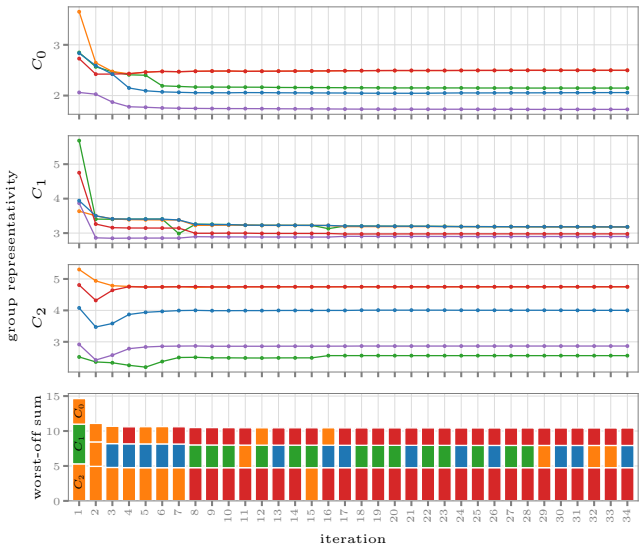


Figure 10: Trends in the 5 groups' representativities for Adult/race (k=3, random state 60) in the 3 clusters due to Fair-Centroid across iterations. Each group is shown in a different colour. The 3 line plots show the trends for all groups in the 3 clusters. The stacked bar plot shows the worst-off group in each cluster and its representativity. The height of each stacked bar is the sum of representativities of the worst-off group in each cluster, *i.e.*, the fair centroid objective.

to representativity of sensitive groups at the cluster-level. Against the backdrop of much recent enthusiasm in using representativity-oriented notions of fairness in clustering, we outlined issues engendered by simplistic aggregations of representativities across clusters, and how our notion alleviates such problems. We proposed Fair-Centroid; a new clustering method that iteratively improves the clustering towards our CCF notion. Given the novel form of our fairness notion, we introduced a new metric that captures disparity between the representativity of groups at the cluster level, accounting for the possibility of different groups being disadvantaged within different clusters. Our experiments on the Adult and CreditCard datasets demonstrate our method's effectiveness in achieving high levels of cluster-level group representativity fairness at low impact to the popular utilitarian cluster coherence metric used within k-means.

Future Work As Fair-Centroid is an initial attempt at operationalising CCF, designing efficient algorithms for CCF is a natural next step, *e.g.*, line search for centroid computation (see Ghadiri, Samadi, and Vempala (2021)), or a LP relaxation (see Abbasi, Bhaskara, and Venkatasubramanian (2021)). Another direction would be to consider numeric sensitive attributes such as *age* – an attribute on which discrimination is well-understood within healthcare (*e.g.*, Dobrowolska et al. (2019)). In terms of clustering paradigms, it would be interesting to apply this work to other non-k-means centroid clustering formulations such as k-medoids (Park and Jun 2009) and fuzzy c-means (Bezdek, Ehrlich, and Full 1984).

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945231; and the Department for the Economy in Northern Ireland. We are grateful for use of the computing resources from the Northern Ireland High Performance Computing (NI-HPC) service funded by EPSRC (EP/T022175).

References

Abbasi, M.; Barrett, C.; Lum, K.; Friedler, S. A.; and Venkatasubramanian, S. 2023. Measuring and Mitigating Voting Access Disparities: A Study of Race and Polling Locations in Florida and North Carolina. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 1038–1048. ACM.

Abbasi, M.; Bhaskara, A.; and Venkatasubramanian, S. 2021. Fair Clustering via Equitable Group Representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 504–514. ACM.

Abraham, S. S.; P, D.; and Sundaram, S. S. 2020. Fairness in Clustering with Multiple Sensitive Attributes. In *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT 2020)*, 287–298. OpenProceedings.org.

Ahmadian, S.; Epasto, A.; Kumar, R.; and Mahdian, M. 2019. Clustering without Over-Representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, 267–275. ACM.

Arthur, D.; and Vassilvitskii, S. 2007. K-Means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, 1027–1035. SIAM.

Becker, B.; and Kohavi, R. 1996. Adult. https://doi.org/10.24432/C5XW20. Accessed: 2024-01-12.

Bentham, J. 1996. *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*. Clarendon Press.

Besse, P.; del Barrio, E.; Gordaliza, P.; Loubes, J.-M.; and Risser, L. 2022. A Survey of Bias in Machine Learning Through the Prism of Statistical Parity. *The American Statistician*, 76(2): 188–198.

Bezdek, J. C.; Ehrlich, R.; and Full, W. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2): 191–203.

Brady, H. E.; and McNulty, J. E. 2011. Turning Out to Vote: The Costs of Finding and Getting to the Polling Place. *American Political Science Review*, 105(1): 115–134.

Buet-Golfouse, F.; and Utyagulov, I. 2022. Towards Fair Unsupervised Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 1399–1409. ACM.

Chen, M. K.; Haggag, K.; Pope, D. G.; and Rohla, R. 2022. Racial Disparities in Voting Wait Times: Evidence from Smartphone Data. *The Review of Economics and Statistics*, 104(6): 1341–1350.

Chhabra, A.; Masalkovaitė, K.; and Mohapatra, P. 2021. An Overview of Fairness in Clustering. *IEEE Access*, 9: 130698–130720.

Chhabra, A.; Singla, A.; and Mohapatra, P. 2022. Fair Clustering Using Antidote Data. In *Proceedings of The Algorithmic Fairness through the Lens of Causality and Robustness*, volume 171 of *Proceedings of Machine Learning Research*, 19–39. PMLR.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2017. Fair Clustering Through Fairlets. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, 5029–5037. Curran Associates, Inc.

Chlamtáč, E.; Makarychev, Y.; and Vakilian, A. 2022. Approximating Fair Clustering with Cascaded Norm Objectives. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2022)*, 2664–2683. SIAM.

Dobrowolska, B.; Jedrzejkiewicz, B.; Pilewska-Kozak, A.; Zarzycka, D.; Ślusarska, B.; Deluga, A.; Kościołek, A.; and Palese, A. 2019. Age discrimination in healthcare institutions perceived by seniors and students. *Nursing Ethics*, 26(2): 443–459.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*, 214–226. ACM.

Esmaeili, S. A.; Duppala, S.; Dickerson, J. P.; and Brubach, B. 2022. Fair Labeled Clustering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '22)*, 327–335. ACM.

Fabris, A.; Silvello, G.; and Susto, G. A. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6): 2074–2152.

Freeman, S. 2018. Rawls on Distributive Justice and the Difference Principle. In *The Oxford Handbook of Distributive Justice*, 13–40. Oxford University Press.

Ghadiri, M.; Samadi, S.; and Vempala, S. 2021. Socially Fair k-Means Clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 438–448. ACM.

Gorantla, S.; Gowda, K. N.; Deshpande, A.; and Louis, A. 2023. Socially Fair Center-Based and Linear Subspace Clustering. In *Machine Learning and Knowledge Discovery in Databases: Research Track (ECML PKDD 2023)*, volume 14169 of *Lecture Notes in Computer Science*, 727–742. Springer Nature Switzerland.

Goyal, D.; and Jaiswal, R. 2023. Tight FPT Approximation for Socially Fair Clustering. *Information Processing Letters*, 182: 106383.

Hotegni, S. S.; Mahabadi, S.; and Vakilian, A. 2023. Approximation Algorithms for Fair Range Clustering. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, 13270–13284. PMLR.

Knittel, M.; Springer, M.; Dickerson, J. P.; and Hajiaghayi, M. 2023. Generalized Reductions: Making any Hierarchical Clustering Fair and Balanced with Low Cost. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, 17218–17242. PMLR.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3): e1452.

Lloyd, S. P. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137.

MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297.

Makarychev, Y.; and Vakilian, A. 2021. Approximation Algorithms for Socially Fair Clustering. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT 2021)*, volume 134 of *Proceedings of Machine Learning Research*, 3246–3264. PMLR.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*, 6755–6764. PMLR.

P, D.; and Abraham, S. S. 2020. Representativity Fairness in Clustering. In *12th ACM Conference on Web Science (WebSci '20)*, 202–211. ACM.

Park, H.-S.; and Jun, C.-H. 2009. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2, Part 2): 3336–3341.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830.

Samadi, S.; Tantipongpipat, U.; Morgenstern, J. H.; Singh, M.; and Vempala, S. 2018. The Price of Fair PCA: One Extra dimension. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, volume 31, 10976–10987. Curran Associates, Inc.

Thorndike, R. L. 1953. Who belongs in the family? *Psychometrika*, 18(4): 267–276.

Wang, J.; Lu, D.; Davidson, I.; and Bai, Z. 2023. Scalable Spectral Clustering with Group Fairness Constraints. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023)*, volume 206 of *Proceedings of Machine Learning Research*, 6613–6629. PMLR.

Yeh, I.-C. 2016. default of credit card clients. https://doi.org/10.24432/C55S3H. Accessed: 2024-01-12.