How Much Does Cluster Fairness Cost? A Counterfactual-based Approach

Antonia Karra[1], Georgios Vardakas[1], Evaggelia Pitoura[1], and Aristidis Likas[1]
{a.karra, g.vardakas, pitoura, arly}@uoi.gr

Department of Computer Science & Engineering, University of Ioannina, Greece

Abstract. Ensuring fairness in clustering algorithms has become a critical concern, particularly when such algorithms are used in high-stakes applications. While several fairness-aware clustering methods have been proposed, a key open question remains: *what is the cost of enforcing fairness in clustering?* In this work, we address this question from both structural and individual perspectives. We introduce two classes of fairness cost measures. First, we define similarity-based costs, which quantify how much a fair clustering diverges from an unfair baseline, measured through Normalized Mutual Information (NMI) and cluster misalignment. Second, we propose a novel counterfactual-based cost, which captures the minimal feature changes needed for individuals to transition from their original to the fair cluster assignment. This counterfactual framework also enables feature-level analysis, revealing which features contribute most to fairness interventions. We apply our approach on four real-world datasets using two fairness criteria, namely balance fairness and social fairness. Our experimental results show that social fairness tends to preserve the original clustering structure better than balance fairness, though it does not always result in lower individual counterfactual costs. Moreover, we uncover implicit biases, with certain features (e.g., marital status) emerging as influential proxies for sensitive attributes.

Keywords: fair clustering · counterfactual explanations · XAI · algorithmic fairness · k-means.

1 Introduction

As Artificial Intelligence (AI) systems are increasingly deployed in sensitive domains such as education, healthcare, and employment, ensuring fairness has become a critical concern. Fair clustering has emerged as an active research area, with several methods proposed to enforce group-level fairness with respect to sensitive attributes [6]. Among these, balance-based approaches seek proportional representation of groups in clusters [7,17,5], while social-based approaches aim for equal cluster quality across groups [9,1,15]. While promising, a key question remains largely unexplored: *What is the cost of these fairness interventions?*

We address this gap by introducing measures to quantify the cost of mitigating clustering unfairness. First, we define similarity-based costs, comparing

the original (unfair) clustering with the fair one using group-level Normalized Mutual Information (NMI) and a misalignment metric that counts data points reassigned across clusters.

However, similarity-based measures do not capture the effort required for an individual to receive a fairer assignment. To address this, we propose a counterfactual-based cost that estimates the minimal changes an individual must undergo to be reassigned to a fair cluster.

Counterfactual explanations, used in classification [23,16,20,14,21], and recently extended to clustering [22,19], describe minimal input changes required to alter the outcome of a model. Given unfair and fair clusterings, we align cluster labels and generate a counterfactual x' for each misaligned individual x, such that x' would receive the fair label. The cost is defined as the distance between x and x' and is computed using a closed-form expression [22]. An additional advantage of this approach is that it enables *feature-level insights*, revealing which attributes contribute most to fairness adjustments and exposing potential proxy variables for sensitive features.

Leveraging our cost definitions, we experimentally address the following research questions:

RQ1: Is social fairness more costly to achieve than balance fairness?

RQ2: What is the relationship between unfairness, similarity-based costs, and counterfactual-based cost?

RQ3: What features contribute most to the cost?

Our results show that social fairness tends to incur lower similarity-based costs than balance fairness. However, neither objective consistently yields lower counterfactual costs. Interestingly, in some cases, the cost of achieving fairness is higher for the group that was not originally disadvantaged than for the group that was. Furthermore, different features contribute to the cost across groups, highlighting asymmetries in how fairness is attained.

The remainder of this paper is structured as follows. In Section 2, we introduce our similarity-based and in Section 3 our counterfactual-based cost formulations. Section 4 reports experimental results addressing our research questions. In Section 5, we discuss related work, and in Section 6, we conclude.

2 Similarity-based Fair Clustering Costs

Let X be a set of n points in $\mathbb{R}^d$. A k-clustering $\mathcal{C}$ of X is a partition of X into k disjoint subsets, $C_1, \ldots, C_k$, called clusters. In this paper, we focus on k-means clustering, although our cost metrics can be defined for other clustering methods as well. We denote the corresponding set of cluster centers by $\{c_1, \ldots, c_k\}$.

Several notions of fairness in clustering have been proposed [6]. While our framework is general, we focus on two widely studied types: (a) *balance fairness*, and (b) *social fairness*. In the following, we formally define these fairness types and then introduce our first class of measures that quantify unfairness of a given (unfair) clustering based on its similarity to a fair solution.

2.1 Fair Clustering

Without loss of generality, let us assume that X is partitioned into two disjoint groups: the blue group B and the red group R, such that $B \cup R = X$ and $B \cap R = \emptyset$. For example, R may represent a protected group defined by a sensitive attribute, such as gender or race. Balance-based fairness requires that each cluster preserves, as closely as possible, the overall group proportions present in the input data [7,5].

Formally, let $C_i \in \mathcal{C}$ be a cluster and $Y \subseteq X$ be a non-empty subset of points. We denote by $Y(C_i)$ the set of points in $Y \cap C_i$. We also define $\rho_Y = \frac{|Y|}{|X|}$, i.e., the representation of Y in the input and $\rho_Y(C_i) = \frac{|Y(C_i)|}{|C_i|}$, i.e., the representation of Y in cluster U_i.

The *balance* of a cluster $C_i \in \mathcal{C}$ is defined as:

$$\text{balance}(C_i) = \min_{Y \in \{R,B\}} \left(\frac{\rho_Y}{\rho_Y(C_i)}, \frac{\rho_Y(C_i)}{\rho_Y} \right). \tag{1}$$

A clustering in which all groups are proportionally represented within every cluster has a balance value of 1, indicating perfect balance. *Balance-based* fair clustering typically tries to achieve this by maximizing the *balance* of the clustering $\mathcal{C}$ defined as the minimum per-cluster balance across all clusters:

$$\text{balance}(\mathcal{C}) = \min_{C_i \in \mathcal{C}} \text{balance}(C_i). \tag{2}$$

Social-based fair clustering follows a different approach: instead of focusing on the group representation inside the clusters, it requires that the clustering cost (e.g. clustering error, quality or utility) is equal across different groups [9]. For k-means clustering, the clustering cost for a non-empty set of points $Y \subseteq X$ is defined as the sum of the squared L_2 distances between each point in Y and the center of its assigned cluster:

$$\Delta(\mathcal{C}, Y) = \sum_{i=1}^{k} \sum_{x \in Y(C_i)} \|x - c_i\|^2. \tag{3}$$

The *fair* k-means objective for two groups B, R is the largest average cost:

$$\Phi(\mathcal{C}) = \max \left\{ \frac{\Delta(\mathcal{C}, B)}{|B|}, \frac{\Delta(\mathcal{C}, R)}{|R|} \right\}. \tag{4}$$

The goal of *social-based* fair clustering is to minimize $\Phi(\mathcal{C})$, i.e., to minimize the highest average cost per group. By doing so, the difference between the cluster cost for the two groups is reduced.

2.2 Similarity-based Costs

Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a clustering of the dataset X, and let $\mathcal{C}^F = \{C_1^F, \ldots, C_k^F\}$ denote the fair clustering produced by applying a fairness-aware clustering algorithm (e.g., balance-based or social-based) to X.

Our first definition of the fairness cost evaluates the difference between the two clusterings based on *Normalized Mutual Information (NMI)*. NMI is a standard metric used to evaluate the similarity between two clusterings by measuring the amount of shared information, normalized to ensure the score lies between 0 (no mutual information) and 1 (perfect alignment). We adapt NMI to capture clustering similarity over specific subsets of points $\emptyset \neq Y \subseteq X$.

Formally, we define:

$$\mathrm{NMI}(\mathcal{C}, \mathcal{C}^F, Y) = \frac{2 \cdot I(\mathcal{C}, \mathcal{C}^F, Y)}{H(\mathcal{C}, Y) + H(\mathcal{C}^F, Y)} \tag{5}$$

where:

$$I(\mathcal{C}, \mathcal{C}^F, Y) = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{|Y(C_i) \cap Y(C_j^F)|}{|Y|} \log_2 \left(\frac{|Y| \cdot |Y(C_i) \cap Y(C_j^F)|}{|Y(C_i)| \cdot |Y(C_j^F)|} \right) \tag{6}$$

$$H(\mathcal{U}, Y) = -\sum_{i=1}^{k} \frac{|Y(U_i)|}{|Y|} \log_2 \left(\frac{|Y(U_i)|}{|Y|} \right), \quad \mathcal{U} \in \{\mathcal{C}, \mathcal{C}^F\} \tag{7}$$

We now present our first measure for fairness evaluation based on the NMI between the unfair $\mathcal{C}$ and fair $\mathcal{C}^{\mathcal{F}}$ partitions.

Definition 1 (NMI-cost). *Let $\mathcal{C}$ and $\mathcal{C}^F$ be respectively a given clustering and a fair clustering of a set of points X. The NMI-cost$(\mathcal{C}, \mathcal{C}^F, Y)$ for a set of points $\emptyset \neq Y \subseteq X$ is defined as $1 - \mathrm{NMI}(\mathcal{C}, \mathcal{C}^F, Y)$.*

To further measure the cost of unfairness, we look at specific points in X. Since clustering is an unsupervised task and cluster labels are assigned arbitrarily, to enable a meaningful comparison between the results of unfair and fair clustering, we apply a cluster alignment procedure that maps the clusters in the unfair clustering $\mathcal{C}$ to the clusters in the fair clustering $\mathcal{C}^F$.

The first step computes a similarity matrix between the clusters of the unfair and fair clustering using the Jaccard similarity:

$$\mathrm{Jaccard}(C_i, C_j^F) = \frac{|C_i \cap C_j^F|}{|C_i \cup C_j^F|},$$

which measures the degree of overlap between each pair of clusters (C_i, C_j^F), with $C_i \in \mathcal{C}$, $C_j^F \in \mathcal{C}^F$, and $1 \leq i, j \leq k$. The second step converts the similarity matrix into a cost matrix via

$$\mathrm{Cost}(i, j) = 1 - \mathrm{Jaccard}(C_i, C_j^F)$$

and solves the resulting *assignment* problem using the *Hungarian algorithm.*

Let $\mathcal{C}^{AF} = \{C_1^{AF}, \ldots, C_k^{AF}\}$ denote the fair clustering after alignment, where each cluster $C_i^{AF} \in \mathcal{C}^{AF}$ is mapped to the corresponding cluster $C_i \in \mathcal{C}$, for $1 \leq i \leq k$.

Let a point $x \in X$ that belongs to C_i in $\mathcal{C}$ and to C_j^{AF} in $\mathcal{C}^{AF}$. We say that x is *misaligned* if $j \neq i$. That is, a point is misaligned if it is assigned different cluster labels in the unfair and the (aligned) fair clustering.

We use the number of misaligned points as an additional similarity-based measure of the unfairness cost.

Definition 2 (Misalignment-cost). *Let $\mathcal{C}$ and $\mathcal{C}^F$ be respectively a clustering and a fair clustering of a set of points X. The misalignment cost mis-cost($\mathcal{C}$, $\mathcal{C}^F$, Y) for a set of points $\emptyset \neq Y \subseteq X$ is equal to the number of points in Y that are misaligned.*

While the NMI-cost offers a *global*, label-invariant measure of structural similarity, the misalignment-cost provides finer-grained insight into which *individual data points* are affected.

Also note that parameterizing our cost definitions with a subset Y of points offers flexibility in their application. When similarity-based costs are applied to the entire dataset (i.e., $Y = X$), they provide an aggregate measure of the overall cost of achieving cluster fairness. When applied separately to individual groups (e.g., $Y = B$ and $Y = R$), they offer a group-wise characterization of the cost. In this case, additional biases may be revealed. For instance, while the clustering may be unfair for one group, the misalignment cost could be higher for the other group, highlighting potential asymmetries in how the fairness cost is distributed between the groups.

3 Counterfactual-based Fair Clustering Cost

Similarity-based costs evaluate how closely the unfair and fair clusterings align, but fail to capture the extent or nature of the interventions needed to achieve fairness. To address this limitation, we turn to counterfactual explanations. Counterfactual explanations have been widely used in classification tasks to explain decisions by identifying the minimal changes required to alter the output of the model [23,16,20,14,21]. Very recently, they have also been explored in the context of clustering [22,19].

In the following, we briefly define counterfactuals for clustering and describe the approach we use to generate them [22]. We then introduce our method for leveraging counterfactuals to quantify the cost of cluster fairness and explain the type of actions needed to achieve fair clustering.

3.1 Counterfactuals for Clustering

Let a clustering solution $\mathcal{C} = \{C_1, \ldots, C_k\}$ with k clusters. Counterfactual explanations enhance interpretability by identifying the minimal changes to the features of a data point x that would result in altering the cluster assignment of x. Concretely, given a point x in cluster C_i and a distance function d, its counterfactual explanation x' for k-means is [22]:

$$x' = \arg\min_{y \in \mathbb{R}^d} d(x, y), \;\; s.t.\; d(y, c_\ell) < d(y, c_i), 1 \leq \ell \leq k. \tag{8}$$

To generate the counterfactuals, we follow the analytical approach proposed in [22] for k-means clustering and the L_2 distance. Given two clusters C_i and C_ℓ with centers c_i and c_ℓ respectively, the cluster boundary set S consists of all points z such that:

$$|z - c_i|^2 = |z - c_\ell|^2. \tag{9}$$

Eq. 9 defines a *hyperplane* that is perpendicular at the middle point ($m_{i\ell} = (c_i + c_\ell)/2$) to the vector $n_{i\ell} = c_\ell - c_i$ connecting the two cluster centers. We refer to this cluster boundary as *separating hyperplane*.

Assuming a point x (factual) of cluster C_i, its closest counterfactual point x'_ℓ with respect to cluster C_ℓ is computed analytically and corresponds to the projection of x to the separating hyperplane. That is:

$$x'_\ell = x - \frac{(x - m_{i\ell})^\top n_{i\ell}}{|n_{i\ell}|^2} n_{i\ell} \tag{10}$$

Moreover, the L_2 distance between x in C_i and its counterfactual x'_ℓ in C_ℓ is given by the *projection distance*:

$$d_{C_i \to C_\ell}(x) = \frac{|(x - m_{il})^\top (c_\ell - c_i)|}{\|c_\ell - c_i\|}. \tag{11}$$

The distance $d_{C_i \to C_l}(x)$ for a data point x mapped to cluster C_i measures the minimum changes in the input features of x that are needed so as x is mapped to C_ℓ instead of C_i.

3.2 Counterfactual-based Cost

Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a given clustering, C^F a fair clustering, and $\mathcal{C}^{AF} = \{C_1^{AF}, \ldots, C_k^{AF}\}$ the aligned fair clustering.

To quantify the cost of repairing the given unfair clustering so that the clustering becomes fair, *we focus on the misaligned data points*. Let x in X be a misaligned data point mapped to C_i in the initial unfair clustering and to C_j^{AF}, $j \neq i$, in the fair clustering.

We define the counterfactual cost, $cfcost(\mathcal{C}, \mathcal{C}^{\mathcal{F}}, x)$ for a *misaligned point* x as:

$$cfcost(\mathcal{C}, \mathcal{C}^{\mathcal{F}}, x) = d_{C_i \to C_j}(x). \tag{12}$$

We define the counterfactual cost, cfcost$(\mathcal{C}, \mathcal{C}^F, Y)$, for a non-empty set of points $Y \subseteq X$, as the *average counterfactual cost* of the misaligned points in Y.

Definition 3 (counterfactual-cost). *Let $\mathcal{C}$ and $\mathcal{C}^F$ be respectively a clustering and a fair clustering of a set of points X. The counterfactual cost, $cfcost(\mathcal{C}, \mathcal{C}^F, Y)$ for a set of points $\emptyset \neq Y \subseteq X$ is defined as: $cfcost(\mathcal{C}, \mathcal{C}^F, Y) = \frac{\sum_{x \in Y'} cfcost(\mathcal{C}, \mathcal{C}^F, x)}{|Y'|}$, where $Y' \subseteq Y$ is the set of misaligned points that belong to Y.*

As opposed to the similarity-based costs that rely on quantifying the difference between the fair and unfair clustering, the counterfactual-based cost accounts also for the *magnitude* of feature updates needed to achieve fairness. Setting $Y = X$ provides an estimation of the total average cost for mitigating unfairness for a data point in X. Restricting Y to the specific B and R groups offers insights for the unfairness mitigation cost per group.

Furthermore, counterfactuals can be used to understand the contribution that individual features have to unfairness. This is achieved by measuring the relative contribution of an individual feature to the counterfactual cost. Let $x = (x_1, \dots x_d)$ be a factual instance and $x' = (x'_1, \dots x'_d)$ its corresponding counterfactual. We define the relative contribution, $r_m(x)$ of feature m as

$$r_m(x) = \frac{(x_m - x'_m)^2}{\sum_{l=1}^{d}(x_l - x'_l)^2} \tag{13}$$

Then the average contribution $r_m(Y)$ of feature m for a set of points $\emptyset \neq Y \subseteq X$ is:

$$r_m(Y) = \frac{\sum_{x \in Y'} r_m(x)}{|Y'|}, \tag{14}$$

where $Y' \subseteq Y$ is the set of misaligned points in Y.

By setting $Y = X$, we attain a global measure of how much each feature contributes, on average, to unfairness. Restricting X to B and R allows us to estimate the contribution of the feature to unfairness for each particular group.

4 Experimental Evaluation

In this section, we use our cost metrics to get insights regarding the cost of achieving fairness in clustering.

4.1 Datasets and Setup

We conducted our experiments using four datasets: Adult[1], Credit Card[2], Bank[3], and Student[4]. For the Adult, Credit Card, and Bank datasets, we randomly sampled 1,000 data points. For the Student dataset, we used the full available 600 instances. The Adult dataset contains demographic and income-related information, with 14 features; the sensitive attribute is sex. The Credit Card dataset includes financial and behavioral attributes of clients (23 features), with MARRIAGE used as the sensitive feature. The Bank dataset contains marketing and demographic data related to a bank campaigns (16 features), with marital status as the sensitive attribute. The Student dataset includes academic and personal

[1] `https://archive.ics.uci.edu/ml/datasets/adult`
[2] `https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients`
[3] `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
[4] `https://archive.ics.uci.edu/dataset/320/student+performance`

background information of students (30 features), using sex as the sensitive attribute. Some statistics for each dataset are summarized in Table 1.

In both approaches, the k-means algorithm is initially applied. Following previous research, we do not consider the sensitive attribute in clustering. We compare the standard k-means algorithm with two fairness-aware approaches: a *balance-based fairness* method, which ensures proportional group representation across clusters [5], and a *social-based fairness* method, implemented via a fairness-aware extension of Lloyd's algorithm (Fair Lloyd), which aims to equalize the clustering cost across sensitive groups [9]. After obtaining the baseline and the fair clusterings, we apply a cluster alignment procedure to enable one-to-one comparison of cluster assignments. Based on this alignment, we identify misaligned points, i.e., data instances assigned to different clusters before and after fairness adjustment. For each misaligned point, we then generate a counterfactual explanation, representing the minimal change in feature space required to achieve reassignment to the target cluster under the fair model.

Table 1: Dataset statistics.

Adult		Student		Bank		Credit Card	
Female	Male	Female	Male	Married	Single	Married	Single
327	673	354	246	708	292	447	553

All results are averaged over 10 runs with different initial cluster centers for each value of k.

Our code is available online[5].

4.2 Experimental Results

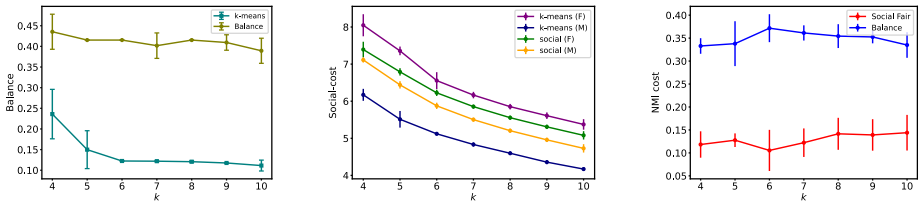
We first evaluate the impact on fairness of applying the fairness-aware clustering algorithms. The balance-based algorithm consistently improves the balance fairness metric (Eq. (2)) for all values of k as shown in Figures 1a, 2a, 3a, and 4a. This metric captures how closely the group proportions within each cluster match their distribution in the overall dataset. A higher balance score reflects better proportionality in the representation of the groups. The consistent increase in balance fairness across datasets demonstrates that balance-based fairness interventions improve group representation, even when disparities in representation persist.

The social cost objective (Eq. (4)) reveals that the standard k-means algorithm introduces disparities in clustering outcomes between demographic groups as shown in Figures 1b, 2b, 3b, and 4b. In the Student and Bank datasets, the social fairness algorithm successfully reduces these disparities, resulting in nearly equal social costs across groups for all values of k. However, this effect is not consistent across all datasets. In the Adult dataset, a noticeable gap remains at

[5] `https://github.com/antwniak/cost_fair_clustering_cf`

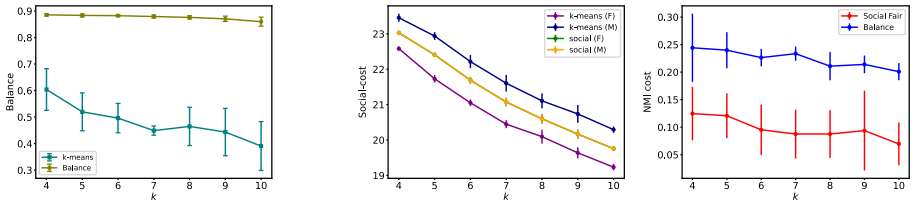
$k = 7$, and in the Credit Card dataset, the approach fails to equalize social costs for most k values. Overall, the effectiveness of the social fairness method varies depending on the dataset and k, highlighting its sensitivity to context.

We now address our research questions.



(a) Balance fairness (b) Social fairness (c) NMI cost

Fig. 1: Adult dataset: Fairness in original and fair k-means for (a) balance and (b) social fairness, (c) the corresponding NMI cost.

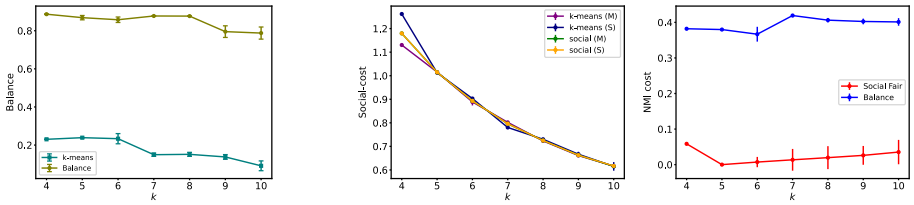


(a) Balance fairness (b) Social fairness (c) NMI cost

Fig. 2: Student dataset: Fairness in original and fair k-means for (a) balance and (b) social fairness, (c) the corresponding NMI cost.

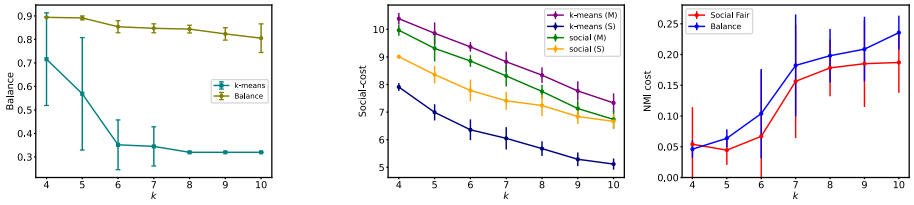
Is balance fairness more costly than social fairness? To evaluate whether balance fairness is more expensive than social fairness, we first compare their cost using the Normalized Mutual Information (NMI) over the whole dataset. As shown in Figures 1c, 2c, 3c, 4c across all four datasets, the social fairness method consistently yields a lower NMI-cost compared to the balance fairness method. This indicates that, although fairness adjustments are applied, the social fairness approach introduces fewer changes to the cluster structure, preserving a closer alignment with the original clustering.

We now look at per group costs. For both per group NMI costs (Figures 5a, 6a, 7a, 8a) and per group misalignment costs (Figure 5b, 6b, 7b, 8b), we observe a similar trend: the social fairness model generally results in lower group



(a) Balance fairness (b) Social fairness (c) NMI cost

Fig. 3: Bank dataset: Fairness in original and fair k-means for (a) balance and (b) social fairness, (c) the corresponding NMI cost.



(a) Balance fairness (b) Social fairness (c) NMI cost

Fig. 4: Credit dataset: Fairness in original and fair k-means for (a) balance and (b) social fairness, (c) the corresponding NMI cost.

NMI and reassignment costs, reinforcing the observation that it introduces fewer structural modifications than balance fairness. Note that the misaligned cost for the whole dataset is just the sum of the misaligned costs for the two groups.

Although the social fairness method typically achieves lower misaligned costs compared to the balance-based method, it induces higher average counterfactual costs (Figures 5c, 6c, 7c, 8c). Across all datasets, the counterfactual cost for both sensitive groups is generally higher under the social method than under the balance method. Counterfactual cost captures how much the features of an individual must change to reach a fairer assignment, and higher values indicate greater deviation from the original instance. This suggests that while social fairness provides better alignment between unfair and fair clusterings, it may lead to counterfactuals that are less realistic or less feasible, revealing a trade-off between alignment quality and counterfactual proximity.

What is the relationship between unfairness, similarity-based costs, and counterfactual-based cost? First, we observe that the two similarity-based costs, namely, the NMI cost and the misalignment cost, shown in 5(a-b),6(a-b), 7(a-b), 8(a-b) per group, tend to follow a similar trend: when the number of misaligned points increases, the NMI cost also rises. This is expected, as both metrics capture structural changes in the clustering assignment caused by fairness interventions. The counterfactual-based cost, depicted in Figures 5c,

6c, and 7c, 8c, reflects the average amount of change required in the features of an individual to achieve a fairer clustering assignment; higher values indicate a greater departure from the original data point. This metric is more sensitive to individual-level changes, since it reflects *how far each individual misaligned point* must move in feature space to satisfy fairness.

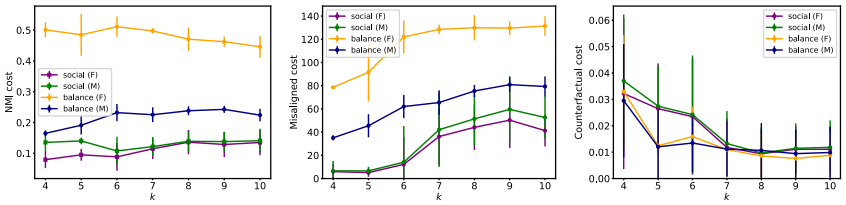
In terms of the relationship between fairness and costs, interestingly, there are cases where although there is unfairness towards one group, the cost to mitigate unfairness is higher for the other group. For example, a higher social cost for one group does not necessarily imply greater similarity-based, and/or counterfactual-based costs. For instance, in the Adult dataset (Figure 1b), we observe that the Female group incurs a higher social cost than the Male group, but both the similarity-based costs (Figures 5a, and 5b) and the counterfactual based costs (Figure 5c) are larger for the Male group than for the Female group, for certain values of k.

Which features contribute to the cost per group? We now examine which features play the most significant role in achieving fairer cluster assignments. Since the sensitive attribute is not used in clustering, this analysis offers insights into implicit bias. Specifically, *explicit bias* refers to decision-making processes that directly consider the protected attribute (e.g., race, gender), whereas *implicit bias* refers to the influence of features correlated with the protected attribute [18].

For this experiment, we report results for the Adult dataset using $k = 5$ clusters, selected based on the optimal value determined by the Silhouette score. We compute the average contribution of each feature to the counterfactual cost Eq.(14). This analysis is performed separately for the two fairness frameworks explored in this study: balance fairness and social fairness.

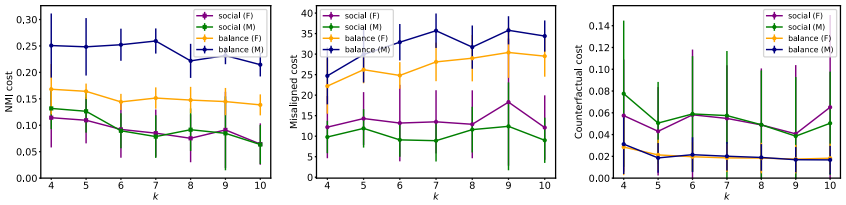
As shown in Figures 9a–9d, the feature marital status emerges as the most influential across all scenarios. Under the balance fairness objective, marital status contributes 35.8% to counterfactual changes for females and 53.81% for males. Similarly, under the social fairness objective, its contribution rises even further—55.34% for females and 48.69% for males. Notably, given that the social cost objective is higher for females in this setting, the greater contribution of marital status suggests that the model relies more heavily on this feature to shift women toward fairer cluster assignments. These findings indicate that, even in the absence of the sensitive attribute, the model relies heavily on marital status as a proxy, particularly when generating fairer outcomes.

Figures 10a–10d show the feature-wise changes applied to each individual during counterfactual generation. Points to the right of the vertical line represent feature increases, while points to the left indicate decreases. These plots confirm that the model alters different features with varying frequency and magnitude depending on the fairness objective and gender, offering a more fine-grained view that complements the aggregate contribution analysis.



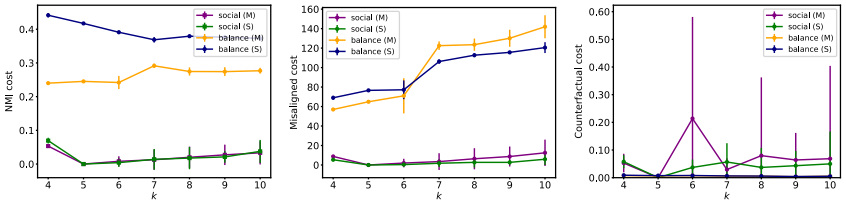
(a) NMI cost (b) Misaligned cost (c) Counterfactual cost

Fig. 5: Adult dataset:(a) NMI cost per sensitive group, (b) misaligned and (c) counterfactual cost.



(a) NMI cost (b) Misaligned cost (c) Counterfactual cost

Fig. 6: Student dataset: (a) NMI cost per sensitive group (b) misaligned and (c) counterfactual cost.

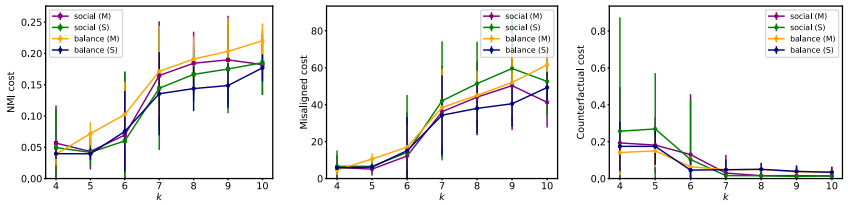


(a) NMI cost (b) Misaligned cost (c) Counterfactual cost

Fig. 7: Bank dataset: (a) NMI cost per sensitive group (b) misaligned and (c) counterfactual cost.

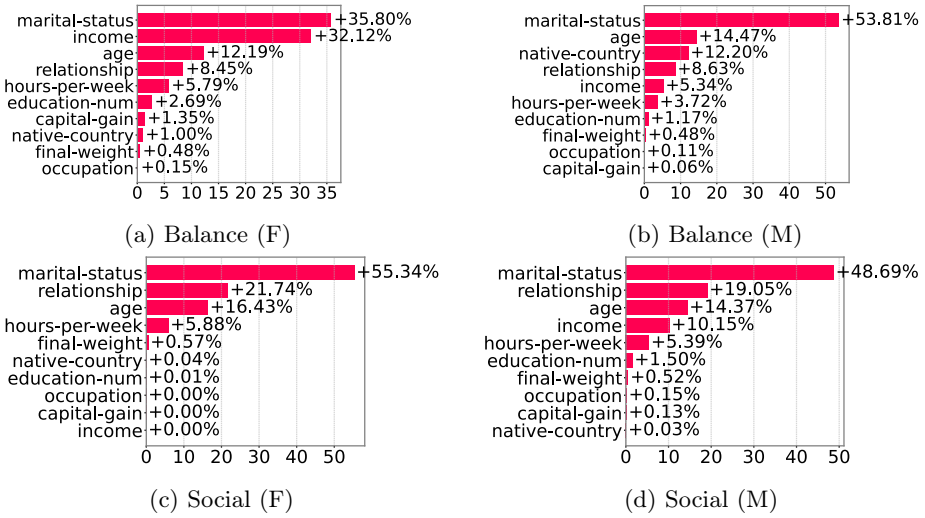
5 Related Work

Fairness in clustering takes different forms depending on the context and application needs. *Balance-based fairness* refers to the requirement that the proportion of individuals belonging to a protected group within each cluster should reflect their proportion in the entire dataset. This ensures that no group is disproportionately over- or under-represented in any given cluster. An important work in fair clustering is that of Chierichetti et al. [7], where the authors introduce the concept of balance and provide approximation algorithms that incorporate this fairness notion for k-center and k-median clustering. Subsequently, other works



(a) NMI cost (b) Misaligned cost (c) Counterfactual cost.

Fig. 8: Credit dataset: (a) NMI cost per sensitive group (b) misaligned and (c) counterfactual cost.



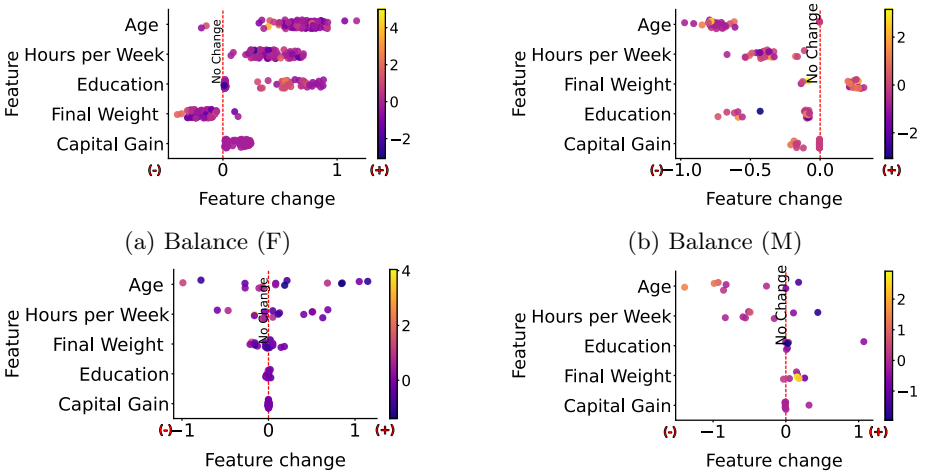
(a) Balance (F) (b) Balance (M)

(c) Social (F) (d) Social (M)

Fig. 9: Adult dataset: Average feature contribution to counterfactual cost for both fairness objectives.

also emerged which extended the original one in terms of the number of sensitive attributes examined and the number of values they can take, such as [17,5,2] as well in terms of the different clustering algorithms like corelation clustering [3], spectral clustering [12], modularity [10], or hierarchical clustering [8].

Social fairness in clustering aims to minimize disparities in clustering cost across demographic groups. Ghadiri et al. [9] introduce equitable group representation, formulating fairness as the minimization of the maximum average clustering cost across groups, and propose bicriteria and approximation algorithms to address this objective. Abbasi et al [1] introduce a fairness notion based on group representativeness, and propose bicriteria approximation algorithms for k-median and facility location problems, offering theoretical guarantees that are especially effective when group sizes are imbalanced. Makarychev et al. [15] extend this line of works by developing alternative approximation algorithms under



(a) Balance (F)

(b) Balance (M)

(c) Social (F)

(d) Social (M)

Fig. 10: Adult dataset: Feature changes for both fairness objectives.

the same fairness criterion, with improved formulations and tighter bounds in certain settings.

Counterfactual explanations have been primarily applied in the context of supervised learning models, particularly classification models [11,4]. Various model types have been explored for this purpose, including gradient-based models [23,16], tree-based ensembles [20,14]. In contrast, their application in unsupervised learning, such as clustering, has received limited attention. For clustering, previous work either adapts algorithms originally developed for counterfactuals in classification tasks [19] and just one work proposes an algorithm specifically designed for the clustering setting [22].

Counterfactual explanations can also be used to evaluate the fairness of model decisions. Some works [18,13] introduce the notion of burden, a more interpretable form of group fairness. The burden reflects how difficult it is for individuals or groups to obtain recourse—i.e., to change their input features in order to receive a favorable outcome. Fairness, in this context, means that this difficulty should be comparable across different sensitive groups.

6 Conclusions

In this work, we provide a comprehensive evaluation of fairness-aware clustering by comparing two key objectives: balance fairness and social cost fairness. Our analysis demonstrates that the social fairness objective leads to fewer disruptions in the clustering structure, as indicated by lower Normalized Mutual Information (NMI) and misalignment costs, which measure the similarity between fair and unfair clusterings and the number of cluster reassignments, respectively. To

assess fairness at the individual level, we leverage counterfactual explanations, using the counterfactual cost to quantify the magnitude of changes required for an individual to reach a fairer cluster assignment. Interestingly, we find that social cost fairness, while preserving the global structure better, often incurs a higher counterfactual cost compared to balance fairness — implying that individuals must undergo more substantial changes to achieve fairer outcomes under this objective.

Future directions of this research include extending the analysis to different clustering algorithms, such as deep clustering and Gaussian mixture models, and applying the framework to a broader range of datasets. Additionally, we aim to explore alternative definitions of fairness and evaluate their effects on clustering behavior and counterfactual explanations. Another important direction involves experimenting with different methods for generating counterfactuals, as well as investigating how the selection of features, i.e., which features are allowed or restricted to change, affects the fairness and interpretability of the resulting counterfactuals.

Acknowledgment

The research project is implemented in the framework of H.F.R.I. call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union - NextGenerationEU (H.F.R.I. ProjectNumber: 15940).

References

1. Abbasi, M., Bhaskara, A., Venkatasubramanian, S.: Fair clustering via equitable group representations. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 504–514 (2021)
2. Abraham, S.S., Sundaram, S.S., et al.: Fairness in clustering with multiple sensitive attributes. arXiv preprint arXiv:1910.05113 (2019)
3. Ahmadi, S., Galhotra, S., Saha, B., Schwartz, R.: Fair correlation clustering. arXiv preprint arXiv:2002.03508 (2020)
4. Artelt, A., Hammer, B.: On the computation of counterfactual explanations–a survey. arXiv preprint arXiv:1911.07749 (2019)
5. Bera, S., Chakrabarty, D., Flores, N., Negahbani, M.: Fair algorithms for clustering. Advances in Neural Information Processing Systems **32** (2019)
6. Chhabra, A., Masalkovaitė, K., Mohapatra, P.: An overview of fairness in clustering. IEEE Access **9**, 130698–130720 (2021)
7. Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. Advances in neural information processing systems **30** (2017)
8. Fogliato, R., G'Sell, M., Chouldechova, A.: Fair hierarchical clustering. In: Conference on Neural Information Processing Systems (2021)
9. Ghadiri, M., Samadi, S., Vempala, S.: Socially fair k-means clustering. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 438–448 (2021)

10. Gkartzios, C., Pitoura, E., Tsaparas, P.: Fair network communities through group modularity. In: Proceedings of the ACM on Web Conference 2025, WWW 2025. pp. 506–517. ACM (2025)
11. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery pp. 1–55 (2022)
12. Kleindessner, M., Samadi, S., Awasthi, P., Morgenstern, J.: Guarantees for spectral clustering with fairness constraints. In: International conference on machine learning. pp. 3458–3467. PMLR (2019)
13. Kuratomi, A., Pitoura, E., Papapetrou, P., Lindgren, T., Tsaparas, P.: Measuring the burden of (un)fairness using counterfactuals. In: ECML/PKDD Workshops (1). Communications in Computer and Information Science, vol. 1752, pp. 402–417. Springer (2022)
14. Lucic, A., Oosterhuis, H., Haned, H., de Rijke, M.: Focus: Flexible optimizable counterfactual explanations for tree ensembles. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 5313–5322 (2022)
15. Makarychev, Y., Vakilian, A.: Approximation algorithms for socially fair clustering. In: Conference on Learning Theory. pp. 3246–3264. PMLR (2021)
16. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
17. Schmidt, M., Schwiegelshohn, C., Sohler, C.: Fair coresets and streaming algorithms for fair k-means clustering. arXiv preprint arXiv:1812.10854 (2018)
18. Sharma, S., Henderson, J., Ghosh, J.: Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. arXiv preprint arXiv:1905.07857 (2019)
19. Spagnol, A., Sokol, K., Barbiero, P., Langheinrich, M., Gjoreski, M.: Counterfactual explanations for clustering models. arXiv preprint arXiv:2409.12632 (2024)
20. Tolomei, G., Silvestri, F., Haines, A., Lalmas, M.: Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 465–474 (2017)
21. Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 650–665. Springer (2021)
22. Vardakas, G., Karra, A., Pitoura, E., Likas, A.: Counterfactual explanations for k-means and gaussian clustering (2025), `https://arxiv.org/abs/2501.10234`
23. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)