H.F.R.I call "Basic Research Funding (Horizontal Support for all Sciences)"
National Recovery and Resilience Plan
(Greece 2.0)

Project Name: Counterfactuals for Clustering: Explainability, Fairness and Quality

Acronym: FairXCluster

Project No: 15940

Deliverable D2.1: State-of-the-art on Fair Clustering

The research project is implemented in the framework of H.F.R.I call "Basic Research Funding (Horizontal Support for all Sciences)" under the National Recovery and Resilience Plan "Greece2.0", funded by the European Union – NextGenerationEU (H.F.R.I project Number: 15940)

State-of-the-art on Fair Clustering

Antonia Karra and Evaggelia Pitoura

Department of Computer Science & Engineering
University of Ioannina, Greece

1 Introduction

Machine learning has become an integral part of modern decision-making systems, shaping various aspects of our daily lives, from healthcare [16] and finance [24] to hiring processes [28], [31] and college admissions [34], [18], [6]. As these models influence critical decisions, concerns about their fairness and potential biases have gained significant attention. Bias in machine learning can emerge from multiple sources, including biased training data, algorithmic design, and societal inequalities, leading to unfair or discriminatory outcomes.

In clustering, fairness is particularly challenging since unsupervised learning does not rely on predefined labels, making it difficult to assess and mitigate biases directly. Unfair clustering results can reinforce social disparities, marginalizing certain groups or misrepresenting structural patterns in the data. To address these challenges, researchers have explored various notions of fairness in clustering, including balance constraints, socially fair clustering, individually fair clustering, and fairness in deep clustering.

This survey provides a comprehensive overview of fair clustering approaches, categorizing existing work into key fairness notions and highlighting recent advancements and open challenges.

2 Fairness in Clustering

Ensuring fairness in machine learning models can be approached at three distinct points in the learning pipeline [27], [10]: before training, during the training process, or after training has been completed.

The pre-training phase involves preprocessing the dataset before it is used for learning, aiming to correct biases at the data level. This ensures that when the model is trained on the adjusted dataset without further modifications, its predictions align with fairness requirements [14, 9, 8, 33]. The in-training phase [21, 36, 11], which is the most widely used approach, involves embedding fairness considerations directly into the model during its learning process. This is achieved by modifying optimization objectives or constraints to produce fairer

outputs without altering the original dataset. Lastly, fairness can also be introduced post-training [9, 4, 20], where the model's predictions are adjusted through post-processing techniques to better align with fairness requirements.

Based on the work of Chhabra et al. [12], the definitions of fairness have been categorized into two different classifications: group-level, individual-level fairness. Group-level fairness ensures that no group of individuals is disproportionately favored or disadvantaged by a clustering algorithm. It originates from the Disparate Impact (DI) doctrine, which states that protected groups (e.g., based on race, gender, or other attributes) should be fairly represented in each cluster relative to their proportion in the overall dataset.

Individual-level fairness ensures that similar individuals are treated similarly by the clustering algorithm. Unlike group-level fairness, it does not rely on predefined protected groups but rather on a similarity metric that quantifies the closeness between individuals. A clustering algorithm satisfies individual fairness if each instance is closer, on average, to members of its assigned cluster than to members of any other cluster. This guarantees that individuals who are similar according to the defined metric are assigned to the same or similar clusters

Preliminaries

Let X be a set of n points in a metric space (X, d), where $d : X \times X \rightarrow \mathbb{R}$ is a distance function, X_i the subset of X belonging to protected group i, $\mathcal{N}$ stands for the set of attributes that are non-sensitive, S for the set of sensitive attributes and let $\mathcal{C}$ represent the clustering procedure. Specifically, $\mathcal{C}$ partitions the dataset into k groups (clusters), denoted as $C=\{C_1, C_2, \ldots, C_k\}$. The center of each cluster C_i is denoted as c_i. Each cluster is formed such that the data objects within the same cluster are similar to one another and (potentially) dissimilar from objects in other clusters. In addition, some works that examine fairness with binary sensitive attributes use color coding as red or blue. Therefore, we define S_r and S_b as the sets of points that are either red or blue, respectively.

Types of Fairness in Clustering

Fairness in clustering takes different forms depending on the context and application needs. Various definitions aim to ensure equitable treatment, either at the group level, or at the individual level.

- **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.
- **Social fairness** focuses on equalizing clustering costs across different groups to avoid systemic disadvantages.

- **Individual fairness** ensures that similar individuals are treated alike, regardless of group membership.
- **Deep Fair Clustering** addresses biases in deep clustering methods by enforcing fairness constraints at both representation and clustering stages.

In the following sections, we analyze key studies in each category highlighting their contributions to fair clustering.

3 Balance-based Fairness

A very important work in fair clustering is that of Chierichetti et al. [14], where the authors introduce the concept of *balance*, which ensures that clusters maintain a similar fraction of protected groups. In their framework, each data point is assigned one of two colors, red or blue, representing the protected group to which it belongs. The balance of a set S is formally defined as:

$$\text{balance}(S) = \min\left(\frac{|S_r|}{|S_b|}, \frac{|S_b|}{|S_r|}\right) \in [0, 1] \tag{1}$$

where $|S_r|$ and $|S_b|$ represent the number of red and blue points, respectively. A clustering $\mathcal{C}$ is (r, b)**-fair** if every cluster $C_i \in \mathcal{C}$ satisfies $\text{balance}(C_i) \geq \frac{b}{r}$, and the overall balance of clustering $\mathcal{C}$ is given by:

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

A perfectly balanced cluster has an equal number of red and blue points (balance = 1), whereas a completely monochromatic cluster has a balance of 0. To simplify the inherently difficult problem of fair clustering, the authors propose a two-step approach that first partitions the dataset into *fairlets* —small groups that preserve the balance ratio of the protected groups—and then applies traditional clustering methods to these fairlets. The process of finding fairlets involves formulating the problem as a *minimum cost flow (MCF) optimization task* over a directed bipartite graph, where nodes represent data points and edges connect individuals from different protected groups, weighted by their distances in the feature space. The goal is to find a perfect matching or an approximate partition that minimizes the total clustering cost while satisfying fairness constraints. This ensures that fairlets are constructed efficiently while maintaining a specified level of balance. Once the fairlets are formed, each is assigned a representative center, and a standard clustering algorithm such as *k-median* or *k-center* is applied to the set of fairlet centers instead of the original data points. This entire process is referred to as a (b, r)-*fairlet decomposition*, and its quality is evaluated based on the minimized k-median or k-center cost, where:

$$\text{k-median cost} = \sum_{x \in X} d(x, \beta(x)) \tag{3}$$

$$\text{k-center cost} = \max_{x \in X} d(x, \beta(x)) \tag{4}$$

with $\beta(x)$ denoting the index of the fairlet to which a point x is assigned. A (b, r)-fairlet decomposition is considered optimal if it minimizes these costs among all possible decompositions. Since achieving perfect balance is often infeasible due to real-world dataset constraints, the authors introduce the concept of approximate balance, allowing a controlled level of imbalance. This is governed by a parameter t, where $t \leq 1$, meaning that smaller values of t permit greater imbalance in protected group representation. For example, if $t = 0.8$, a cluster may contain up to 80% of one group and 20% of the other, thus enabling fair clustering in datasets with unequal group sizes.

In Schmidt et al. [29], the authors extend the definition of fairness of previous work to accommodate sensitive attributes with multiple potential values. More specifically, they introduce *fair coresets*, which are small, weighted subsets of the dataset that preserve fairness constraints and provide an efficient approximation for the k-means objective. Additionally, they represent the sensitive attribute as a color and, to handle multiple sensitive attributes, assign a unique color to each combination of their possible values. To scale their approach to large datasets, the authors propose streaming algorithms that process data sequentially, updating the coreset dynamically while preserving fairness constraints. When a new data point is added, the algorithm ensures that the proportional representation of sensitive groups is maintained by carefully adjusting the weights of points in the coreset. These updates are designed to prevent overrepresentation or underrepresentation of any sensitive group as the dataset grows. The authors provide theoretical guarantees that their methods achieve near-optimal k-means costs while maintaining a fair clustering structure. To enforce fairness in clustering, they define a constraint that ensures each cluster contains a balanced representation of sensitive attributes relative to the entire dataset. Given a dataset with a sensitive attribute having multiple values, fairness is ensured by requiring that for each cluster C_i and each sensitive group j, the fraction of group j within the cluster remains within a factor of its overall proportion in the dataset:

$$\alpha \cdot \xi(j) \leq \frac{|\{p \in C_i : c(p) = j\}|}{|C_i|} \leq \beta \cdot \xi(j), \quad \forall C_i, j \tag{5}$$

where $\xi(j)$ is the fraction of data points in the entire dataset belonging to group j, and α, β are fairness parameters controlling the deviation allowed from this proportion. This formulation ensures that no sensitive group is disproportionately over- or underrepresented in any cluster, thereby maintaining fairness while preserving clustering quality. To explicitly enforce these fairness constraints, the authors introduce a *coloring constraint matrix* K, which encodes the exact number of points of each sensitive group assigned to each cluster. The matrix K has dimensions $k \times \ell$, where k is the number of clusters and ℓ is the number of sensitive groups (colors). Each entry $K_{i,j}$ specifies the number of points from group j that must be assigned to cluster C_i. This structured

approach transforms the fairness requirement into a constrained optimization problem. More formally, the fairness constraint can be rewritten in terms of K:

$$\alpha \cdot \xi(j) \leq \frac{|\{p \in C_i : c(p) = j\}|}{|C_i|} \tag{6}$$

$$= \frac{K_{i,j}}{\sum_{h=1}^{\ell} K_{i,h}} \leq \beta \cdot \xi(j), \quad \forall i \in \{1, \ldots, k\}, j \in \{1, \ldots, \ell\}. \tag{7}$$

Despite these efforts, existing fairness notions were insufficient for clustering problems due to the limitations of standard coreset definitions. If fairness constraints are enforced only through proportionality, then combining multiple core-sets may lead to suboptimal or unfair solutions. More precisely, given two point sets P_1 and P_2 with their corresponding coresets S_1 and S_2, replacing individual points by representative points in a naive coreset construction introduces an error of $O(\epsilon\Delta)$. While this error is small for each coreset individually, when the point sets are combined, the optimal cost drops significantly, whereas the cost of the coreset remains large ($\Omega(\epsilon\Delta)$). This discrepancy violates the core property required for composable coresets. To address this, the authors introduce a stricter fairness definition based on the *coloring constraint matrix* K and define the *color-k-means cost*, which ensures that fairness constraints are preserved even when coresets are combined. The *color-k-means cost*, denoted as $\text{colcost}(P, K, C)$, represents the minimum clustering cost while satisfying the fairness constraints encoded in K:

$$\text{colcost}(P, K, C) = \min_{\substack{C_1,\ldots,C_k \\ \text{s.t. } |\{p \in C_i : c(p) = j\}| = K_{i,j} \; \forall i,j}} \sum_{i=1}^{k} \sum_{p \in C_i} \|p - c_i\|^2. \tag{8}$$

By incorporating the coloring constraint matrix into the optimization process, the proposed method guarantees a *composable and scalable* solution for fair k-means clustering while maintaining both fairness and clustering efficiency.

In Backurs et al. [8], the authors drew inspiration from Chierichetti et al. [14], who developed polynomial-time algorithms to incorporate fairness into traditional clustering methods, such as k-center and k-median. In this work, the authors introduce a more efficient algorithm with *nearly linear* runtime, enabling the use of fairlet decomposition on larger data. Their focus is on the k-median formulation, which is less sensitive to outliers compared to k-center, while ensuring balanced and fair clustering under predefined fairness constraints. The proposed algorithm consists of two steps. In the first step, the algorithm maps the input points into a γ-HST (Hierarchically Separated Tree), a hierarchical structure that facilitates efficient clustering. Each node v in the tree is annotated with S_r and S_b, representing the number of red and blue points in the subtree rooted at v. The goal is to partition the dataset into (r, b)-fairlets that maintain a balanced representation of protected groups. To compute the total cost of fairlet decomposition, the authors use the k-median cost function, defined as $\text{cost}_{\text{median}}(S) = \min_{p \in S} \sum_{q \in S} d(p, q)$, where $d(p, q)$

represents the distance between points p and q in the tree. The objective is to minimize this clustering cost while ensuring fairness constraints are met. Once fairlets are constructed, they apply a standard k-median algorithm to cluster them.The next phase of the their approach is the same with the original work, it follows the same procedure of merging (r, b)-fairlets into k-clusters by selecting a center for each fairlet, applying a β-approximate k-median algorithm, and ensuring that all points in a fairlet are assigned to the cluster of their respective center.

Building on the clustering fairness framework introduced by Chierichetti et al. [14], Bera et al. [9], extends the approach by incorporating overlapping protected groups, thereby addressing real-world fairness challenges more effectively. Also, unlike previous works that imposed rigid fairness constraints, this paper allows user-defined fairness parameters. These parameters define the minimum and maximum representation of protected groups in clusters. More specifically, the authors propose a fairness model defined by two key properties: **Restricted Dominance** (RD), which limits the maximum proportion of any group i in a cluster S_j by enforcing the constraint $\frac{|X_i \cap S_j|}{|S_j|} \leq \beta_i$, and **Minority Protection** (MP), which ensures a minimum representation for underrepresented groups by requiring that $\frac{|X_i \cap S_j|}{|S_j|} \geq \alpha_i$, where X_i is the subset of data points in X that belong to the protected group i. Using parameters β_i and α_i, these constraints can vary across different groups, offering flexibility. They also introduce a parameter Δ to control the extent of group overlap, allowing individuals to belong to multiple groups. To enforce fairness, the authors present a black-box transformation that modifies any standard clustering algorithm to satisfy the RD and MP constraints while maintaining near-optimal clustering quality. Specifically, given a metric space (X, d), a set of potential cluster centers $C \subseteq X$, and an integer k, the objective is to select a subset $S \subseteq F$ of at most k cluster centers and assign each point $v \in X$ to one of these centers using a function $\phi : X \to S$. The goal is to find an assignment $\phi : X \to S$ in order to satisfy the fairness constraints **RD** and **MP** and minimize the clustering cost $L_p(S; \phi)$, the distance between points and their assigned centers, while ensuring compliance with the RD and MP fairness constraints. The authors prove that their algorithm achieves a $(\rho + 2)$-approximation to the best fair clustering solution while allowing a small additive violation in fairness constraints.

Abraham et al. [2] tackles the challenge of fair clustering in the context of datasets containing multiple sensitive attributes. The authors propose a new clustering framework and algorithm designed to incorporate group fairness across multiple sensitive attributes. These attributes can be numerical, binary, or categorical, broadening the applicability of fair clustering techniques to more complex datasets. More specifically, they present the **FairKM** technique, which constructs an objective function to satisfy fairness constraints on the sensitive attributes. The FairKM objective function integrates two key components to

balance clustering quality and fairness. The objective function O is defined as:

$$O = \underbrace{\sum_{C \in \mathcal{C}} \sum_{x \in C_i} \text{dist}_{\mathcal{N}}(x, c_i)}_{\text{K-Means Term over attributes in } \mathcal{N}} + \lambda \underbrace{\text{deviation}_{\mathcal{S}}(C, \mathcal{X})}_{\text{Fairness Term over attributes in } \mathcal{S}} \tag{9}$$

The first term, **K-Means Loss Term**, measures the distance between each data point x in a cluster C_i and the cluster center c_i, considering only a subset of attributes $\mathcal{N}$. The second term, **Fairness Loss Term**, ensures that the clustering remains fair with respect to the sensitive attributes $\mathcal{S}$. The function $deviation_S(C, X)$ penalizes deviations from a fair representation of sensitive attributes within each cluster. The parameter λ controls the trade-off between clustering quality and fairness. The goal is to minimize O to achieve fair clustering while maintaining meaningful clusters. If $\lambda = 0$, the function reduces to standard K-Means clustering (without fairness considerations), whereas higher λ values prioritize fairness over pure clustering compactness. To optimize this objective function, FairKM follows an iterative **round-robin assignment update approach**, balancing clustering quality and fairness constraints. The process begins with an initial clustering, where centroids and the distribution of sensitive attributes in each cluster are computed. Each data point x is then evaluated for reassignment by minimizing the total objective function, which consists of two competing components: the K-Means term, which ensures intra-cluster similarity, and the Fairness term, which enforces balanced representation of sensitive attributes. The algorithm evaluates the impact of moving x to each possible cluster and selects the assignment that results in the lowest objective function value. When updating the cluster assignments, FairKM computes the change in the objective function:

$$\delta O = \delta(\text{K-Means term}) + \lambda \cdot \delta(\text{Fairness term}) \tag{10}$$

where the first term captures the effect on clustering compactness, and the second term adjusts for fairness deviations. The **round-robin** approach ensures that each point is updated sequentially, allowing fairness constraints to be gradually incorporated. After reassigning a point, the cluster centroids are updated to reflect the new memberships, and fairness metrics are recomputed based on the new cluster compositions. The process is repeated until either the clustering assignments stabilize or a predefined number of iterations is reached.

More specifically, the **FairKM** algorithm extends the classical K-Means clustering approach to include fairness constraints, following an iterative process: it begins by randomly initializing k clusters and computing initial centroids. For each data point X, the cluster assignment is updated by minimizing the combined loss, which includes both clustering quality and fairness terms. After updating the assignments, the cluster centroids are recalculated, and fairness constraints are adjusted accordingly. This cycle continues until convergence, ensuring that the final clusters are both compact and fair in terms of sensitive attribute representation. By integrating fairness directly into the optimization

process, FairKM offers a principled approach to mitigating bias in clustering while maintaining meaningful structure in the data.

In the paper of Kleindessner et al. [21] the integration of fairness into spectral clustering is explored. Unlike Chierichetti et al. [14], which guarantees fairness at the cost of clustering quality, this paper introduces a spectral clustering approach that aims to balance fairness and clustering quality, only achieving fairness if it doesn't significantly degrade the objective value. Spectral clustering is a popular graph-based method for partitioning data into clusters by leveraging the structure of the graph's similarity matrix. The unnormalized spectral clustering algorithm minimizes the RatioCut objective, which balances the separation of clusters with the similarity within them. It works by computing the graph's Laplacian matrix, finding its smallest eigenvalues and eigenvectors, and applying k-means clustering to project the graph into a lower-dimensional space. While this method focuses on optimizing clustering quality, its original design does not account for fairness constraints. The unnormalized spectral clustering algorithm can be extended to incorporate fairness constraints by ensuring proportional representation of protected groups within each cluster. This is achieved by introducing a group-membership matrix F. Each column of F indicates whether a vertex belongs to a specific protected group. The next step is modifying the Laplacian matrix to respect fairness constraints via projection into the null space of the membership matrix. The algorithm computes a new embedding that incorporates these constraints and applies k-means clustering to partition the data. This extension allows spectral clustering to balance fairness and clustering quality, providing a framework to handle equity concerns in graph-based data partitioning.

This work of Ahmadian et al. [5] extends the concept of fairness in clustering to hierarchical clustering, where the goal is to construct a tree structure that optimizes a specific objective (e.g., revenue, value, or cost) while ensuring fairness. The proposed algorithm for fair hierarchical clustering builds on the concept of fairlet decomposition. The algorithm employs a local search method, inspired by Chierichetti et al. [14], to optimize the cost of forming these fairlets while maintaining fairness constraints. The work also draws inspiration from Ahmadian et al. [4], which introduced the concept of preventing over-representation in clustering through fairness constraints, particularly in the *a-capped k-center problem*. This idea is extended to hierarchical clustering by ensuring that at every level of the hierarchy, no group dominates, much like the bounded representation parameter in flat clustering. Once the fairlets are constructed, the hierarchical clustering tree is built by clustering these fairlets using average-linkage clustering, ensuring that fairness is maintained at all levels of the hierarchy. The method is designed to optimize one of three objectives—revenue, value, or cost—where revenue focuses on maximizing similarity within clusters, value minimizes dissimilarity, and cost reduces the total clustering expense. The authors demonstrate that fairness can be integrated into hierarchical clustering with negligible impact on the quality of the results.

The paper Ahmadi et al. [3] tackles the challenge of ensuring fairness in correlation clustering, a method that clusters objects based on similarity and

dissimilarity relationships without fixing the number of clusters in advance. The authors propose a fairlet-based reduction technique that converts a fair correlation clustering problem into a standard one using a graph transformation, enabling solutions that achieve fairness and reduce cluster imbalance effectively. First, the graph is decomposed into fairlets, small subsets that locally satisfy fairness constraints, minimizing the cost of clustering within and between fairlets. The fairlet decomposition cost combines two components: **Internal cost** (FCOST^{in}) which is the cost of edges within each fairlet, penalizing negative edges and **External cost** ($\text{FCOST}^{\text{out}}$) which is the cost of edges between fairlets, penalizing positive edges. The goal is to minimize the total fairlet decomposition cost. The authors use approximation techniques to construct small fairlets efficiently. Once fairlets are created, a reduced graph is constructed and each fairlet becomes a single vertex in the graph. Then, they apply an existing correlation clustering algorithm to minimize the cost of clustering on the reduced graph. The final step is the assignment of each fairlet to the same cluster as its corresponding vertex in the reduced graph.

4 Social Fairness

While traditional fairness notions in clustering, such as balance, focus on ensuring proportional representation of different demographic groups, they do not account for disparities in the clustering cost experienced by these groups. In many real-world applications, simply ensuring that each group has an equal presence in clusters does not prevent certain groups from consistently being placed in higher-cost clusters, leading to unfair treatment. Social fairness addresses this issue by ensuring that the clustering process distributes costs equitably among all demographic groups, preventing any single group from bearing a disproportionate burden.

The research of Ghadiri et al. [15] addresses fairness issues in traditional k-means clustering, particularly when applied to datasets containing diverse demographic groups. Standard k-means aims to minimize the sum of squared distances for all data points, but this approach can lead to disproportionately high clustering costs for certain subgroups. To mitigate this, **Fair-k-means** introduces a fairness-aware objective that minimizes the **maximum average clustering cost** across all demographic groups. For m demographic groups, where $X = A_1 \cup \cdots \cup A_m$, the objective function is defined as:

$$\Phi(U, \mathcal{C}) = \max\left(\frac{\Delta(U, \mathcal{C} \cap A_1)}{|A_1|}, \frac{\Delta(U, \mathcal{C} \cap A_2)}{|A_2|}, \ldots, \frac{\Delta(U, \mathcal{C} \cap A_m)}{|A_m|}\right) \tag{11}$$

where U represents the set of cluster centers, $\mathcal{C}$ is the partitioning of the dataset into clusters, and A_i for $i \in \{1, 2, \ldots, m\}$ represents different demographic groups. The term $\mathcal{C} \cap A_i$ denotes the subset of clustered points belonging to group A_i, while $\Delta(U, \mathcal{C} \cap A_i)$ is the **total clustering cost** for group A_i, calculated as the sum of squared distances of its members to their assigned cluster

centers. Finally, $|A_i|$ is the number of data points in group A_i. To efficiently determine fair cluster centers, the method performs a line search along the optimal center path, ensuring that cluster placement balances fairness while maintaining effective clustering. Adjusting the cluster centers accordingly, **Fair-k-means** ensures that differences in clustering cost between demographic groups are minimized, leading to a more equitable clustering outcome. This work is the first to study clustering fairness from the perspective of demographic subgroups, proposing a novel fairness criterion that goes beyond previous methods, which primarily focused on proportional representation.

In the paper of Abbasi et al. [1], the authors highlight equitable group representation as the central notion of fairness in clustering. Unlike traditional fairness constraints that focus solely on demographic balance, equitable representation ensures that the clustering outcome reflects the actual distribution and characteristics of the groups within the data. This approach goes beyond merely balancing cluster sizes, aiming instead for cluster centers that are representative of each group's distribution. To formalize fairness, the authors introduce two key error metrics: Absolute Representation Error (AbsError) and Relative Representation Error (RelError). AbsError measures the total clustering cost for each group, ensuring that distances between a group's points and their assigned cluster centers are minimized and balanced across groups. It is defined as:

$$\text{AbsError}_C(X) = \sum_{x \in X} d(x, C) \tag{12}$$

where C is the set of cluster centers, X is the set of all data points, and $d(x, C)$ represents the distance from data point x to its nearest cluster center in C. Meanwhile, RelError normalizes clustering cost relative to the best possible clustering cost for each group, ensuring that groups with naturally higher intrinsic clustering costs are not disproportionately penalized. It is given by:

$$\text{RelError}_C(A_i) = \frac{\sum_{x \in A_i} d(x, C)}{\sum_{x \in A_i} d(x, \text{Opt}(A_i))} \tag{13}$$

where $\text{Opt}(A_i)$ represents the optimal placement of cluster centers if only group A_i were clustered.This distinction between absolute and relative errors is crucial, as the two objectives may be incompatible—minimizing one does not necessarily minimize the other. To enforce fairness, the authors formulate an optimization problem that ensures equitable representation by minimizing the maximum average clustering cost across groups. The fair clustering problem is expressed as:

$$\min_{C} \max_{i \in \{1,\ldots,m\}} \left(\frac{\sum_{x \in A_i} d(x, C)}{|A_i|} \right) \tag{14}$$

where $A_1, A_2, \ldots, A_m$ represent the different demographic groups, $d(x, C)$ is the distance between data point x and the closest cluster center, and $|A_i|$ is the number of data points in group A_i. The outer max function ensures that no group is disproportionately affected by clustering. To efficiently solve this problem, the

authors develop approximation algorithms for fair k-median and k-means clustering using linear programming (LP) relaxation and rounding techniques. The LP relaxation allows fractional point assignments, which simplifies the optimization process while still maintaining fairness constraints. This approach ensures that fairness is enforced during clustering rather than as a post-processing step.

The main goal of the work of Makarychev and Vakilian [26] is to develop clustering algorithms that ensure fairness across different demographic groups by balancing the clustering cost among them. They try to improve previous methods, such as those of Abbasi et al. [1] and Ghadiri et al. [15] by developing more efficient approximation algorithms that ensure fairness by controlling the clustering cost across multiple groups. Given a dataset partitioned into l groups $X_1, X_2, \dots X_l$ the clustering cost for each group X_j with respect to a set of cluster centers C is given by:

$$\text{cost}(C, w_j) = \sum_{p \in P_j} w_j(p) \cdot d(p, C)^p \tag{15}$$

where the parameter p determines the clustering objective, where $p = 1$ corresponds to the k-median problem, and $p = 2$ corresponds to the k-means problem and the w_j is the demand function represents the weight or importance of a point in the clustering process for different groups, ensuring that fairness constraints are respected. The fair clustering algorithm begins by transforming the original problem instance into a modified version with adjusted demand functions. To simplify the clustering problem while preserving fairness, **location consolidation** is performed, where nearby points are merged based on a fractional distance measure $\mathcal{R}(u)$ that determines the effective separation and is defined as follows:

$$\mathcal{R}(u) := \left(\sum_{v \in X} d(u, v)^p \cdot x_{uv} \right)^{1/p} \tag{16}$$

where x_{uv} represents the assignment variable that determines the relationship between u and v.This process creates a reduced set of representative points that are well-separated, making the problem more manageable. Once the new demand structure is established, the clustering model is adjusted so that only the selected representative points serve as cluster centers. The cost of the modified clustering setup is carefully analyzed to ensure that it does not significantly exceed the cost of the original problem. Finally, to obtain a solution that assigns each point to a specific cluster, a **rounding procedure** is applied, converting the fractional clustering representation into a final integer solution. This rounding step guarantees that the number of selected cluster centers does not exceed a predefined limit while maintaining fairness constraints. A deterministic rounding approach is also introduced as an alternative, ensuring that fairness is preserved while keeping the cost within an acceptable range. By integrating demand adjustments, location consolidation, and rounding techniques, the procedure constructs a socially fair clustering solution that balances fairness, efficiency, and cost-effectiveness.

5 Individual Fairness

Individual fairness is centered on guaranteeing that data points deemed similar are treated alike, without relying on predefined protected categories. In the context of clustering, this means that a fair model should assign individuals with comparable characteristics to the same or similar clusters, based on a predefined measure of similarity.

In the work of Anderson et al. [7], the concept of individual fairness in clustering is examined for the first time. This need arises to address the possibility that standard clustering algorithms or clustering algorithms that enhance group fairness may still be unfair to similar individuals. The contribution of this work is to introduce a method that ensures similar individuals are assigned to clusters in a statistically fair way using divergence functions. In particular, a framework is presented for assigning individuals, embedded in a metric space, to *probability distributions over a bounded number of cluster centers*. The fairness criterion is that individuals that are close to each other in a given fairness space are mapped to *statistically similar probability distributions*. The framework adapts any l_p-norm clustering algorithm to maintain individual fairness while guaranteeing an approximate solution. Additionally, the study explores the relationship between individual and group fairness in clustering, ensuring fairness among individuals within protected groups while maintaining computational feasibility. The authors present the $ALG - IF(\mathcal{I})$ algorithm where $\mathcal{I}$ is an instance. Tha algorithm consists of two main steps. In the first step, a ρ-approximation algorithm is applied to solve the (k, p)-clustering problem in order to find a set of cluster centers $\mathcal{C}$ without considering fairness constraints. In the second step, they solve an optimization problem $(FAIR - ASSGN)$ to assign individuals to clusters while enforcing fairness constraints. These constraints ensure that individuals with similar characteristics are assigned to statistically similar distributions. The problem is defined as:

$$\text{FAIR-ASSGN}(\mathcal{J}): \quad \min \sum_{j \in X} \sum_{c \in C} x_{cj} d(c, j)^p \tag{17}$$

$$\text{s.t.} \quad \sum_{c \in C} x_{cj} = 1, \quad \forall j \in X \tag{18}$$

$$D_f(\bar{x}_{j_1} || \bar{x}_{j_2}) \leq \mathcal{F}(j_1, j_2), \quad \forall j_1, j_2 \in V \tag{19}$$

$$0 \leq x_{cj} \leq 1, \quad \forall j \in X, \forall c \in C \tag{20}$$

where X is the set of individuals (data points) that need to be clustered, and C is the set of cluster centers. The variable x_{cj} represents the probability that individual j is assigned to cluster center c. The function $d(c, j)$ denotes the distance between an individual j and a cluster center c. The parameter p determines the clustering objective, where $p = 1$ corresponds to the k-median problem, and $p = 2$ corresponds to the k-means problem. The term $D_f(\bar{x}_{j_1} || \bar{x}_{j_2})$ represents the *f-divergence*, which quantifies the statistical difference between the assignment distributions of two individuals j_1 and j_2. The fairness function $\mathcal{F}(j_1, j_2)$ defines an upper bound on the divergence, ensuring that individuals who are

similar receive statistically similar assignments. The second constraint ensures that each individual j is assigned to exactly one cluster, meaning that the sum of probabilities over all cluster centers must be equal to 1. The third constraint enforces fairness by limiting the divergence between the probability distributions of any two individuals. Finally, the fourth constraint ensures that all probability values remain valid, meaning they must be within the range $[0, 1]$. This function $\mathcal{F}$ thus guarantees that similar individuals receive assignments that are statistically close, maintaining individual fairness in the clustering process.

Another individual-level notion of fairness is proposed in Kleindessner et al. [22] which argues that a data point is considered individually fair if its average distance to points within its assigned cluster does not exceed its average distance to points in any other cluster. They assume X to be finite. Their definition of individual fairness for clustering is presented as follows. Let $\mathcal{C} = (C_1, \ldots, C_k)$ be a k-clustering of X, that is $X = C_1 \cup \cdots \cup C_k$ and $C_i \neq \emptyset$ for $i \in [k]$. For $x \in X$, they write $C(x)$ for the cluster C_i that x belongs to. They say that $x \in X$ is treated individually fair if either $C(x) = \{x\}$ or

$$\frac{1}{|C(x)| - 1} \sum_{y \in C(x)} d(x, y) \leq \frac{1}{|C_i|} \sum_{y \in C_i} d(x, y) \tag{21}$$

for all $i \in [k]$ with $C_i \neq C(x)$. The clustering $\mathcal{C}$ is *individually fair* if every $x \in X$ is treated individually fair.

6 Deep Fair Clustering

Fairness in deep clustering has gained increasing attention as traditional clustering methods often inherit biases from the data, leading to unfair representations of different demographic groups. Unlike classical fair clustering approaches, which primarily focus on balancing group proportions, recent works integrate deep learning techniques to learn fair representations while optimizing clustering objectives. Existing methods employ strategies such as adversarial learning to remove sensitive information from feature representations, fair distance constraints to ensure unbiased cluster centroids, and optimization-based fairness enforcement to maintain equity across groups. These advancements provide scalable and adaptive solutions, bridging the gap between deep representation learning and fairness-aware clustering, ensuring that the learned clusters remain both meaningful and unbiased.

The work of Wang and Davidson [33] was the first to integrate deep clustering with fairness constraints. The authors propose a Deep Fair Clustering method that leverages deep learning to generate embeddings that simultaneously (i) structure data into meaningful clusters and (ii) ensure a balanced representation of protected attributes within each cluster. The key innovation of this approach lies in the introduction of fairoids, which serve as fairness reference points that represent the average latent embeddings of protected groups. The clustering process begins by learning a latent representation of the input

data through a deep encoder. Specifically, the input dataset X is transformed into a lower-dimensional space using an encoder function $F_W(X)$, producing a latent representation Z. This step ensures that clustering is performed on a more structured and meaningful feature space. Once the latent space is established, cluster centroids μ_k and fairoids π_t are initialized. The centroids define the center of each cluster, while fairoids represent the central embeddings of protected groups. To ensure fairness, cluster centroids are positioned equidistantly from all fairoids, preventing overrepresentation of any single group. The model optimizes cluster assignments using a soft clustering approach, where an assignment function $\alpha(Z)$ probabilistically maps each instance in the latent space to a cluster. This is achieved using a Student's t-distribution kernel, ensuring smooth and flexible cluster formation. In parallel, the method enforces fairness constraints by minimizing the disparity between cluster centroids and fairoids, preventing clusters from being biased toward specific subgroups. To quantify and mitigate bias, the method employs Maximum Mean Discrepancy (MMD), which measures the distributional difference between protected groups across clusters. By minimizing MMD, the model ensures that the learned clusters remain both unbiased and well-structured. This approach establishes a scalable and effective framework for deep fair clustering, offering a practical solution for enforcing fairness in deep clustering tasks while maintaining high clustering performance.

The paper of Zhang and Davidson [35] introduces a deep fair clustering framework that ensures both fairness and compact clustering results. The authors propose the following key ideas. They use a probabilistic discriminative clustering method that learns feature representations to create compact and clearly defined clusters. The fairness objective in this work is expressed as an integer linear programming (ILP) problem. To ensure group-level fairness in clustering, the ILP formulation adds constraints that control the representation of sensitive groups within each cluster. The framework combines three main components: a clustering loss, fairness constraints, and contrastive learning. At the beggining, a neural network is trained to assign data points to clusters. The network predicts a probability distribution over the clusters for each data point. To ensure good clustering quality, the algorithm uses a clustering loss that makes clusters compact (points within a cluster are similar) and prevents all points from being assigned to a single cluster. The next step is to ensure that each cluster has a balanced representation of different groups (e.g., no single group dominates a cluster). To achieve this, the problem is formulated as an ILP problem, which optimizes the cluster assignments while enforcing fairness rules. Once the fair clustering assignments are obtained, they are used as "pseudo-labels" to update the network. A fairness loss is introduced to ensure the network's predictions align with these fair assignments. To make the clustering more robust and improve feature quality, the framework uses contrastive learning. This method compares original data points with slightly modified (perturbed) versions and ensures they are assigned to the same cluster. The steps are repeated (training the network, adjusting for fairness, and refining the features) until the clustering is both fair and of high quality.

The work of Li et al. [23] introduced the Deep Fair Clustering (DFC) method that simultaneously learns unbiased and well-structured representations, ensuring that sensitive attributes do not influence clustering while maintaining high performance. Unlike previous methods that impose fairness constraints directly in the input space, which may hinder clustering quality, DFC optimizes representations to enhance both fairness and clustering effectiveness. The paper introduces a more stringent definition of fairness, requiring that the clustering assignments $C(X)$—where $C(X)$ represents the clustering assignment produced by a (randomized) clustering algorithm applied to dataset X—be statistically independent of the sensitive attribute G. In other words, the cluster to which a data point is assigned should not reveal any information about its sensitive attribute. Formally, given a dataset X sampled from an underlying distribution $\mathcal{D}$, where $G = G(X)$ is the sensitive attribute that takes categorical values, and the clustering algorithm partitions X into K disjoint clusters, this fairness condition can be expressed as:

$$\mathbb{E}_{X\sim\mathcal{D}}[G \mid C(X) = c] = \mathbb{E}_{X\sim\mathcal{D}}[G], \quad \forall c \in \{1, \ldots, K\} \tag{22}$$

which ensures that the expected value of G in each cluster is the same as its expected value in the whole dataset. The objective function of DFC consists of three main components: fairness-adversarial loss, structural preservation loss, and clustering regularizer. In the **fairness-adversarial loss**, an encoder generates data representations, while a discriminator attempts to predict protected subgroup membership. Fairness is achieved when the discriminator fails to distinguish between groups, ensuring that sensitive attributes are not encoded in the learned representations. However, fairness alone is insufficient for effective clustering. To address this, **structural preservation loss** maintains the relationships between data points within protected subgroups, ensuring that the learned features remain meaningful for partitioning. Since clustering is unsupervised, a self-supervised strategy is employed, using pseudo soft assignments from individual clustering processes within each subgroup to guide learning. Finally, the **clustering regularizer** refines cluster assignments by reinforcing confidence scores while preventing clusters from being dominated by specific protected subgroups. This ensures that clusters are well-formed and balanced in representation.

7 Fairness in Graph Clustering

Fairness in graph clustering is a crucial challenge, as traditional clustering methods often reinforce biases in graph structures, leading to the underrepresentation of certain demographic groups. Standard techniques, such as spectral clustering and modularity-based community detection, do not account for fairness constraints, which can result in biased partitions. To address these issues, researchers have introduced fairness notions specifically for graph clustering. *Group fairness* ensures that different demographic groups are proportionally

represented in each cluster, with approaches such as fairness-aware spectral clustering by Kleindessner et al. [21] and the *Fairlets* method by Chierichetti et al. [14], which preprocesses the graph into fair subgroups before clustering. A more recent approach by Gkartzios et al. [17] introduces *Group Modularity*, extending traditional modularity-based clustering to explicitly enforce fairness constraints, ensuring balanced demographic representation in community detection by modifying the clustering objective. *Individual fairness* requires that nodes with similar attributes or structural roles receive similar cluster assignments, a concept explored by Mahabadi et al. [25], who adapted fairness notions from supervised learning to clustering. To mitigate bias, multiple techniques have been proposed, including optimization-based debiasing, where Backurs et al. [8] developed fairness-aware k-median clustering to integrate fairness constraints into the objective function; graph rewiring, which modifies the graph topology to balance demographic representation, as suggested by Mehrabi et al. [27]; and regularization-based approaches, such as Jiang et al. [19], which adjust node embeddings to promote fairness. Given the widespread use of graph clustering in applications such as *social network analysis, recommender systems, and knowledge graphs*, ensuring fairness is essential for equitable outcomes and remains a key research challenge.

8 Summary and Discussion

Table 1 provides an overview of various datasets used in fairness-aware clustering research. It includes key details such as the dataset name, description, possible protected attributes, a reference link, and the number of dimensions (features) for each dataset. The protected attributes column indicates which sensitive demographic or categorical attributes are considered in fairness evaluations, such as gender, race, age, financial status, or education level. Some datasets have multiple protected attributes, while others focus on a single one. The table also highlights the variety of data sources, including census records, financial transactions, social networks, and image datasets, demonstrating the wide applicability of fairness-aware clustering techniques.

Table 2 summarizes different fair clustering algorithms along with their key properties. It categorizes the algorithms based on their fairness type (balance or individual fairness), approach (pre-processing, in-processing, or post-processing), computational complexity, and whether they consider single or multiple protected attributes. The complexity column provides insights into the efficiency of each method, with some algorithms having quadratic, polynomial, or logarithmic time complexity, while others involve more computationally intensive procedures. Additionally, the table differentiates between binary and multi-valued protected attributes, indicating whether fairness constraints apply to groups with two categories (e.g., male/female) or multiple categories (e.g., different racial or income groups). This table provides useful comparison information for understanding the trade-offs between different fair clustering methods in terms of efficiency, fairness type, and applicability to various datasets.

Table 1: Datasets used to evaluate classical fair clustering algorithms.

Dataset	**Description**	**Protected Group**	**Link**	**Dim**
Diabetes	Info on medical features related to diabetes progression [8, 14, 9, 29]	Gender, race, age	Link	8
Adult Income Dataset	Census database from 1994	Sex, race, age	Link	14
Bank	Info about bank clients [8, 14, 9, 29]	Marital status, education, housing	Link	15
Census	Info about individuals, education, work hours [8, 14, 9]	Sex, age, marital status	Link	3.5
Census II	Extended census data	Sex, age, marital status	Link	68
Credit Card	Credit transactions, payment history [9]	Financial status	Link	22
4area	computer science researchers and their areas of study[4]	main area of research	Link	8
Victorian	Text fragments from 19th-century authors [4]	Authors	Link	1,000
Reuters	RCV1 subset for authorship identification [4]	Authors	Link	10,000
Iris	Flower species dataset for ML tasks [1]	Species	Link	4
NC Voters	Voter registration data in North Carolina [1]	Race	Link	30
Kinematics	Motion analysis dataset [2]	Problem type	Link	8
FriendshipNet	Social connections dataset [21]	Gender	-	-
FacebookNet	High school friendship dataset [21]	Gender	-	-
DrugNet	Drug user network adjacency matrix [21]	Sex, ethnicity	Link	2
MNIST-USPS	Combined MNIST and USPS images [35]	Sample source	Link1, Link2	1,024

Dataset	**Description**	**Protected Group**	**Link**	**Dim**
Color Reverse-MNIST	Domain adaptation dataset [35]	Image color format	-	1,024
Office-31	31-class image dataset for domain adaptation [13]	Domain source	Link	50,176
Human Action Recognition (HAR)	Human motion capture dataset	Participant IDs	Link	561
Daily and Sports Activity	Sensor records of human activity	Participant names	Link	5,625

Table 2: Summary of the Fair Clustering Algorithms

Algorithm Name	**Fairness Type**	**Approaches**	**Complexity**	**Protected Attributes**	**Type**
Fair k-center[14]	Balance	Pre-proccesing	Quadratic	Single	Binary
Fair k-median[14]	Balance	Pre-proccesing	Quadratic	Single	Binary
Fair coresets k-means clustering [29]	Balance	Pre-proccesing	O(nlogn)	Single	Multi-valued
FAIR(k,p)-CLUSTERING [9]	Balance	Pre-processing	-	Multiple	Binary
HST-based fair clustering algorithm (k-median) [8]	Balance	Pre-proccesing	Nearly-linear $O(dnlogn + T(n,d,k))$	Single	Binary
Fair algorithms for clustering [9]	Balance	Post-processing	$O(T(n,d,k)+ nlogn)$	Multiple	Binary
Clustering without over-representation [4]	Balance	Post-processing	-	Single	Multi-valued
LP-Fair k-Median [1]	Balance	In-processing	polynomial time	Multiple	Multi-valued
LS-Fair k-Median (Local Search Heuristic) [1]	Balance	In-processing	$O(n^2k)$	Multiple	Multi-valued

Algorithm Name	Fairness Type	Approaches	Complexity	Protected Attributes	Type
FairKM [2]	Balance	In-processing	$O(\|X\|^2\|N\|kl+ \|X\|\|S\|mkl)$	Multiple	Multi-valued
Fair Correlation Clustering [3]	Balance	Pre-proccesing	-	Multiple	Multi-valued
Spectral clustering [21]	Balance	In-proccesing	-	Single	Multi-valued
ALG-IF [7]	Individual Fairness	In-processing	$O(T(A_1) + T(A_2)$	Multiple	-
A notion of individual fairness for clustering [22]	Individual Fairness	In-processing	-	-	-
DFC- Faroids [33]	Deep Fair Clustering	Pre-processing	-	Single	Multi-valued
DFC-ILP [35]	Deep Fair Clustering	In-processing	$O(LdN + N^{2.5} + N^2)$	Multiple	Multi-valued
DFC [23]	Deep Fair Clustering	In-proccesing	-	Single	Multi-valued
Fair k-Means [15]	Social Fairness	In-proccesing	$O(nkd + kdTopt)$	Multiple	Multi-valued
FairLP-AbsError [1]	Social Fairness	In-proccesing	-	Multiple	-
FairLP-RelError [1]	Social Fairness	In-proccesing	-	Multiple	-
Socially Fair l_p-Clustering [26]	Social Fairness	In-proccesing	-	Multiple	-

9 Conclusions

Fair clustering is a rapidly evolving area that addresses biases in unsupervised learning, ensuring equitable representation and treatment of individuals within clustered groups. This survey inspired from the work [12] explored the main fairness notions in clustering: balance, which ensures proportional representation of protected groups; social fairness, which equalizes clustering costs across demographic groups; individual fairness, which guarantees that similar data points receive similar treatment; and deep fair clustering, which integrates fairness constraints into deep clustering methods. Through our literature review, we highlighted key methodological advancements, from fairlet decomposition and fairness-aware spectral clustering to deep learning approaches that enforce fairness through adversarial learning or constrained optimization.

Despite these recent advancements, challenges remain. Many existing meth-

ods assume binary protected attributes, limiting their applicability in real-world scenarios with multiple intersecting demographic factors. Additionally, there is a trade-off between fairness and clustering quality, with strict fairness constraints often leading to higher clustering costs. Finally, although counterfactual explanations have been explored for fairness in classification Sharma et al. [30], there is no related work for clustering. A promising future direction would be leveraging counterfactual explanations recently proposed for clustering Vardakas et al. [32] to provide new definitions of individual and cluster fairness in clustering that would take into account the cost for attaining fairness.

References

[1] Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 504–514, 2021.

[2] Savitha Sam Abraham, Sowmya S Sundaram, et al. Fairness in clustering with multiple sensitive attributes. *arXiv preprint arXiv:1910.05113*, 2019.

[3] Saba Ahmadi, Sainyam Galhotra, Barna Saha, and Roy Schwartz. Fair correlation clustering. *arXiv preprint arXiv:2002.03508*, 2020.

[4] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019.

[5] Sara Ahmadian, Alessandro Epasto, Marina Knittel, Ravi Kumar, Mohammad Mahdian, Benjamin Moseley, Philip Pham, Sergei Vassilvitskii, and Yuyan Wang. Fair hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:21050–21060, 2020.

[6] Amal AlGhamdi, Amal Barsheed, Hanadi AlMshjary, and Hanan AlGhamdi. A machine learning approach for graduate admission prediction. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, pages 155–158, 2020.

[7] Nihesh Anderson, Suman K Bera, Syamantak Das, and Yang Liu. Distributional individual fairness in clustering. *arXiv preprint arXiv:2006.12589*, 2020.

[8] 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. PMLR, 2019.

[9] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

[11] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International conference on machine learning*, pages 1032–1041. PMLR, 2019.

[12] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.

[13] Anshuman Chhabra, Peizhao Li, Prasant Mohapatra, and Hongfu Liu. Robust fair clustering: A novel fairness attack and defense framework. In *The Eleventh International Conference on Learning Representations*, 2022.

[14] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

[15] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 438–448, 2021.

[16] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

[17] Christos Gkartzios, Evaggelia Pitoura, and Panayiotis Tsaparas. Fair network communities through group modularity. In *The Web Conference 2025*, 2025. URL `https://openreview.net/forum?id=JWRQawkyz7`.

[18] Narender Gupta, Aman Sawhney, and Dan Roth. Will i get in? modeling the graduate admission process for american universities. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 631–638. IEEE, 2016.

[19] Yushun Jiang, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness-aware feature propagation for graph-based learning. In *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[20] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457. PMLR, 2019.

[21] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International conference on machine learning*, pages 3458–3467. PMLR, 2019.

[22] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. A notion of individual fairness for clustering. *arXiv preprint arXiv:2006.04960*, 2020.

[23] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020.

[24] Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai. Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436, 2011.

[25] Sepideh Mahabadi, Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Individual fairness for k-clustering. In *Advances in Neural Information Processing Systems*, 2020.

[26] Yury Makarychev and Ali Vakilian. Approximation algorithms for socially fair clustering. In *Conference on Learning Theory*, pages 3246–3264. PMLR, 2021.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[28] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[29] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854*, 2018.

[30] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[31] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. Hiring algorithms: An ethnography of fairness in practice.(2019). 2019.

[32] Georgios Vardakas, Antonia Karra, Evaggelia Pitoura, and Aristidis Likas. Counterfactual explanations for k-means and gaussian clustering. *arXiv preprint arXiv:2501.10234*, 2025.

[33] Bokun Wang and Ian Davidson. Towards fair deep clustering with multi-state protected variables. *arXiv preprint arXiv:1901.10053*, 2019.

[34] Austin Waters and Risto Miikkulainen. Grade: Machine learning support for graduate admissions. *Ai Magazine*, 35(1):64–64, 2014.

[35] Hongjing Zhang and Ian Davidson. Deep fair discriminative clustering. *arXiv preprint arXiv:2105.14146*, 2021.

[36] Imtiaz Masud Ziko, Eric Granger, Jing Yuan, and Ismail Ben Ayed. Clustering with fairness constraints: A flexible and scalable approach. *arXiv preprint arXiv:1906.08207*, 2019.