Serving Clients Fairly: On Facility Location and k-Median with Fair Outliers

Rajni Dabas
Northwestern University, Evanston, IL, USA

Samir Khuller
Northwestern University, Evanston, IL, USA

Emilie Rivkin
Northwestern University, Evanston, IL, USA

Abstract

Classical clustering problems such as *Facility Location* and *k-Median* aim to efficiently serve a set of clients from a subset of facilities – minimizing the total cost of facility openings and client assignments in Facility Location, and minimizing assignment (service) cost under a facility count constraint in k-Median. These problems are highly sensitive to outliers, and therefore researchers have studied variants that allow excluding a small number of clients as outliers to reduce cost. However, in many real-world settings, clients belong to different demographic or functional groups, and unconstrained outlier removal can disproportionately exclude certain groups, raising fairness concerns, especially when the facilities correspond to critically needed facilities for emergencies such as fire stations, hospitals and other emergency services.

We study *Facility Location with Fair Outliers*, where each group is allowed a specified number of outliers, and the objective is to minimize total cost while respecting group-wise fairness constraints. We present a bicriteria approximation with a $O(1/\epsilon)$ approximation factor and $(1 + 2\epsilon)$ factor violation in outliers per group. For *k-Median with Fair Outliers*, we design a bicriteria approximation with a $4(1 + \omega/\epsilon)$ approximation factor and $(\omega + \epsilon)$ violation in outliers per group improving on prior work by avoiding dependence on k in outlier violations. We also prove that the problems are W[1]-hard parameterized by ω.

We complement our algorithmic contributions with a detailed empirical analysis, demonstrating that fairness can be achieved with negligible increase in cost and that the integrality gap of the standard LP is small in practice.

2012 ACM Subject Classification Theory of computation → Facility location and clustering

Keywords and phrases Approximation algorithms, fairness

Digital Object Identifier 10.4230/LIPIcs.FORC.2026.9

Supplementary Material *Dataset*: `https://coral.ise.lehigh.edu/larry/research/data-sets-for-reliability-models-for-facility-location-the-expected-failure-cost-case/` [40]
Dataset: `https://archive.ics.uci.edu/dataset/2/adult` [8]
Dataset: `https://archive.ics.uci.edu/dataset/222/bank+marketing` [37]
Software: `https://github.com/EmmiRiv/Fair-FL`
archived at `swh:1:dir:e77279ffae99ff5dbac071c6dfc22ea9153df02e`

Funding Supported by NSF 2216970 (IDEAL) as well as a gift from Adobe Labs.

1 Introduction

The *facility location* problem (FL) is a classical problem in combinatorial optimization in which we are given a collection of potential facility locations $\mathcal{F}$, each with an opening cost f_i. The goal is to serve a set of clients C using the closest open facility with the objective function of minimizing the sum of the costs of opening facilities and the sum of the distances of all clients to their closest open facility. Once we select the facilities to open, the rest of the cost is completely determined, since each client simply connects to its closest open facility.

7th Symposium on Foundations of Responsible Computing (FORC 2026).
Editor: Huijia (Rachel) Lin; Article No. 9; pp. 9:1–9:19
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

This classical problem has been extensively studied, with a variety of heuristics as well as fast approximation algorithms being developed – not just that – it has been the bedrock of showcasing different techniques for the problem, including greedy algorithms, LP-rounding algorithms, randomized rounding, Primal-Dual algorithms, etc [43, 41]. So much so, that many valuable techniques can be traced back to early work on this problem. After many years of research, we finally have very fast and practical algorithms, even with an amazing understanding of their worst-case behavior (both lower and upper bounds).

Moreover, this problem is closely related to fundamental clustering problems, such as k-median (kM), which asks that k cluster centers be chosen to minimize the sum of distances of all points to the closest cluster center. In fact, one of the first approximation algorithms for the k-median works by essentially reducing it to a collection of facility location instances with different facility opening costs with each point being viewed as a client. Multiple solutions are then combined by a randomized algorithm to produce a feasible solution [31]. In most basic clustering problems, the user has to specify k, the number of clusters in advance without full knowledge of the data – with facility location by setting different values for facility costs we can obtain a spectrum of solutions that essentially creates clusters (nodes in C that are connected to a common open facility).

All of these measures are highly sensitive to outliers in the data and can lead to very skewed solutions. This led researchers to consider the problem of clustering with outliers [13]. Here, we wish to find the best possible clustering for a given (user-specified) fraction of the points. Several basic questions were considered – including k-centers, k-median and Facility Location. Over time, hundreds of papers have appeared that address the issue of outliers in many contexts [13, 29, 15, 20, 34, 23, 12, 24, 10, 14].

Our work is motivated by Almanza et al. [1] who wrote "Clustering problems and clustering algorithms are often overly sensitive to the presence of outliers: even a handful of points can greatly affect the structure of the optimal solution and its cost. This is why many algorithms for robust clustering problems have been formulated in recent years. These algorithms discard some points as outliers, excluding them from the clustering. However, outlier selection can be unfair: some categories of input points may be disproportionately affected by the outlier removal algorithm."

Returning to the basic facility location problem, we are simply asked to serve a given number of clients with the open facilities (see Figure 1) so as to minimize the cost of opening facilities and serving a given number of clients in the cheapest possible way. Thus, we have no way to control which clients become outliers.

In many applications, we deal with clients possessing diverse demographic or functional attributes. For example, urban services must be planned to serve distinct communities, yet traditional outlier-aware algorithms offer no control over which groups are disproportionately excluded. This lack of control has documented real-world consequences: a 2016 Bloomberg investigation revealed that Amazon's Prime Free Same-Day Delivery service excluded predominantly Black neighborhoods – such as Roxbury in Boston and the Bronx in New York – while fully covering the surrounding white neighborhoods [28]. Similarly, car-sharing fleets are often concentrated in affluent zones, failing to provide accessibility to underserved communities [17]. Such disparities highlight the urgent need for group-level outlier control to ensure that fairness is not sacrificed for cost efficiency. In our model, we incorporate this fairness by explicitly bounding the number of allowed outliers within each community.

In particular, our model incorporates fairness across groups by explicitly controlling the number of outliers allowed in each community. Let $(\mathcal{M}, d)$ denote a metric space where $\mathcal{M}$ is a finite set of points and $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ is a distance function satisfying triangle inequality and symmetry. The model can now be formally defined as follows:

Figure 1 A car rental company wants to create a new rental location in a major city in order to target residents who will temporarily rent a car for a day trip. In City A, most residents live alone with no children (purple circles), while in City B there is an even split between single residents and families with children (teal squares). The car rental company wants to promote its cars as being useful to both single adults and families with children. Building the center in City B will accomplish this goal of being available to both demographic groups with only a small increase in cost.

▶ **Definition 1** (Facility Location with Fair Outliers). We are given a set $\mathcal{F} \subseteq \mathcal{M}$ of potential facility locations and a set of clients $C \subseteq \mathcal{M}$ partitioned into ω disjoint groups: $C_1, C_2, \ldots, C_\omega$. Each facility $i \in \mathcal{F}$ has an associated opening cost f_i. Additionally, for each group C_g, we are given an upper bound ℓ_g on the number of clients that may be designated as outliers. The goal is to select a subset of facilities $F \subseteq \mathcal{F}$ to open and, for each group C_g, a subset of outliers $C'_g \subseteq C_g$ with $|C'_g| \leq \ell_g$ to minimize the total cost:

$$\sum_{i \in F} f_i + \sum_{g=1}^{\omega} \sum_{j \in C_g \setminus C'_g} d_{j,F}$$

where $d_{j,F}$ denotes the cost of assigning client j to its nearest open facility in F.

The k-Median with Fair Outliers problem is same as Facility Location with Fair Outliers except that, instead of facility opening costs, the k-Median variant imposes a hard budget k on the number of facilities that may be opened. Formally,

▶ **Definition 2** (k-Median with Fair Outliers). We are given a set $\mathcal{F} \subseteq \mathcal{M}$ of potential facility locations, and a set of clients $C \subseteq \mathcal{M}$ partitioned into ω disjoint groups: $C_1, C_2, \ldots, C_\omega$. For each group C_g, we are given an upper bound ℓ_g on the number of clients that may be designated as outliers. Additionally, we are given an integer k bounding the number of facilities that may be opened. The goal is to select a subset of facilities $F \subseteq \mathcal{F}$ to open with $|F| \leq k$ and, for each group C_g, a subset of outliers $C'_g \subseteq C_g$ with $|C'_g| \leq \ell_g$ to minimize:

$$\sum_{g=1}^{\omega} \sum_{j \in C_g \setminus C'_g} d_{j,F}$$

These problems are NP-hard and so researchers have resorted to the design of heuristics and approximation algorithms.

1.1 Our Contributions

We first study facility location with fair outliers problem for arbitrary number of groups in Section 3 to give a $O(1/\epsilon)$ factor approximation for the problem with $(1+2\epsilon)$ factor violation in the outliers for any group. In particular, we present Theorem 3.

Inamdar and Varadarajan [27] gave an $O(\log \omega)$-approximation for the facility location problem with fair outliers for arbitrary ω. More recently, Bajpai et al. [4] studied the case of a constant number of groups and obtained a 4-approximation. However, both these results rely on solving an exponentially large linear program using the ellipsoid method, and are therefore impractical to implement even for small values of $\omega > 1$.

▶ **Theorem 3.** *There exists a polynomial time (α/ϵ)-factor approximation algorithm with $(1 + 2\epsilon)$ factor violation in outliers for each group for facility location with fair outliers problem where α is the approximation factor for classical facility location problem, number of groups is arbitrary and $\epsilon > 0$ is a fixed parameter.*

We will show that the standard LP formulation for the problem has unbounded integrality gap even when $\omega = 1$ and hence we present a bi-criteria algorithm using LP rounding technique. Moreover, for arbitrary ω, there is a known hardness of $\Omega(\log \omega)$ due to [27] unless $P = NP$ ruling out any proper (as opposed to bi-criteria) polynomial-time approximation algorithm for an arbitrary number of groups.

In this paper, we also give a stronger hardness result. In particular, we show via a series of reductions that the problem Facility Location with Outliers is *W[1]-hard* parameterized by ω.

A problem is W[1]-hard with respect to a parameter k if no algorithm can solve it in time $f(k) \cdot \text{poly}(n)$, where $f(k)$ is an arbitrary function of k, and n is the size of the input.

▶ **Theorem 4.** *Facility Location with fair outliers problem is W[1]-hard parametrized by ω.*

We next study the k-median with fair outliers problem in Section 4, where we present a bi-criteria approximation algorithm. To the best of our knowledge, the only prior practical result for this problem is the bi-criteria approximation by [1]. They gave a constant-factor approximation, however, their algorithm allows a violation of factor in $3k + 2$ outliers per group, which can be prohibitive in practice, especially when k is large. In contrast, our algorithm eliminates this dependence on k in the violation bounds, making it more practical and scalable. The result is formally stated in Theorem 5.

The k-Median formulation presents unique algorithmic challenges and remains essential for practical applications. Historically, k-Median with outliers has proven far more complex to approximate than its Facility Location counterpart (FLO). While FLO admits straightforward 2-factor greedy or 3-factor primal-dual approaches, k-Median with outliers requires more sophisticated techniques, such as Lagrangian relaxation combined with iterative local search to achieve a constant-factor approximation [15], or iterative rounding on strengthened LPs to reach a 7.081 factor [23]. Furthermore, k-Median cannot always be emulated by FL because it enforces a *hard* cardinality budget k. In real-world scenarios with absolute funding or land-use limits – such as siting a fixed number of fire stations or hospitals – an FL formulation may produce a solution that opens more facilities than the budget allows, rendering it physically or financially infeasible. By explicitly bounding the facility count, k-Median ensures the solution remains within strict resource boundaries despite the added computational difficulty.

▶ **Theorem 5.** *There exists a polynomial time $4\,(1 + \omega/\epsilon)$ factor approximation algorithm with a $(\omega + \epsilon)$ factor violation in outliers for each group for k-median with fair outliers problem where number of groups is arbitrary and $\epsilon > 0$ is a fixed parameter.*

We also show that the hardness result for facility location extends to the k-median with fair outliers problem with slight modification in the reduction. In particular, we prove the following.

▶ **Theorem 6.** *k-median with fair outliers is W[1]-hard parameterized by ω.*

Finally, we implement our facility location and k-median algorithms and evaluate their performance on a real-world and synthetic datasets. For the facility location problem with fairness constraints, even for small values of ω (e.g., 2 groups), existing approximation algorithms [27, 4] – though polynomial-time – are impractical due to their reliance on solving exponentially large linear programs via the ellipsoid method. In contrast, our bi-criteria algorithm is significantly more efficient in practice. Empirically, it consistently outperforms its worst-case guarantees in both cost and fairness.

In summary, we observe that fairness our algorithms can ensure fairness with minimal increase in the cost of the solution with respect to non-fair algorithms (facility location with outliers).

For the k-median problem, we implement the bi-criteria algorithm and compare its performance – in both cost and fairness – against the non-fair variant, namely k-median with outliers [13]. Similar to the facility location case, we find that (i) our algorithm consistently outperforms its worst-case guarantees in practice, and (ii) it achieves significantly better fairness, while incurring virtually no increase in cost compared to the unfair baseline. We do not directly compare with Almanza et al. [1], as their experiments focus on the k-means objective, whereas our focus is on k-median. Nonetheless, our empirical findings support their conclusions in the context of the k-median objective as well.

1.2 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we review the related work. Sections 3 and 4 give details of approximation algorithms for facility location with fair outliers and k-median with fair outliers respectively. In Section 5 we layout our experimental results for facility location. The experimental results for k-median can be found in Section 6. The proof of W[1] hardness is discussed in Appendix A. Proofs of statements marked by † can be found in the appendix.

2 Related Work

Our work lies at the intersection of two important areas: the well-studied domain of clustering with outliers and fairness in clustering. We review relevant literature from both individual domains, as well as prior work at their intersection.

Clustering with Outliers. The concept of outliers in facility location problem was first introduced by Charikar et al. [13]. They observed that the standard linear programming (LP) relaxation for the Facility Location with Outliers (FLO) problem has an unbounded integrality gap. To address this issue, Charikar et al. [13] proposed a technique that involves guessing the cost of the most expensive facility in the optimal solution. After guessing the most expensive facility, they leverage the primal-dual framework of Jain and Vazirani [31] to develop a 3-factor approximation algorithm. The approximation ratio was later improved to 2 by Jain et al. [29] through a simple greedy algorithm, analyzed using the dual-fitting technique.

Charikar et al. [13] also studied the outlier variant of k-Center and k-Median Problem. For the k-Center[1] with Outliers problem, they presented a 3-factor approximation algorithm based on a greedy strategy, which was subsequently improved to a 2-factor approximation

[1] k-Center is same as k-Median except in k-Center the goal is to minimize the maximum distance instead of total distance.

via LP rounding [12, 24]. In the case of k-Median with Outliers (kMO), they proposed a bi-criteria approximation algorithm that achieves a cost approximation factor of $4(1+\frac{1}{\epsilon})$, while allowing a $(1+\epsilon)$ violation in the number of outliers. It is important to note that our approach builds directly upon this foundational work of Charikar et al. [13]. We extend their approach by incorporating group-wise fairness constraints into the underlying formulation, allowing the rest of the algorithmic process to follow their established technique seamlessly. We complement these theoretical extensions with detailed empirical experiments, demonstrating that our fair approach works efficiently in practice.

As with FLO, the standard LP relaxation for kMO also exhibits an unbounded integrality gap. However, unlike FLO, it is not straightforward to overcome this challenge, making the kMO problem more challenging and less well-understood. The first constant-factor approximation for kMO was obtained by Chen [15] using a Lagrangian relaxation approach, inspired by the framework of Jain and Vazirani [31], combined with iterative local search. However, the approximation factor in Chen's result, while constant, is relatively large and unspecified. Krishnaswamy et al. [34] significantly improved upon this by applying iterative rounding on a strengthened LP formulation, yielding a 7.081 approximation. This result was further refined by Gupta et al. [23], who improved the approximation factor to $(6.994+\epsilon)$ through enhancements in the iterative rounding technique. Friggstad et al. [20] employed natural multi-swap local search heuristics to address outliers in the k-Median problem. Their approach provides a $(3+\epsilon)$-factor approximation with a $(1+\epsilon)$-factor violation in the cardinality. They also show that any constant size multi-swap local search algorithm has unbounded locality gap for the problem, which implies that the violation in k or number of outliers is inevitable.

Incorporating outliers into clustering problems such as FL and kM significantly increases the complexity of these problems. Techniques that perform well in the standard (non-outlier) setting – such as LP rounding, primal-dual methods, and local search – do not extend straightforwardly to their outlier variants, partially because;

1. the standard linear programming (LP) relaxations for the outlier variants exhibit unbounded integrality gaps as shown in Section 3 for facility location and by Charikar et al. [13] for k-Median and,
2. multi-swap local search algorithms of constant size have unbounded locality gaps [20].

Consequently, the approximation guarantees in the presence of outliers are substantially worse and require more sophisticated algorithmic techniques. Currently, the best-known approximation ratios for FLO and kMO are 2 [29] and $6.994+\varepsilon$ [23], respectively. In comparison, their non-outlier counterparts – FL and kM – admit significantly better approximations of 1.488 [35] and 2.613 [21], respectively. Notably, the outlier variants do not currently have stronger lower bounds also; the best-known bounds remain at 1.463 [22] for FL and 1.763 [30] for kM.

Fairness for Outliers. In a parallel line of research, fairness has also been explored in clustering with outliers, where the goal is to ensure that the exclusion of certain data points as outliers does not unfairly impact individuals or communities. Traditional algorithms for clustering with outliers lack any control over which clients are discarded, which may inadvertently lead to biased decisions as shown empirically in [1] and in our experiments.

The idea of fairness for groups in the presence of outliers was first introduced in the context of the vertex cover problem[2] by Bera et al. [9]. They called the problem as *partition vertex cover problem* and gave an $O(\log \omega)$ approximation for the problem which is best

[2] Given a graph $G=(V,E)$, find a minimum-weight subset $U \subseteq V$ such that, for all $e \in E$, at least one endpoint is in U.

possible when ω is an arbitrary integer. Bandyapadhyay et al. [5] studied the problem when the vertices are unweighted. They achieve a $(2+\epsilon)$-approximation in time $n^{O(\omega/\epsilon)}$. Hong and Kao [25] studied the problem in hypergraphs and presented a $(f \cdot H_\omega + H_\omega)$-approximation, where f is the maximum edge size and H_ω is the ω^{th} harmonic number.

In the context of facility location, Inamdar and Varadarajan [27] studied facility location with fair outliers. They showed that this problem is $(\log \omega)$-hard to approximate for an arbitrary ω. They also presented a matching upper bound, $\mathcal{O}(\log \omega)$, approximation algorithm. Recently, Bajpai et al. [4] obtained a 4-approximation for the problem with constant number of groups. For the k-Center with fair outliers problem, the first algorithm was proposed by Bandyapadhyay et al. [6], providing a polynomial time 2-approximation while allowing $k+\omega$ centers. A subsequent result by Anegg et al. [2] eliminated the violation in k and obtained a 4-approximation in time $O(n^\omega)$, which was later improved to a 3-approximation in time $O(n^{\omega^2})$ by Jia et al. [32] .

For k-Median with fair outliers, Almanza et al. [1] presented a bi-criteria approximation algorithm, that is, they presented an $O(1)$ approximation algorithm with $(3k+2)$ factor violation in outliers of every group.

We continue this line of research for facility location and k-Median problem. As noted by Almanza et al. [1], this setting is “quite flexible and allows one to enforce popular fairness constraints such as demographic parity [7], calibration within groups [39], statistical parity [19], diversity rules (e.g., 80 percent rule) [11], and proportional representation rules [36].”

3 Facility Location with Fair Outliers

In this section, we present a bi-criteria approximation algorithm for the Facility Location with Fair Outliers problem. We begin by formulating the problem as an integer linear program.

In this formulation, y_i indicates whether facility i is open, z_j indicates whether client j is an outlier, and x_{ij} denotes whether client j is served by facility i. Constraints (1) and (2) ensure that each client is either assigned to an open facility or designated as an outlier. Constraints (3) impose bounds on the total number of outliers for each group.

$$\text{Minimize} \quad \sum_{i\in\mathcal{F}} f_i y_i + \sum_{j\in C}\sum_{i\in\mathcal{F}} d_{ij}x_{ij}$$

subject to

$$\sum_{i\in\mathcal{F}} x_{ij} + z_j \geq 1 \qquad \forall j \in C \tag{1}$$

$$x_{ij} \leq y_i \qquad \forall j \in C, i \in \mathcal{F} \tag{2}$$

$$\sum_{j\in C_g} z_j \leq \ell_g \qquad \forall g \in [\omega] \tag{3}$$

$$x_{ij}, y_i, z_j \in \{0,1\} \qquad \forall i \in \mathcal{F}, j \in C$$

We relax the integer constraints, allowing x_{ij}, y_i, z_j to take values in the continuous range $[0,1]$, resulting in the LP relaxation.

The LP formulation exhibits an unbounded integrality gap even when $\omega = 1$. Consider the following example: suppose there is a single facility with opening cost f, and M clients co-located at that facility. If the number of allowed outliers is $M-1$, the LP can fractionally open the facility to an extent of $1/M$ and serve each client to the same extent, effectively serving one full client in total. In contrast, any integral solution would need to fully open the facility and serve one client, incurring a cost of f, which becomes arbitrarily large relative to the LP cost as M increases – thus leading to an unbounded integrality gap. The example can be modified to allow a smaller number of outliers relative to the total number of clients by adding additional groups of clients served by facilities with zero opening cost. For these added groups, both the LP and the integral solutions coincide, so they do not affect the integrality gap, which still arises from the original group.

Let $\rho^* = \langle x^*, y^*, z^* \rangle$ be an LP optimal solution for the LP. For any solution $\rho = \langle x, y, z \rangle$ to the LP, let $\text{cost}(\rho)$ denote its cost.

3.1 Identifying the Outliers

We now identify the set of clients to be treated as outliers in our solution. The idea is to declare a client as an outlier if it is *predominantly* an outlier in the LP solution ρ^*. For a given $\epsilon > 0$, we partition the client set C into:

(i) $C_o = \{j \in C : z_j^* \geq 1 - \epsilon\}$
(ii) $C_r = C \setminus C_o$

We define a new solution $\hat{\rho} = \langle \hat{x}, \hat{y}, \hat{z} \rangle$ as follows:

(i) For $j \in C_o$, set $\hat{z}_j = 1$ and $\hat{x}_{ij} = 0$ for all $i \in \mathcal{F}$.
(ii) For $j \in C_r$, set $\hat{z}_j = z_j^*$ and $\hat{x}_{ij} = x_{ij}^*$ for all $i \in \mathcal{F}$.
(iii) For all $i \in \mathcal{F}$, set $\hat{y}_i = y_i^*$.

For any group $g \in [\omega]$, we observe:

$$\sum_{j \in C_g} \hat{z}_j \leq \frac{1}{1-\epsilon} \sum_{j \in C_g} z_j^* \leq (1 + 2\epsilon)\ell_g$$

Moreover, $\text{cost}(\hat{\rho}) \leq \text{cost}(\rho^*)$.

3.2 Reduction to Facility Location

We now scale the assignment variables for $j \in C_r$ so that each such client is fully served, while maintaining feasibility. Let $\rho' = \langle x', y', z' \rangle$ be the updated solution:

1. For all $j \in C_r$ and all $i \in \mathcal{F}$, set:

$$x'_{ij} = \frac{\hat{x}_{ij}}{\sum_{i \in \mathcal{F}} \hat{x}_{ij}}.$$

2. For each $i \in \mathcal{F}$, set:

$$y'_i = \min\left\{1, \hat{y}_i \cdot \max_{j \in C_r : \hat{x}_{ij} > 0} \left\{\frac{x'_{ij}}{\hat{x}_{ij}}\right\}\right\}.$$

Note that for $j \in C_r$, $\sum_{i \in \mathcal{F}} \hat{x}_{ij} \geq \epsilon$, so:

$$x'_{ij} \leq \frac{\hat{x}_{ij}}{\epsilon}, \quad \text{and} \quad \hat{y}_i \leq y'_i \leq \frac{\hat{y}_i}{\epsilon}.$$

It follows that ρ' is a fractional feasible solution to the standard Facility Location problem with client set C_r, and:

$$\text{cost}(\rho') \leq \frac{\text{cost}(\hat{\rho})}{\epsilon} \leq \frac{\text{cost}(\rho^*)}{\epsilon}.$$

Let $\langle \bar{x}, \bar{y} \rangle$ be a solution to the Facility Location problem on C_r with cost at most $\alpha \cdot \text{LP}_{C_r}$, where LP_{C_r} is the cost of the optimal solution to the LP relaxation restricted to C_r. Then, the solution $\langle \bar{x}, \bar{y}, \hat{z} \rangle$ constitutes a $\frac{\alpha}{\epsilon}$-approximate solution to the Facility Location with Fair Outliers problem, violating the group outlier bounds by at most a factor of $(1 + 2\epsilon)$. Hence, we obtain the following theorem.

▶ **Theorem 3.** *There exists a polynomial time (α/ϵ)-factor approximation algorithm with $(1 + 2\epsilon)$ factor violation in outliers for each group for facility location with fair outliers problem where α is the approximation factor for classical facility location problem, number of groups is arbitrary and $\epsilon > 0$ is a fixed parameter.*

Note that once the set of outliers has been identified, any standard facility location algorithm [29, 3, 16] can be applied to determine which facilities to open for the remaining clients. In our experiments, after removing the outliers, we use a simple heuristic for facility location in which we directly round the fractional facility openings and assign remaining clients to nearest facility. The variable values are in fact close to 0/1. This is the primary reason we do not violate the outlier or cost constraints once we solve the LP.

4 k-Median with Fair Outliers

In this section, we focus on the k-Median with Fair Outliers problem. We develop a bi-criteria approximation algorithm leveraging the well-studied framework of k-Median with Penalties. The core idea is to encode fairness constraints via appropriately scaled penalties and then apply a known constant-factor approximation algorithm for the penalty-based variant. This approach allows us to recover both cost and fairness guarantees in the original problem.

We begin by defining the k-Median with Penalties problem.

▶ **Definition 7** (k-Median with Penalties)**.** Given a metric space $(\mathcal{M}, d)$, a set of facilities $\mathcal{F} \subseteq \mathcal{M}$, a set of clients $C \subseteq \mathcal{M}$, penalties $p_j \geq 0$ for each client $j \in C$, and an integer k bounding the number of open facilities, the goal is to select (i) a subset $F \subseteq \mathcal{F}$ with $|F| \leq k$, and (ii) a subset $C' \subseteq C$ of outliers paying penalties, minimizing the total cost

$$\sum_{j \in C \setminus C'} d_{j,F} + \sum_{j \in C'} p_j,$$

where $d_{j,F}$ is the distance from client j to its nearest open facility in F.

Charikar et al. [13] provide a 4-approximation algorithm for this problem:

▶ **Theorem 8** ([13])**.** *Let C be the assignment cost of the returned solution and P be the total penalty paid by outliers. If OPT_P denotes the cost of the optimal solution to the penalty-based problem, then $C + 4P \leq 4\,\mathrm{OPT}_P$.*

4.1 Our Algorithm

We design our approximation for k-Median with Fair Outliers as follows:

1. Guess the optimal cost OPT_C of the k-Median with Fair Outliers instance within a factor $(1 + \varepsilon)$.
2. For each group $g \in [\omega]$ and each client $j \in C_g$, set the penalty $p_j = \frac{\mathrm{OPT}_C}{\gamma \cdot \ell_g}$, where $\gamma > 0$ is a parameter to be chosen.
3. Solve the corresponding k-Median with Penalties instance with penalties $\{p_j\}$ using the 4-approximation algorithm from Theorem 8.
4. Let F be the set of opened facilities, C_o be the set of outliers paying penalties, and σ be the assignment of remaining clients $C \setminus C_o$, then output F as the set of opened facilities, C_o as the set of outliers, and σ as the assignment of remaining clients $C \setminus C_o$ for the k-Median with Fair Outliers instance.

4.2 Analysis

▶ **Proposition 9.** *For k-median with ω groups and ℓ_g outliers per group, the optimal cost OPT_C can be guessed within a factor $(1+\varepsilon)$ using polynomially many trials.*

Proof. Let $d_{\min}$ and $d_{\max}$ denote the smallest and largest pairwise distances in the metric, and $\ell = \sum_{g=1}^{\omega} \ell_g$. Since each non-outlier client contributes at least $d_{\min}$ and at most $d_{\max}$ to the cost, we have$(n-\ell)d_{\min} \leq \mathrm{OPT}_C \leq (n-\ell)d_{\max}$. By considering powers of $(1+\varepsilon)$ within this range, we require only $O\big(\log_{1+\varepsilon} n\big)$ guesses to identify an estimate within a $(1+\varepsilon)$ factor. ◀

Now, we can slightly reword Theorem 5:

▶ **Theorem 10.** *For the k-Median with Fair Outliers problem with ω groups, there exists a polynomial-time algorithm that returns a solution with cost at most*

$$4\left(1+\frac{\omega}{\gamma}\right)\mathrm{OPT}_C,$$

and with at most $(\omega+\gamma)\ell_g$ outliers in each group g.

Proof. Consider the optimal fair solution with cost OPT_C and ℓ_g outliers per group g. Using the penalty assignment defined above, this solution is feasible for the k-Median with Penalties problem so,

$$\mathrm{OPT}_P \leq \mathrm{OPT}_C + \sum_{g=1}^{\omega} \frac{\mathrm{OPT}_C}{\gamma \ell_g}\ell_g = \mathrm{OPT}_C\left(1+\frac{\omega}{\gamma}\right).$$

We can also take this instance of k-medians with penalties and run it through the black box algorithm. It returns a solution of cost C with ℓ' outliers. We partition the outliers by color. By Theorem 8 and the above inequality,

$$C + 4\sum_{c=1}^{\omega}\frac{\mathrm{OPT}_C}{\gamma \ell_c}\ell'_g \leq 4\,\mathrm{OPT}_P \leq 4\,\mathrm{OPT}_C\left(1+\frac{\omega}{\gamma}\right).$$

Since all terms are non-negative, we have

$$C \leq 4\,\mathrm{OPT}_C\left(1+\frac{\omega}{\gamma}\right) \text{ and } \sum_{c=1}^{\omega}\frac{1}{\gamma \ell_g}\ell'_g \leq 1+\frac{\omega}{\gamma}.$$

Again, each term is non-negative, so $\ell'_g \leq \ell_g\,(\omega+\gamma)$.

By Proposition 9, we can guess OPT_C within a $(1+\varepsilon)$ factor, so the guarantees on C and ℓ'_g for each group g hold within the same factor. ◀

5 Experiments for facility location

5.1 Datasets

We use a dataset inspired by Snyder and Daskin [40]. In their paper, they study the facility location from the lens of reliability after the failure of some constructed facilities. We use two of their datasets and pull key pieces of information for our application. The two datasets we use are a 49-node set on the capital cities of the continental United States and an 88-node set on the first set with the 50 most populated cities added.[3] These numbers are from the 1990 census. Each city is also a facility, and the facility opening costs are the median property value in the city. The distances are the pairwise great-circle distances.

We then add more information to the data for our purposes. We partition the cities of the 49-node set based on the political party of the state governor at the time ($|C_R| = 23$, $|C_D| = 26$). In the 88-node set, the additional cities get their own group ($|C_C| = 39$). Finally, we scale down the property values by a factor of 10^4 to account for the fact that we do not incorporate the demand used in the original paper.

5.2 Algorithms

We evaluate our *bi-criteria algorithm* based on LP rounding. We compare the cost and fairness of this algorithm against classical facility location with outliers that does not enforce group-level fairness constraints. Additionally, we compare the solution costs of all algorithms to the LP optimal value for benchmarking.

We next describe the two algorithms in detail:

- **LPR-F (LP Rounding with Fairness)**: This is our main bi-criteria algorithm. It begins by solving the fair LP relaxation with group-level outlier constraints and declares clients outliers if they are outliers to an extent of over $1 - \epsilon$. In the second phase, it uses the simple 4-approximation from Chudak et al. [16].
- **LPR-NF (LP Rounding, Non-Fair)**: A baseline variant of LPR-F that does not enforce fairness. The LP includes only a single constraint on the total number of outliers. The rounding process follows as in LPR-F.
- **LPH-F (LP Rounding Heuristic with Fairness)**: The first phase follows LPR-F. In the second phase, it applies a simple heuristic for facility location problem: facilities with fractional opening values close to 1 are opened, and each client is assigned to its nearest open facility.
- **LPH-NF (LP Rounding Heuristic, Non-Fair)**: A variant of LPH-F that does not enforce group fairness constraints.

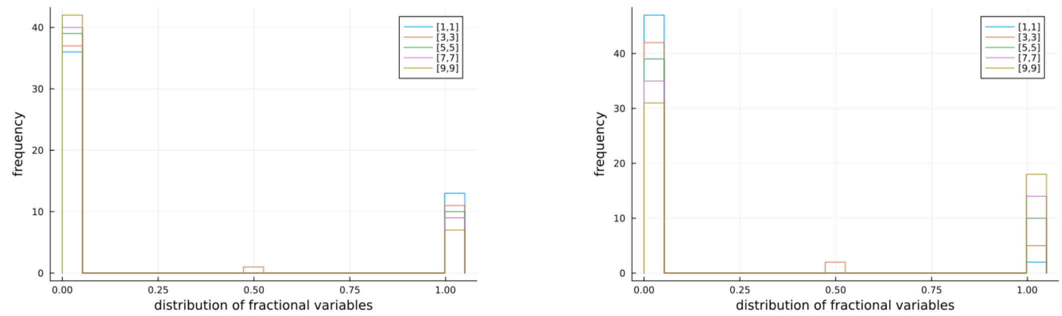
5.3 Results and insights

We define the *unfairness* as $\max\left\{\max_g \left\{\ell'_g/\ell_g\right\}, 1\right\}$, where ℓ'_g is the number of outliers assigned to group C_g by the algorithm. This definition reflects a situation in which each group is aware of its designated number of outliers and feels underrepresented when this budget is exceeded. However, a group does not experience unfairness if it is overrepresented in the covered population.

We say that an outlier budget is proportional if $\ell_g/|C_g| \approx \ell_h/|C_h|$ for all groups C_g, C_h. Similarly, we define a proportional outlier budget as one where $\ell_g \approx \alpha \cdot |C_g|$ for all groups C_g and fixed $0 \leq \alpha \leq 1$.[4]

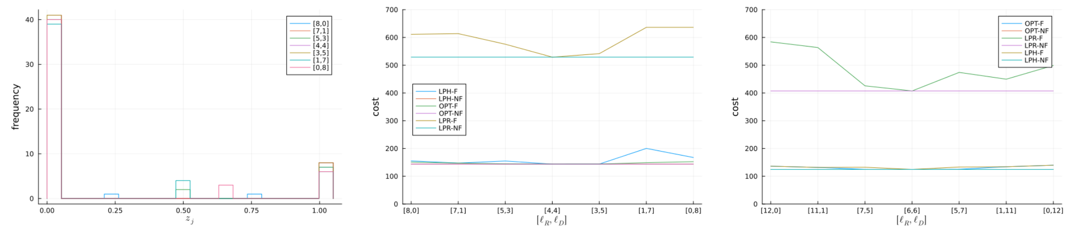
- We observe in Figure 2 that the initial LP solution differentiates between "strong outliers" with z_j close to 1 and "weak outliers" with z_j close to 0. Moreover, the facility opening variables y_i show this same bimodality. In this real-world instance, the LP produces near-integer variables, without the need for rounding, when the outlier budgets are close to proportional.

[3] Eliminating repeated cities.
[4] We use $\approx$ to account for rounding issues.



(a) Distribution of y_i values. **(b)** Distribution of z_j values.

Figure 2 Histogram showing the distribution of fraction LP variables when changing outlier budgets $[\ell_R, \ell_D]$. Uses the 49-node set.



(a) 8 total outliers. **(b)** 8 total outliers. **(c)** 12 total outliers.

Figure 3 Effects of redistributing the same number of outliers among the two groups. In Figure 3a, the histogram for fractional z_j variables appears. When the outliers are not distributed proportionately, there is a larger number of cities that are not outliers to an integral extent. Uses the 49-node set and $\epsilon = 0.1$.

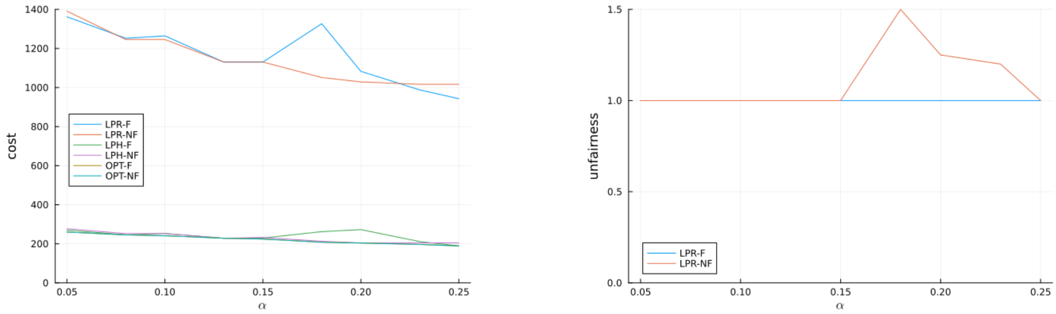
- Because the fractional variables are so close to integral in the first LP, simply rounding the facility variables in the second stage performs extremely well. In all cases, the heuristic rounding procedure exceeds the cost of the LP optimal solution by a factor of at most 1.4.
- The cost is affected by the way the per-group outliers are distributed. When the total number of outliers is fixed, moving the per-group outlier budget away from proportional increases the cost when following the budget. When the per-group budgets are close to proportional, the cost of a fair assignment is close to the cost of simply ignoring the total number of outliers. In Figure 3, this trend is demonstrated most strongly by LPR-F.
- We also observe that the algorithm without fairness can fail to provide a fair assignment while our fair algorithm consistently does. Although the bicriteria allows for violation in outliers, the z_j variables tend towards integral values, so we do not violate the outlier budgets. Moreover, ensuring proportional fairness only increases the cost by a factor of 1.3 at most.

▶ **Corollary 11.** *k-median with fair outliers is W[1]-hard parametrized by ω.*

6 Experiments for k-Median

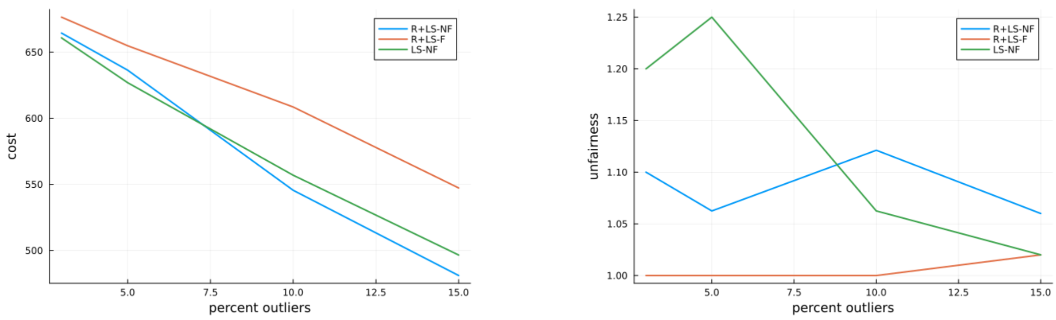
6.1 Datasets

We use publicly available datasets, as preprocessed in [1]. The *Adult* dataset [18] contains U.S. census information, where we use the `sex` attribute as the group label. The *Bank* dataset [38] comprises data from a direct marketing campaign by a bank, with group labels



(a) Costs. **(b)** Unfairness.

Figure 4 The effect of an increasing fraction of outliers on the cost and unfairness. Uses the 88-node set and $\epsilon = 0.1$.



(a) Cost vs. percentage of outliers. **(b)** Unfairness vs. percentage of outliers.

Figure 5 Performance of different algorithms on the *adult* dataset grouped by *sex* for the k-median problem.

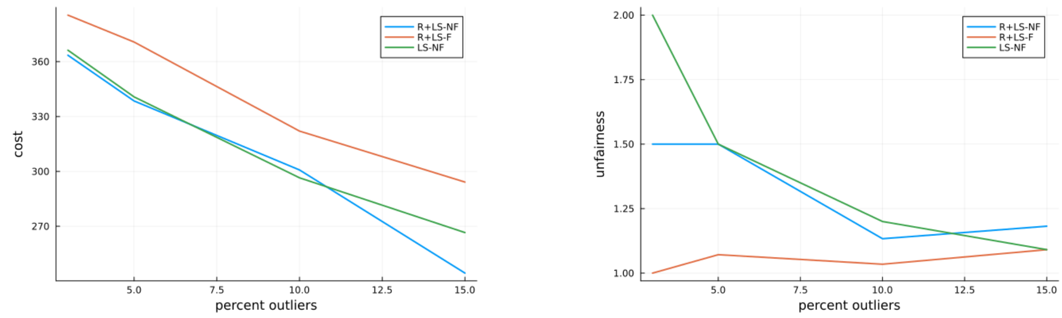
based on the `marital status` attribute. For both datasets, we retain only the available numeric attributes, each of which is normalized independently. Due to the computational overhead of our algorithm on a non-commercial machine, we use a random subset of 500 points, and we generate a set of candidate facility locations using the k-means algorithm of Khandelwal et al. [33]. For all experiments, we use $k = 5$.

Note that both datasets are fully unsupervised – that is, they lack ground-truth cluster labels. Consequently, we cannot directly evaluate clustering accuracy. To address this limitation, we also evaluate our algorithms on synthetic datasets with known cluster structure. The *Synthetic* dataset includes two group designations: "in" and "out". In-group members are drawn from the distribution $\mathcal{N}(0, 10)^{\otimes 2}$, while out-group members are drawn from $\mathcal{N}(10, 20)^{\otimes 2}$. We generate 500 "in" and 50 "out" points.

6.2 Algorithms

For k-Median with fair outliers, we implement the algorithm described in Section 4 with a minor modification and compare the fairness and cost of our algorithm against two baselines for k-Median with outliers (non-fair). The three algorithms we evaluate are:

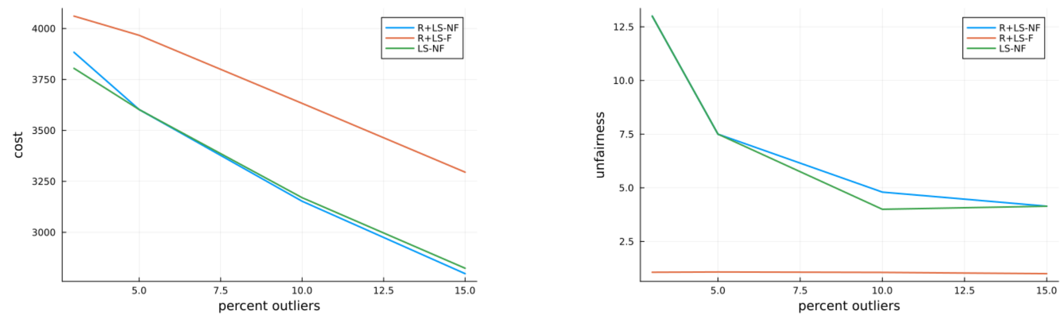
- **R+LS-F (Reduction + Local Search, Fair):** We implement the algorithm described in Section 4 with a minor modification: instead of using the primal-dual algorithm of Charikar et al. [13] for k-Median with penalties, we employ the local search algorithm proposed by Wang et al. [42]. The primal-dual algorithm is computationally expensive due



(a) Cost vs. percentage of outliers.

(b) Unfairness vs. percentage of outliers.

Figure 6 Performance of different algorithms on the *bank* dataset grouped by *marital status* for the k-median problem.



(a) Cost vs. percentage of outliers.

(b) Unfairness vs. percentage of outliers.

Figure 7 Performance of different algorithms on the *synthetic* dataset with known group structure for the k-median problem.

to the use of Lagrangian relaxation; specifically, it requires solving two facility location subproblems – one using fewer than k facilities and one using more – through a binary search, with each step involving a call to a facility location solver. This significantly increases the runtime in practice. In contrast the local search is efficient. The algorithm begins with an arbitrary set of k open facilities and repeatedly considers swapping one currently open facility with one currently closed facility. Specifically, for each pair $(f_{\text{out}}, f_{\text{in}})$ where f_{out} is an open facility and f_{in} is closed, it computes the cost of the solution obtained by replacing f_{out} with f_{in}, reassigning each client to its nearest open facility. If such a swap results in a decrease in total cost of at least 1%, it is accepted. This process continues until no improving swap exists, at which point the algorithm terminates. Importantly, substituting the local search algorithm for the primal-dual method does not materially affect the worst-case approximation guarantees: the cost bound remains the same, and the violation bound increases only slightly.

- **R+LS-NF (Reduction + Local Search, Non Fair)** The algorithm is same as R+LS except we have no notion of group-wise fairness. The penalties are defined according to the total number of outliers. This is the algorithm for k-median with outlier given by Charikar et al. [13].
- **LS-NF (Local Search, Non-Fair):** This algorithm performs standard local search for the k-Median objective, starting with an arbitrary set of k open facilities and repeatedly considering facility swaps as in R+LS. After convergence, the farthest ℓ clients (based on their distances to the nearest open facility) are dropped as outliers.

We use the same *unfairness* metric to measure the fairness of a solution and evaluate cost as the total client-to-facility connection cost. We plot the same charts as in the facility location experiments: cost vs. percentage of outliers and unfairness vs. percentage of outliers for all three algorithms.

6.3 Results and Insights

- The fair algorithm (R+LS-F) achieve *unfairness close to 1*, while the non-fair baselines R+LS-NF and LS-NF exhibit significant group-level unfairness of up to 1.12 and 1.25 for adult dataset, 1.50 and 2 for bank dataset, and 13 and 13 for synthetic dataset, respectively (see Figures 5b, 6b and 7b).
- Enforcing fairness results in some increase in cost, it is not prohibitively large and is justified by the improved group-level guarantees (see Figures 5a, 6a and 7a).

As with the facility location experiments in this paper and the k-means results of Almanza et al. [1], our findings for k-Median demonstrate that *fairness can be achieved with no significant increase in cost.*

7 Conclusion

Our algorithms succeed at maintaining fairness constraints without significant increase in cost. However, the worst-case guarantees of these algorithms are poor. A further direction of study is to use a combinatorial algorithm to achieve both scalability and a constant-factor approximation. This work can also be generalized to the case where a client can have multiple group attributes.

References

1 Matteo Almanza, Alessandro Epasto, Alessandro Panconesi, and Giuseppe Re. k-clustering with fair outliers. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 5–15. ACM, 2022. doi:10.1145/3488560.3498485.

2 Georg Anegg, Haris Angelidakis, Adam Kurpisz, and Rico Zenklusen. A technique for obtaining true approximations for k-center with covering constraints. *Math. Program.*, 192(1):3–27, 2022. doi:10.1007/S10107-021-01645-Y.

3 Vineet Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for k-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004. doi:10.1137/S0097539702416402.

4 Tanvi Bajpai, Chandra Chekuri, and Pooja Kulkarni. Covering a few submodular constraints and applications, 2025. doi:10.48550/arXiv.2507.09879.

5 Sayan Bandyapadhyay, Aritra Banik, and Sujoy Bhore. On colorful vertex and edge cover problems. *Algorithmica*, 85(12):3816–3827, 2023. doi:10.1007/S00453-023-01164-6.

6 Sayan Bandyapadhyay, Tanmay Inamdar, Shreyas Pai, and Kasturi R. Varadarajan. A constant approximation for colorful k-center. In Michael A. Bender, Ola Svensson, and Grzegorz Herman, editors, *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany*, volume 144 of *LIPIcs*, pages 12:1–12:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.ESA.2019.12.

7 Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL: http://www.fairmlbook.org.

8 Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

9 Suman Kalyan Bera, Shalmoli Gupta, Amit Kumar, and Sambuddha Roy. Approximation algorithms for the partition vertex cover problem. In Subir Kumar Ghosh and Takeshi Tokuyama, editors, *WALCOM: Algorithms and Computation, 7th International Workshop, WALCOM 2013, Kharagpur, India, February 14-16, 2013. Proceedings*, volume 7748 of *Lecture Notes in Computer Science*, pages 137–145. Springer, 2013. doi:10.1007/978-3-642-36065-7_14.

10 Aditya Bhaskara, Sharvaree Vadgama, and Hong Xu. Greedy sampling for approximate clustering in the presence of outliers. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11146–11155, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/73983c01982794632e0270cd0006d407-Abstract.html.

11 Daniel Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Routledge, 2 edition, 2006. doi:10.4324/9781315263298.

12 Deeparnab Chakrabarty, Prachi Goyal, and Ravishankar Krishnaswamy. The non-uniform k-center problem. *CoRR*, abs/1605.03692, 2016. arXiv:1605.03692.

13 Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In S. Rao Kosaraju, editor, *Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, January 7-9, 2001, Washington, DC, USA*, pages 642–651. ACM/SIAM, 2001. URL: http://dl.acm.org/citation.cfm?id=365411.365555.

14 Jiecao Chen, Erfan Sadeqi Azer, and Qin Zhang. A practical algorithm for distributed clustering and outlier detection. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2253–2262, 2018. URL: https://proceedings.neurips.cc/paper/2018/hash/f7f580e11d00a75814d2ded41fe8e8fe-Abstract.html.

15 Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In Shang-Hua Teng, editor, *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, pages 826–835. SIAM, 2008. URL: http://dl.acm.org/citation.cfm?id=1347082.1347173.

16 Fabián A. Chudak and David B. Shmoys. Improved approximation algorithms for the uncapacitated facility location problem. *SIAM J. Comput.*, 33(1):1–25, 2003. doi:10.1137/S0097539703405754.

17 Jennifer Dill and Nathan McNeil. Are shared vehicles shared by all? a review of equity and vehicle sharing. *Journal of Planning Literature*, 36 (1):5–30, 2020. doi:10.1177/0885412220966732.

18 Dheeru Dua and Casey Graff. Adult data set. https://archive.ics.uci.edu/ml/datasets/adult, 2017. UCI Machine Learning Repository.

19 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In Shafi Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012. doi:10.1145/2090236.2090255.

20 Zachary Friggstad, Kamyar Khodamoradi, Mohsen Rezapour, and Mohammad R. Salavatipour. Approximation schemes for clustering with outliers. In Artur Czumaj, editor, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 398–414. SIAM, 2018. doi:10.1137/1.9781611975031.27.

21 Kishen N. Gowda, Thomas W. Pensyl, Aravind Srinivasan, and Khoa Trinh. Improved bi-point rounding algorithms and a golden barrier for k-median. In Nikhil Bansal and Viswanath Nagarajan, editors, *SODA*, pages 987–1011. SIAM, 2023. doi:10.1137/1.9781611977554.ch38.

22 Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. In Howard J. Karloff, editor, *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 25-27 January 1998, San Francisco, California, USA*, pages 649–657. ACM/SIAM, 1998. URL: http://dl.acm.org/citation.cfm?id=314613.315037.

23 Anupam Gupta, Benjamin Moseley, and Rudy Zhou. Structural iterative rounding for generalized k-median problems. In Nikhil Bansal, Emanuela Merelli, and James Worrell, editors, *ICALP*, volume 198 of *LIPIcs*, pages 77:1–77:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPIcs.ICALP.2021.77.

24 David G. Harris, Thomas W. Pensyl, Aravind Srinivasan, and Khoa Trinh. A lottery model for center-type problems with outliers. *ACM Trans. Algorithms*, 15(3):36:1–36:25, 2019. doi:10.1145/3311953.

25 Eunpyeong Hung and Mong-Jen Kao. Approximation algorithm for vertex cover with multiple covering constraints. *Algorithmica*, 84(1):1–12, 2022. doi:10.1007/S00453-021-00885-W.

26 Tanmay Inamdar, Daniel Lokshtanov, Saket Saurabh, and Vaishali Surianarayanan. Parameterized complexity of fair bisection: (fpt-approximation meets unbreakability). In Inge Li Gørtz, Martin Farach-Colton, Simon J. Puglisi, and Grzegorz Herman, editors, *31st Annual European Symposium on Algorithms, ESA 2023, September 4-6, 2023, Amsterdam, The Netherlands*, volume 274 of *LIPIcs*, pages 63:1–63:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.ESA.2023.63.

27 Tanmay Inamdar and Kasturi R. Varadarajan. On the partition set cover problem. *CoRR*, abs/1809.06506, 2018. arXiv:1809.06506.

28 David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it?, April 2016. Accessed: 2026-01-20. URL: https://www.bloomberg.com/graphics/2016-amazon-same-day/.

29 Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. *J. ACM*, 50(6):795–824, 2003. doi:10.1145/950620.950621.

30 Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 731–740. ACM, 2002. doi:10.1145/509907.510012.

31 Kamal Jain and Vijay V. Vazirani. Primal-dual approximation algorithms for metric facility location and k-median problems. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 2–13. IEEE Computer Society, 1999. doi:10.1109/SFFCS.1999.814571.

32 Xinrui Jia, Kshiteej Sheth, and Ola Svensson. Fair colorful k-center clustering. *Math. Program.*, 192(1):339–360, 2022. doi:10.1007/S10107-021-01674-7.

33 Siddhesh Khandelwal and Amit Awekar. Faster k-means cluster estimation. *CoRR*, abs/1701.04600, 2017. arXiv:1701.04600.

34 Ravishankar Krishnaswamy, Shi Li, and Sai Sandeep. Constant approximation for k-median and k-means with outliers via iterative rounding. *CoRR*, abs/1711.01323, 2017. arXiv:1711.01323.

35 Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. In Luca Aceto, Monika Henzinger, and Jirí Sgall, editors, *Automata, Languages and Programming - 38th International Colloquium, ICALP 2011, Zurich, Switzerland, July 4-8, 2011, Proceedings, Part II*, volume 6756 of *Lecture Notes in Computer Science*, pages 77–88. Springer, 2011. doi:10.1007/978-3-642-22012-8_5.

36 Burt L. Monroe. Fully proportional representation. *American Political Science Review*, 89(4):925–940, 1995. doi:10.2307/2082518.

37 S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5K306.

38 Sérgio Moro, Raul Laureano, and Paulo Cortez. Bank marketing data set. https://archive.ics.uci.edu/ml/datasets/bank+marketing, 2014. UCI Machine Learning Repository.

39 Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5680–5689, 2017. URL: `https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html`.

40 Lawrence V. Snyder and Mark S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39(3):400–416, 2005. `doi:10.1287/trsc.1040.0107`.

41 Vijay V. Vazirani. *Approximation algorithms*. Springer, 2001. URL: `http://www.springer.com/computer/theoretical+computer+science/book/978-3-540-65367-7`.

42 Yishui Wang, Dachuan Xu, Donglei Du, and Chenchen Wu. Local search algorithms for k-median and k-facility location problems with linear penalties. In *Proceedings of the 9th International Conference on Combinatorial Optimization and Applications - Volume 9486*, COCOA 2015, pages 60–71, Berlin, Heidelberg, 2015. Springer-Verlag. `doi:10.1007/978-3-319-26626-8_5`.

43 David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms*. Cambridge University Press, 2011. URL: `http://www.cambridge.org/de/knowledge/isbn/item5759340/?site_locale=de_DE`.

A Lower Bounds

▶ **Theorem 12.** *Facility location with fair outliers is W[1]-hard parametrized by ω.*

We use the following lemmas to prove this statement. First, we need to define the problems we will use along the way.

▶ **Definition 13** (Multidimensional Subset Sum [26])**.** This is a known W[1]-hard problem. An instance $(V, T$ of MSS consists of n d-dimensional vectors with non-negative entries in $\mathbb{Z}$ and a target vector T. We want to determine if there is a subset $U \subseteq V$ such that $\sum_{v \in U} v = T$.

▶ **Definition 14** (k MSS)**.** This is a variant on MSS where we want to know if there is a suitable U with cardinality k.

▶ **Definition 15** ($k\,\text{MSS}_{\geq}$)**.** Given a set of d-dimensional vectors V with non-negative, integer-valued entries and a target vector T, we want to find a subset $U \subseteq V$ such that $|U| \leq k$ and for all $j = 1, \ldots, d$, $\sum_{v_j \in U} \geq t_j$.

▶ **Lemma 16.** *k MSS is W[1]-hard.*

Proof. Take an instance (V, t) of MSS. For each $k = 1, \ldots, n$, use the decider for k MSS to determine if (V, t, k) is an instance of k MSS. If $(V, t) \in$ MSS, then there exists some k for which $(V, t, k) \in k$ MSS. If $(V, t) \notin$ MSS, then there is no such k. Therefore, we can use the decider for k MSS to solve an instance of MSS. ◀

▶ **Lemma 17.** *$k\,\text{MSS}_{\geq}$ is W[1]-hard.*

Proof. Take an instance (V, t, k) of k MSS. For a sufficiently large constant L and for each $k' = 1, \ldots, k$, consider an instance (V', t', k') of $k\,\text{MSS}_{\geq}$ where $v' = (v_1, \ldots, v_d, L - v_1, \ldots, L - v_d)$ for every $v \in V$ and $t' = (t_1, \ldots, t_d, k'L - t_1, \ldots, k'L - t_d)$.

We can use the decider for $k\,\text{MSS}_{\geq}$ to determine if $(V, t, k) \in k$ MSS. We assumed that L is sufficiently large (perhaps $k \cdot t_1 \cdots t_d$). As a result, we do not have to worry about the case where $k'L - t_i \leq (k' - 1)L$; we can treat the sum to $k'L$ and the sum to $-t_i$ separately.

If $(V', t', k') \in k\,\mathrm{MSS}_{\geq}$, then there are exactly k' vectors in the set U' with

$$\sum_{v' \in U'} v'_j = \sum_{v' \in U'} L - v_j \geq t'_j = k'L - t_j$$

for $j = d+1, \ldots, 2d$. Since we know that the $k'L$ is only coming from the contribution of L from each vector, this implies that

$$\sum_{v' \in U'} v_j \leq t_j.$$

Moreover, those same vectors in U' have

$$\sum_{v' \in U'} v'_j = \sum_{v' \in U'} v_j \geq t_j$$

for $j = 1, \ldots, d$. If a set U' exists, then it is a size $k' \leq k$ subset of V' with

$$\sum_{v' \in U'} v_j \leq t_j.$$

Therefore, $(V, t, k) \in k\,\mathrm{MSS}$. If there is no $k' \leq k$ where $(V', t', k') \in k\,\mathrm{MSS}_{\geq}$, then $(V, t, k) \notin k\,\mathrm{MSS}$. ◀

Now, we can prove the theorem.

Proof. For each vector, create a facility with co-located clients. These facilities are some large distance away from each other. Specifically, for each vector $v \in V$, create a facility i with $f_i = 1$. For each dimension $g = 1, \ldots, d$, place v_g clients of group g colocated with the facility. Now, we have an instance of facility location with fair outliers, (F, C, ℓ), where we have to cover at least ℓ_g clients from each group.

By using the decider for the facility location with fair outliers problem with cost no more than k, we can solve the instance of $k\,\mathrm{MSS}_{\geq}$. ◀