Approximation Schemes for Clustering with Outliers

Zachary Friggstad*† Kamyar Khodamoradi* Mohsen Rezapour*
Mohammad R. Salavatipour*‡

Abstract

Clustering problems are well-studied in a variety of fields such as data science, operations research, and computer science. Such problems include variants of centre location problems, k-median, and k-means to name a few. In some cases, not all data points need to be clustered; some may be discarded for various reasons. For instance, some points may arise from noise in a data set or one might be willing to discard a certain fraction of the points to avoid incurring unnecessary overhead in the cost of a clustering solution.

We study clustering problems with outliers. More specifically, we look at UNCAPACITATED FACILITY LOCATION (UFL), k-MEDIAN, and k-MEANS. In these problems, we are given a set $\mathcal{X}$ of data points in a metric space $\delta(.,.)$, a set $\mathcal{C}$ of possible centres (each maybe with an opening cost), maybe an integer parameter k, plus an additional parameter z as the number of outliers. In UNCAPACITATED FACILITY LOCATION with outliers, we have to open some centres, discard up to z points of $\mathcal{X}$ and assign every other point to the nearest open centre, minimizing the total assignment cost plus centre opening costs. In k-MEDIAN and k-MEANS, we have to open up to k centres but there are no opening costs. In k-MEANS, the cost of assigning j to i is $\delta^2(j,i)$.

We present several results. Our main focus is on cases where δ is a doubling metric (this includes fixed dimensional Euclidean metrics as a special case) or is the shortest path metrics of graphs from a minor-closed family of graphs. For UNIFORM-COST UFL with outliers on such metrics we show that a multiswap simple local search heuristic yields a PTAS. With a bit more work, we extend this to bicriteria approximations for the k-MEDIAN and k-MEANS problems in the same metrics where, for any constant $\epsilon > 0$, we can find a solution using $(1+\epsilon)k$ centres whose cost is at most a $(1+\epsilon)$-factor of the optimum and uses at most z outliers. Our algorithms are all based on natural multiswap local search heuristics. We also show that natural local search heuristics that do not violate the number of clusters and outliers for k-MEDIAN (or k-MEANS) will have unbounded gap even in Euclidean metrics.

Furthermore, we show how our analysis can be extended to general metrics for k-MEANS with outliers to obtain a $(25+\epsilon, 1+\epsilon)$-approximation: an algorithm that uses at most $(1+\epsilon)k$ clusters and whose cost is at most $25+\epsilon$ of optimum and uses no more than z outliers.

1 Introduction

Clustering is a fundamental problem in the field of data analysis with a long history and a wide range of applications in very different areas, including data mining [9], image processing [41], biology [28], and database systems [17]. Clustering is the task of partitioning a given set of data points into clusters based on a specified similarity measure between the data points such that the points within the same cluster are more similar to each other than those in different clusters.

In a typical clustering problem, we are given a set of n data points in a metric space, and an integer k which specifies the desired number of clusters. We wish to find a set of k points to act as *centres* and then assign each point to its nearest centre, thereby forming k clusters. The quality of the clustering solution can be measured by using different objectives. For example, in the k-MEANS clustering (which is the most widely used clustering model), the goal (objective function) is to minimize the sum of squared distances of each data point to its centre, while in k-MEDIAN, the goal is to minimize the sum of distances of each data point to its centre. The UNCAPACITATED FACILITY LOCATION problem is the same as k-MEDIAN except that instead of a cardinality constraint bounding the number of centres, there is an additional cost for each centre included in the solution. Minimizing these objective functions exactly is NP-hard [2, 16, 21, 29, 38, 42], so there has been substantial work on obtaining provable upper bounds (approximability) and lower bounds (inapproximability) for these objectives; see [1, 10, 21, 29, 34, 35] for the currently best bounds. Although inapproximability

*Department of Computing Science, University of Alberta

†This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and an NSERC Discovery Grant.

‡Supported by NSERC.

results [21, 29, 34] prevent getting polynomial time approximation schemes (PTASs) for these problems in general metrics, PTASs are known for these problems in fixed dimensional Euclidean metrics [4, 14, 19]. Indeed, PTASs for k-MEDIAN and UNCAPACITATED FACILITY LOCATION in fixed dimension Euclidean space [4] have been known for almost two decades, but getting a PTAS for k-MEANS in fixed dimension Euclidean space had been an open problem until recent results of [14, 19].

In spite of the fact that these popular (centre based) clustering models are reasonably good for noise-free data sets, their objective functions (specially the k-MEANS objective function) are extremely sensitive to the existence of points far from cluster centres. Therefore, a small number of very distant data points, called *outliers*, –if not discarded– can dramatically affect the clustering cost and also the quality of the final clustering solution. Dealing with such outliers is indeed the main focus of this paper.

Clustering with outliers has a natural motivation in applications where outliers and unexpected data contained in the data may apply a strong influence over the clustering quality. In some specific applications, we need to find out the anomaly patterns (outliers) in the data for further investigation [26]. This finds applications, for example, in detecting fraudulent usage of credit cards – by monitoring transactions data in order to detect exceptional cases perhaps in type of purchase, location, timeframe, etc. – or in detecting suspicious trades in the equity markets, or in monitoring medical condition or more; see [26] for more details.

We restrict our attention to the outlier version of the three well studied clustering problems: k-MEANS with outliers (k-MEANS-OUT), k-MEDIAN with outliers (k-MEDIAN-OUT), and UNCAPACITATED FACILITY LOCATION with outliers (UFL-OUT). Formally, in these problems, we are given a set $\mathcal{X}$ of n data points in a metric space, a set $\mathcal{C}$ of possible centres, and the number of desired outliers z. Both k-MEANS-OUT and k-MEDIAN-OUT aim at finding k centres $C = \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and a set of (up to) z points Z to act as outliers. The objective is to minimize the clustering cost. In k-MEANS-OUT, this is the sum of squared distances of each data point in $\mathcal{X} \setminus Z$ to its nearest centre, i.e., $\sum_{x \in \mathcal{X} \setminus Z} \delta(x, C)^2$, while in k-MEDIAN-OUT this is just the sum of distances, i.e., $\sum_{x \in \mathcal{X} \setminus Z} \delta(x, C)$, where $\delta(x, C)$ indicates the distance of point x to its nearest centre in C. UFL-OUT is the same as k-MEDIAN-OUT except that instead of a cardinality constraint, we are given opening cost f_c for each centre $c \in \mathcal{C}$. The problem hence consists of finding centres (facilities) C and z outliers Z that minimizes $\sum_{x \in \mathcal{X} \setminus Z} \delta(x, C) + \sum_{c \in C} f_c$. We present PTASs for these problems on doubling metrics i.e. metrics with fixed doubling dimensions (which include fixed dimension Euclidean metrics as special case) and shortest path metrics of minor closed graphs (which includes planar graphs as special case)[1]. Recall that a metric (V, δ) has doubling dimension d if each ball of radius $2r$ can be covered with 2^d balls of radius r in V. We call it a *doubling metric* if d can be regarded as a constant; Euclidean metrics of constant (Euclidean) dimension are doubling metrics.

Despite a very large amount of work on the clustering problems, there has been only little work on their outlier versions. To the best of our knowledge, the clustering problem with outliers was introduced by Charikar et al. [11]. They devised a factor 3-approximation for UFL-OUT, and also a bicriteria $4(1 + \frac{1}{\epsilon})$-approximation scheme for k-MEDIAN-OUT that drops $z(1 + \epsilon)$ outliers. They obtained these results via some modifications of the Jain-Vazirani algorithm [30]. The first true approximation algorithm for k-MEDIAN-OUT was given by Chen [13] who obtained this by combining very carefully the Jain-Vazirani algorithm and local search; the approximation factor is not specified but seams to be a very large constant. Very recently, the first bicriterion approximation algorithm for k-MEANS-OUT is obtained by Gupta et al. [23]. They devised a bicriteria 274-approximation algorithm for k-MEANS-OUT that drops $O(kz \log(n\Delta))$ outliers, where Δ denotes the maximum distance between data points. This is obtained by a simple local search heuristic for the problem.

1.1 Related work k-MEANS is one of the most widely studied problems in the Computer Science literature. The problem is usually considered on d-dimensional Euclidean space $\mathbb{R}^d$, where the objective becomes minimizing the variance of the data points with respect to the centres they are assigned to. The most commonly used algorithm for k-MEANS is a simple heuristic known as Lloyd's algorithm (commonly referred to as the k-MEANS algorithm) [37]. Although this algorithm works well in practice, it is known that the the cost of the solutions computed by this algorithm can be arbitrarily large compared to the optimum solution [31]. Under some additional assumptions about the initially chosen centres, however, Arthur and Vassilvitskii [5] show that the approximation ratio of Lloyd's algorithm is $O(\log k)$. Later, Ostrovsky et al. [40] show that the approximation ratio is bounded by a constant if the input points obey some special properties. Under no such assumptions, Kanungo et al. [31] proved that a simple local search heuristic (that swaps only a constant

[1] For brevity, we will call such graphs *minor closed*, understanding this means they belong to a fixed family of graphs that closed under minors.

number of centres in each iteration) yields an $(9+\epsilon)$-approximation algorithm for Euclidean k-MEANS. Recently, Ahmadian et al. [1] improved the approximation ratio to $6.357+\epsilon$ by primal-dual algorithms. For general metrics, Gupta and Tangwongsan [22] proved that the local search algorithm is a $(25+\epsilon)$-approximation. This was also recently improved to $9+\epsilon$ via primal-dual algorithms [1].

In order to obtain algorithms with arbitrary small approximation ratios for Euclidean k-MEANS, many researchers restrict their focus on cases when k or d is constant. For the case when both k and d are constant, Inaba et al. [27] showed that k-MEANS can be solved in polynomial time. For fixed k (but arbitrary d), several PTASs have been proposed, each with some improvement over past results in terms of running time; e.g., see [15, 18, 24, 25, 32, 33]. Despite a large number of PTASs for k-MEANS with fixed k, obtaining a PTAS for k-MEANS in fixed dimensional Euclidean space had been an open problem for a along time. Bandyapadhyay and Varadarajan [8] presented a bicriteria PTAS for the problem that finds a $(1+\epsilon)$-approximation solution which might use up to $(1+\epsilon)k$ clusters. The first true PTAS for the problem was recently obtained by [14, 19] via local search. The authors show that their analysis also works for metrics with fixed doubling dimension [19], and the shortest path metrics of minor closed graphs [14].

There are several constant factor approximation algorithms for k-MEDIAN in general metrics. The simple local search (identical with the one for k-MEANS) is known to give a $3+\epsilon$ approximation by Arya et al. [6, 7]. The current best approximation uses different techniques and has an approximation ratio of $2.675+\epsilon$ [36, 10]. For Euclidean metrics, this was recently improved to $2.633+\epsilon$ via primal-dual algorithms [1]. Arora et al. [4], based on Arora's quadtree dissection [3], gave the first PTAS for k-MEDIAN in fixed dimensional Euclidean metrics. We note [4] also gives a PTAS for UFL-OUT and k-MEDIAN-OUT in constant-dimensional Euclidean metrics, our results for Euclidean metrics in particular are therefore most meaningful for k-MEANS-OUT. The recent PTASs (based on local search) for k-MEDIAN by [14, 19] work also for doubling metrics [19] and also for minor-closed metrics [14]. No PTAS or even a bicriteria PTAS was known for such metrics for even uniform-cost UFL with outliers (UNIFORM-UFL-OUT) or k-MEDIAN-OUT.

Currently, the best approximation for UNCAPACITATED FACILITY LOCATION in general metrics is a 1.488-approximation [35]. As with k-MEDIAN, PTASs are known for UNCAPACITATED FACILITY LOCATION in fixed dimensional Euclidean metrics [4], metrics with fixed doubling dimension [19], and for the shortest path metrics of minor closed graphs [14]; however the results by [14] only work for UNCAPACITATED FACILITY LOCATION with uniform opening cost.

1.2 Our results We present a general method for converting local search analysis for clustering problems without outliers to problems with outliers. Roughly speaking, we preprocess and then aggregate test swaps used in the analysis of such problems in order to incorporate outliers. We demonstrate this by applying our ideas to UNIFORM-UFL-OUT, k-MEDIAN-OUT, and k-MEANS-OUT.

A quick comment on the running times of our procedures is in order. In each theorem statement below, we mention that a ρ-swap local search algorithm provides some approximation where ρ is some constant (depending on ϵ and perhaps some other quantity like the dimension of the Euclidean space or the size of an excluded minor). This is an algorithm that tries all possible ways to close up to ρ centres from the local optimum and open up to ρ centres not currently in the local optimum. There are $|\mathcal{C}|^{O(\rho)}$ such swaps to consider, which is polynomial. The number of iterations is also polynomial in the input size when using the standard trick of performing a swap only if it improves by a $(1+\epsilon/|\mathcal{C}|)$-factor (mentioned in §2), so the overall algorithms run in polynomial time.

Most of our results are for metrics of fixed doubling dimensions as well as shortest path metrics of minor-closed graphs. First, we show that on such metrics a simple multi-swap local search heuristic yields a PTAS for UNIFORM-UFL-OUT.

THEOREM 1.1. *A $\rho=\rho(\epsilon,d)$-swap local search algorithm yields a PTAS for* UNIFORM-UFL-OUT *for doubling metrics and minor-closed graphs. Here, d is either the doubling constant of the metric or a constant that depends on a minor that is exclude from the minor-closed family.*

We then extend this result to k-MEDIAN and k-MEANS with outliers (k-MEDIAN-OUT and k-MEANS-OUT) and obtain bicriteria PTASs for them. More specifically:

THEOREM 1.2. *A $\rho=\rho(\epsilon,d)$-swap local search algorithm yields a bicriteria PTAS for* k-MEDIAN-OUT *and* k-MEANS-OUT *on doubling metrics and minor-closed graphs; i.e. finds a solutions of cost at most $(1+\epsilon)\cdot OPT$ with at most $(1+\epsilon)k$ clusters that uses at most z outliers where OPT is the cost of optimum k-clustering with z outliers.*

In fact, in minor-closed metrics a true local optimum in

the local search algorithm would find a solution using $(1+\epsilon)k$ clusters with cost at most OPT in both k-MEDIAN-OUT and k-MEANS-OUT, but a $(1+\epsilon)$-factor must be lost due to a standard procedure to ensure the local search algorithm terminates in polynomial time.

We show how these results can be extended to the setting where the metric is the ℓ_q^q-norm, i.e. cost of connecting two points i, j is $\delta^q(i,j)$ (e.g. k-MEDIAN-OUT is when $q=1$, k-MEANS-OUT is when $q=2$). Finally, we show that in general metrics we still recover bicriteria constant-factor approximation schemes for k-MEDIAN-OUT and k-MEANS-OUT. While not our main result, it gives much more reasonable constants under bicriteria approximations for these problems.

THEOREM 1.3. *A $1/\epsilon^{O(1)}$-swap local search algorithm finds a solution using $(1+\epsilon)k$ clusters and has cost at most $(3+\epsilon)\cdot OPT$ for k-MEDIAN-OUT or cost at most $(25+\epsilon)\cdot OPT$ for k-MEANS-OUT.*

It should be noted that a true constant-factor approximation for k-MEDIAN-OUT is given by Chen [12], though the constant seems to be very large. More interestingly, even a constant-factor bicriteria approximation scheme for k-MEANS-OUT that uses $(1+\epsilon)k$ clusters and discards the correct number of outliers has not been observed before (recall that the algorithm of [11] for k-MEDIAN-OUT has ratio $O(1/\epsilon)$ for k-MEDIAN using at most $(1+\epsilon)z$ outliers). It is not clear that Chen's algorithm can be extended to give a true constant-factor approximation for k-MEDIAN-OUT; one technical challenge is that part of the algorithm reassigns points multiple times over a series of $O(\log n)$ iterations. So it is not clear that the algorithm in [12] extends to k-MEANS.

To complement these results we show that for UFL-OUT (i.e. non-uniform opening costs) any multi-swap local search has unbounded gap. Also, for k-MEDIAN-OUT and k-MEANS-OUT we show that without violating the number of clusters or outliers, any multi-swap local search will have unbounded gap even on Euclidean metrics.

THEOREM 1.4. *Multi-swap local search has unbounded gap for* UFL-OUT *and for k-*MEDIAN-OUT *and k-*MEANS-OUT *on Euclidean metrics.*

Outline of the paper: We start with preliminaries and notation. Then in §2, we prove Theorem 1.1. In §3 we prove Theorem 1.2 for the case of k-MEDIAN on doubling metrics. Theorem 1.3 is proven in §4. Finally, the proof of Theorem 1.4 appears in §5.

The extension to ℓ_q^q norms is reasonably straightforward and follows arguments in [20]. The details are left to the full version of this paper.

1.3 Preliminaries and Notation In UNIFORM-UFL-OUT we are given a set of $\mathcal{X}$ points, a set $\mathcal{C}$ of centres and z for the number of outliers. Our goal is to select a set $C \subset \mathcal{C}$ to open and a set $Z \subset \mathcal{X}$ to discard and assign each $j \in \mathcal{X} - Z$ to the nearest centre in C to minimize $\sum_{x\in\mathcal{X}\setminus Z} \delta(x,C) + |C|$, where $\delta(x,C)$ is the distance of x to nearest $c \in C$. In k-MEDIAN-OUT and (discrete) k-MEANS-OUT, along with $\mathcal{X}$, $\mathcal{C}$, and z, we have an integer k as the number of clusters. We like to find k centres $C = \{c_1, \ldots, c_k\} \subseteq \mathcal{C}$ and a set of (up to) z points Z to act as outliers. In k-MEDIAN-OUT, we like to minimize $\sum_{x\in\mathcal{X}\setminus Z} \delta(x,C)$ and in k-MEANS-OUT, we want to minimize $\sum_{x\in\mathcal{X}\setminus Z} \delta(x,C)^2$.

For all these three problems, if we have the ℓ_q^q-norm then we like to minimize $\sum_{x\in\mathcal{X}\setminus Z} \delta^q(x,C)$. We should note that in classical k-MEANS (in $\mathbb{R}^d$), one is not given a candidate set of potential centres, but they can be chosen anywhere. However, by using the classical result of [39], at a loss of $(1+\epsilon)$ factor we can assume we have a set $\mathcal{C}$ of "candidate" centres from which the centres can be chosen (i.e. reduce to the discrete case considered here). This set can be computed in time $O(n\epsilon^{-d}\log(1/\epsilon))$ and $|\mathcal{C}| = O(n\epsilon^{-d}\log(1/\epsilon))$.

2 Uniform-Cost UFL with Outliers in Doubling Metrics

We start with presenting an approximation scheme for UNIFORM-UFL-OUT in doubling metrics (Theorem 1.1). Recall there is already a PTAS for UFL-OUT in constant-dimensional Euclidean metrics using dynamic programming for UNCAPACITATED FACILITY LOCATION through quadtree decompositions [4]. However, our approach generalizes to many settings where quadtree decompositions are not known to succeed such as when the assignment cost between a point j and a centre i is $\delta(j,i)^q$ for constant $1 < q < \infty$ including k-MEANS distances ($q=2$) and also to shortest-path metrics of edge-weighted minor-closed graphs. Furthermore, our technique here extends to k-MEANS-OUT (as seen in the next section). Still, we will initially present our approximation scheme in this simpler setting to lay the groundwork and introduce the main ideas.

Recall that we are given a set $\mathcal{X}$ of points and a set $\mathcal{C}$ of possible centres in a metric space with doubling dimension d and a number z bounding the number of admissible outliers. As the opening costs are uniform, we may scale all distances and opening costs so the opening cost of a centre is 1. For any $\emptyset \subsetneq \mathcal{S} \subseteq \mathcal{C}$, order the points $j \in \mathcal{X}$ as $j_1^{\mathcal{S}}, j_2^{\mathcal{S}}, \ldots, j_n^{\mathcal{S}}$ in nondecreasing order of distance $\delta(j_i^{\mathcal{S}}, \mathcal{S})$. The cost of $\mathcal{S}$ is then $\text{cost}(\mathcal{S}) := \sum_{\ell=1}^{n-z} \delta(j_\ell^{\mathcal{S}}, \mathcal{S}) + |\mathcal{S}|$. That is, after discarding the z points that are furthest from $\mathcal{S}$

the others are assigned to the nearest centre in $\mathcal{S}$: we pay this total assignment cost for all points that are not outliers and also the total centre opening cost $|\mathcal{S}|$. The goal is to find $\emptyset \subsetneq \mathcal{S} \subseteq \mathcal{C}$ minimizing cost($\mathcal{S}$).

Let $\epsilon > 0$ be a constant. Let $\rho' := \rho'(\epsilon, d)$ be some constant we will specify later. We consider a natural multiswap heuristic for UNIFORM-UFL-OUT, described in Algorithm 1.

Each iteration can be executed in time $|\mathcal{X}| \cdot |\mathcal{C}|^{O(\rho')}$. It is not clear that the number of iterations is bounded by a polynomial. However, the standard trick from [6, 31] works in our setting. That is, in the loop condition we instead perform the swap only if cost$((\mathcal{S} - Q) \cup P) \leq (1 - \epsilon/|\mathcal{C}|) \cdot$ cost($\mathcal{S}$). This ensures the running time is polynomial in the input size as every $|\mathcal{C}|/\epsilon$ iterations the cost decreases by a constant factor.

Our analysis of the local optimum follows the standard template of using test swaps to generate inequalities to bound cost(S). The total number of swaps we use to generate the final bound is at most $|\mathcal{C}|$, so (as in [6, 31]) the approximation guarantee of a local optimum will only be degraded by an additional $(1+\epsilon)$-factor. For the sake of simplicity in our presentation we will bound the cost of a local optimum solution returned by Algorithm 1.

2.1 Notation and Supporting Results from Previous Work We use many results from [19], so we use the same notation. In this section, these results are recalled and a quick overview of the analysis of the multiswap local search heuristic for UNIFORM-COST UFL is provided; this is simply Algorithm 1 where the cost function is defined appropriately for UNIFORM-COST UFL. Recall that in UNCAPACITATED FACILITY LOCATION, each centre $i \in \mathcal{C}$ has an *opening cost* $f_i \geq 0$. Also let cost$(\mathcal{S}) = \sum_{j \in \mathcal{X}} \delta(j, \mathcal{S}) + \sum_{i \in \mathcal{S}} f_i$. In UNIFORM-COST UFL all opening costs are uniform.

Let $\mathcal{S}$ be a local optimum solution returned by Algorithm 1 and $\mathcal{O}$ be a global optimum solution. As it is standard in local search algorithms for uncapacitated clustering problems, we may assume $\mathcal{S} \cap \mathcal{O} = \emptyset$. This can be assumed by duplicating each $i \in \mathcal{C}$, asserting $\mathcal{S}$ uses only the original centres and $\mathcal{O}$ uses only the copies. It is easy to see $\mathcal{S}$ would still be a local optimum solution. Let $\sigma : \mathcal{X} \to \mathcal{S}$ be the point assignment in the local optimum and $\sigma^* : \mathcal{X} \to \mathcal{O}$ be the point assignment in the global optimum. For $j \in \mathcal{X}$, let $c_j = \delta(j, \mathcal{S})$ and $c_j^* = \delta(j, \mathcal{O})$ (remember there are no outliers in this review of [19]).

For each $i \in \mathcal{S} \cup \mathcal{O}$, let D_i be the distance from i to the nearest centre of the other *type*. That is, for $i \in \mathcal{S}$ let $D_i = \delta(i, \mathcal{O})$ and for $i^* \in \mathcal{O}$ let $D_{i^*} = \delta(i^*, \mathcal{S})$. Note for every $j \in \mathcal{X}$ and every $i' \in \{\sigma(j), \sigma^*(j)\}$ that

$$D_{i'} \leq \delta(\sigma(j), \sigma^*(j)) \leq c_j^* + c_j. \tag{2.1}$$

A special pairing $\mathcal{T} \subseteq \mathcal{S} \times \mathcal{O}$ was identified in [19] (i.e. each $i \in \mathcal{S} \cup \mathcal{O}$ appears at most once among pairs in $\mathcal{T}$) with the following special property.

LEMMA 2.1. (LEMMA 3 IN [19], PARAPHRASED) *For any $A \subseteq \mathcal{S} \cup \mathcal{O}$ such that A contains at least one centre from every pair in $\mathcal{T}$, $\delta(i, A) \leq 5 \cdot D_i$ for every $i \in \mathcal{S} \cup \mathcal{O}$.*

Next, a *net* is cast around each $i \in \mathcal{S}$. The idea is that any swap that has $i \in \mathcal{S}$ being closed would have something open near every $i^* \in \mathcal{O}$ that itself is close to i. More precisely, [19] identifies a set $\mathcal{N} \subseteq \mathcal{S} \cup \mathcal{O}$ with the following properties. For each $i \in \mathcal{S}$ and $i^* \in \mathcal{O}$ with $\delta(i, i^*) \leq D_i/\epsilon$ and $D_{i^*} \geq \epsilon \cdot D_i$ there is some pair $(i, i') \in \mathcal{N}$ with $\delta(i', i^*) \leq \epsilon \cdot D_{i^*}$. The set $\mathcal{N}$ contains further properties to enable Theorem 2.1 (below), but these are sufficient for our discussion.

The last major step in [19] before the final analysis was to provide a structure theorem showing $\mathcal{S} \cup \mathcal{O}$ can be partitioned into test swaps that mostly allow the redirections discussed above.

THEOREM 2.1. (THEOREM 4 IN [19], SLIGHTLY ADJUSTED) *For any $\epsilon > 0$, there is a constant $\rho := \rho(\epsilon, d)$ and a randomized algorithm that samples a partitioning π of $\mathcal{O} \cup \mathcal{S}$ such that:*

- *For each part $P \in \pi$, $|P \cap \mathcal{O}|, |P \cap \mathcal{S}| \leq \rho$.*
- *For each part $P \in \pi$, $\mathcal{S} \triangle P$ includes at least one centre from every pair in $\mathcal{T}$.*
- *For each $(i^*, i) \in \mathcal{N}$,*
  Pr$[i, i^*$ lie in different parts of $\pi] \leq \epsilon$.

There is only a slight difference between this statement and the original statement in [19]. Namely, the first condition of Theorem 4 in [19] also asserted $|P \cap \mathcal{O}| = |P \cap \mathcal{S}|$. As noted at the end of [19], this part can be dropped by skipping one final step of the proof that "balanced" parts of the partition that were constructed (also see [20] for further details).

The analysis in [19] used Theorem 2.1 to show cost$(\mathcal{S}) \leq (1 + O(\epsilon)) \cdot OPT$ generated an inequality by swapping each part $P \in \pi$. Roughly speaking, for each $j \in \mathcal{X}$ with probability at least $1 - \epsilon$ (over the random construction of π), the swap that closes $\sigma(j)$ will open something very close to $\sigma(j)$ or very close to $\sigma^*(j)$. With the remaining probability, we can at least move j a distance of at most $O(c_j^* + c_j)$. Finally, if j was never moved to something that was close to $\sigma^*(j)$

Algorithm 1 UNIFORM-COST UFL ρ'-Swap Local Search

Let $\mathcal{S}$ be an arbitrary non-empty subset of $\mathcal{C}$
while $\exists$ sets $P \subseteq \mathcal{C} - \mathcal{S}$, $Q \subseteq \mathcal{S}$ with $|P|, |Q| \leq \rho'$ s.t. $\text{cost}((\mathcal{S} - Q) \cup P) < \text{cost}(\mathcal{S})$ **do**
  $\mathcal{S} \leftarrow (\mathcal{S} - Q) \cup P$
return $\mathcal{S}$

this way then we ensure we move j from $\sigma(j)$ to $\sigma^*(j)$ when $\sigma^*(j)$ is swapped in.

In our analysis for clustering with outliers, our reassignments for points that are not outliers in either the local or global optimum are motivated by this approach. Details will appear below in our analysis of Algorithm 1.

2.2 Analysis for Uniform-Cost UFL with Outliers: An Outline Now let $\mathcal{S}$ be a locally optimum solution for Algorithm 1, let $\mathcal{X}^a$ be the points in $\mathcal{X}$ that are assigned to $\mathcal{S}$ and $\mathcal{X}^o$ be the points in $\mathcal{X}$ that are outliers when opening $\mathcal{S}$. Similarly, let $\mathcal{O}$ be a globally optimum solution, let $\mathcal{X}^{a^*}$ be the points in $\mathcal{X}$ that are assigned to $\mathcal{O}$ and $\mathcal{X}^{o^*}$ be the points in $\mathcal{X}$ that are outliers when opening $\mathcal{O}$. Note $|\mathcal{X}^o| = |\mathcal{X}^{o^*}| = z$.

Let $\sigma : \mathcal{X} \to \mathcal{S} \cup \{\perp\}$ assign $j \in \mathcal{X}^a$ to the nearest centre in $\mathcal{S}$ and $j \in \mathcal{X}^o$ to $\perp$. Similarly, let $\sigma^* : \mathcal{X} \to \mathcal{O} \cup \{\perp\}$ map each $j \in \mathcal{X}^{a^*}$ to the nearest centre in $\mathcal{O}$ and each $j \in \mathcal{X}^{o^*}$ to $\perp$. For $j \in \mathcal{X}$, we let $c_j = 0$ if $j \in \mathcal{X}^o$ and, otherwise, let $c_j = \delta(j, \mathcal{S}) = \delta(j, \sigma(j))$. Similarly, let $c_j^* = 0$ if $j \in \mathcal{X}^{o^*}$ and, otherwise, let $c_j^* = \delta(j, \mathcal{O}) = \delta(j, \sigma(j))$.

Our starting point is the partitioning scheme described in Theorem 2.1. The new issue to be handled is in reassigning the outliers when a part is swapped. That is, for any $j \in \mathcal{X}^{o^*}$ any swap that has $\sigma(j)$ swapped out cannot, in general, be reassigned anywhere cheaply.

Really the only thing we can do to upper bound the assignment cost change for j is to make it an outlier. We can try assigning each $j \in \mathcal{X}^o$ to $\sigma^*(j)$ if it is opened, thereby allowing one $j \in \mathcal{X}^{o^*}$ with $\sigma(j)$ being swapped out to become an outlier. However there may not be enough $j' \in \mathcal{X}^o$ that have $\sigma^*(j)$ opened after the swap. That is, we might not remove enough outliers from the solution $\mathcal{S}$ to be able to let all such j become outliers.

Our approach is to further combine parts P of the partition π and perform the swaps simultaneously for many of these parts. Two of these parts in a larger grouping will not actually be swapped out: their centres in $\mathcal{O}$ will be opened to free up more spaces for outliers yet their centres in $\mathcal{S}$ will not be swapped out. Doing this carefully, we ensure that the total number of $j \in \mathcal{X}^{o^*}$ that have $\sigma(j)$ being closed is at most the total number of $j \in \mathcal{X}^o$ that have $\sigma^*(j)$ being opened.

These larger groups that are obtained by combing parts of π are not disjoint. However, the overlap of centres between larger groups will be negligible compared to $|\mathcal{S}| + |\mathcal{O}|$.

2.3 Grouping the Parts We will assume $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$. This is without loss of generality as $\mathcal{S}$ would still be a local optimum in the instance with $\mathcal{X}^o \cap \mathcal{X}^{o^*}$ removed and z adjusted. Recall we are also assuming $\mathcal{S} \cap \mathcal{O} = \emptyset$.

For each part P of π, let $\Delta_P := |\{j \in \mathcal{X}^o : \sigma^*(j) \in P\}| - |\{j \in \mathcal{X}^{o^*} : \sigma(j) \in P\}|$. This is the difference between the number of outliers we can reclaim by moving them to $\sigma^*(j)$ (if it is open after swapping P) and the number of outliers j that we must create because $\sigma(j)$ was closed when swapping P.

Consider the following refinements of π: $\pi^+ = \{P \in \pi : \Delta_P > 0\}, \pi^- = \{P \in \pi : \Delta_P < 0\}$ and $\pi^0 = \{P \in \pi : \Delta_P = 0\}$. Intuitively, nothing more needs to be done to prepare parts $P \in \pi^0$ for swapping as this would create as many outliers as it would reclaim in our analysis framework. We work toward handling π^+ and π^-.

Next we construct a bijection $\kappa : \mathcal{X}^o \to \mathcal{X}^{o^*}$. We will ensure when $\sigma(j)$ is swapped out for some $j \in \mathcal{X}^{o^*}$ that $\sigma^*(\kappa^{-1}(j))$ will be swapped in. So there is space to make j an outlier in the analysis. There are some cases in our analysis where we never swap out $\sigma(j)$ for some $j \in \mathcal{X}^{o^*}$, but we will still ensure $\sigma^*(\kappa^{-1}(j))$ is swapped in at some point so we can still make j an outlier while removing $\kappa^{-1}(j)$ as an outlier to get the negative dependence on c_j in the final inequality.

To start defining κ, for each $P \in \pi$ we pair up points in $\{j \in \mathcal{X}^{o^*} : \sigma(j) \in P\}$ and $\{j \in \mathcal{X}^o : \sigma^*(j) \in P\}$ arbitrarily until one of these two groups is exhausted. These pairs define some mapping of κ. The number of unpaired points in $\{j \in \mathcal{X}^{o^*} : \sigma(j) \in P\}$ is exactly $-\Delta_P$ if $\Delta_P < 0$ and the number of unpaired points in $\{j \in \mathcal{X}^o : \sigma^*(j) \in P\}$ is exactly Δ_P.

Having done this for each P, we begin pairing unpaired points in $\mathcal{X}^o \cup \mathcal{X}^{o^*}$ between parts. Arbitrarily order π^+ as $P_1^+, P_2^+, \ldots, P_m^+$ and π^- as $P_1^-, P_2^-, \ldots, P_\ell^-$. We will complete the pairing κ and also construct edges in a bipartite graph H with π^+ on one side and π^- on the other side using Algorithm 2. To avoid confusion with edges in the distance metric, we call the edges of H between π^+ and π^- *superedges*.

In Algorithm 2, we say a part $P \in \pi^+ \cup \pi^-$ has an unpaired point $j \in \mathcal{X}^{o^*} \cup \mathcal{X}^o$ and that j is an unpaired point of P if, currently, κ has not paired j and $\sigma(j) \in P$ or $\sigma^*(j) \in P$ (whatever is relevant). The resulting graph over π^+, π^- is depicted in Figure 1, along with other features described below.

We now group some superedges together. Let $\alpha = 4\rho/\epsilon$ denote the *group size*. We will assume $|\mathcal{P}| > \alpha$, as otherwise $|\pi^+ \cup \pi^-| \leq 2\alpha$ and we could simply merge all parts in $\pi^+ \cup \pi^-$ into a single part P' with $\Delta_{P'} = 0$ with $|P' \cap \mathcal{S}|, |P' \cap \mathcal{O}| \leq 2\alpha\rho$. The final local search algorithm will use swap sizes greater than $2\alpha\rho$ and the analysis showing $\text{cost}(\mathcal{S}) \leq (1 + O(\epsilon)) \cdot \text{cost}(\mathcal{O})$ would then follow almost exactly as we show[2].

Order edges $e_0, e_1, \ldots, e_{|\mathcal{P}|-1}$ of $\mathcal{P}$ according to when they were formed. For each integer $s \geq 0$ let $E_s = \{e_i : \alpha \cdot s \leq i < \alpha \cdot (s+1)\}$. Let s' be the largest index with $E_{s'+1} \neq \emptyset$ (which exists by the assumption $|\mathcal{P}| > \alpha$). Merge the last two groups by replacing $E_{s'}$ with $E_{s'} \cup E_{s'+1}$. Finally, for each $0 \leq s \leq s'$ let $G_s \subseteq \mathcal{S} \cup \mathcal{O}$ consist of all centres $i \in \mathcal{S} \cup \mathcal{O}$ belonging to a part P that is an endpoint of some superedge in E_s. The grouping is $\mathcal{G} = \{G_s : 0 \leq s \leq s'\}$. The groups of superedges E_s are depicted in Figure 1. Note each part $P \in \pi^+ \cup \pi^-$ is contained in at least one group of $\mathcal{G}$ because $\Delta_P \neq 0$ (so a superedge edge in $\mathcal{P}$ was created with P as an endpoint). The following lemma is easy and the proof is deferred to the full version.

LEMMA 2.2. *For each $G_s \in \mathcal{G}$, $\alpha - 1 \leq |G_s| \leq 8\rho\alpha$.*

Note: In fact, it is not hard to see that H is a forest where each component of it is a caterpillar (a tree in which every vertex is either a leaf node or is adjacent to a "stalk" node; stalk nodes form a path). The parts P that are split between different groups G_s are the high degree nodes in H and these parts belong to the "stalk" of H.

DEFINITION 2.1. *For each group $G_s \in \mathcal{G}$, say a part $P \in \pi$ is split by G_s if $P \subseteq G_s$ and $N_{\mathcal{P}}(P) \not\subseteq E_s$, where $N_{\mathcal{P}}(P)$ are those parts that have a superedge to P.*

We simply say P is split if the group G_s is clear from the context. The following lemma is also easy and is deferred to the full version.

LEMMA 2.3. *For each $G_s \in \mathcal{G}$, there are at most two parts P split by G_s.*

[2]One very minor modification is that the presented proof of Lemma 2.5 in the final analysis ultimately uses $|\mathcal{P}| \geq \alpha$. It still holds otherwise in a trivial way as there would be no overlap between groups.

2.4 Analyzing Algorithm 1 Now suppose we run Algorithm 1 using $\rho' = 4\alpha\rho$. For each $P \in \pi^0$, extend $\mathcal{G}$ to include a "simple" group $G_s = P$ for each $P \in \pi^0$ where s is the next unused index. Any $P \in \pi^0$ is not split by any group. Each $P \in \pi$ is then contained in at least one group of $\mathcal{G}$ and $|G_s \cap \mathcal{S}|, |G_s \cap \mathcal{O}| \leq \rho'$ for each $G_s \in \mathcal{G}$ by Lemma 2.2.

We are now ready to describe the swaps used in our analysis of Algorithm 1. Simply, for $G_s \in \mathcal{G}$ let $\mathcal{S}_s$ be the centres in $G_s \cap \mathcal{S}$ that are not in a part P that is split and let $\mathcal{O}_s$ simply be $G_s \cap \mathcal{O}$. We consider the swap $\mathcal{S} \to (\mathcal{S} - \mathcal{S}_s) \cup \mathcal{O}_s$.

To analyze these swaps, we further classify each $j \in \mathcal{X}^a \cap \mathcal{X}^{a*}$ in one of four ways. This is the same classification from [19]. Label j according to the first property that it satisfies below.

- **lucky**: both $\sigma(j)$ and $\sigma^*(j)$ in the same part P of π.
- **long**: $\delta(\sigma(j), \sigma^*(j)) > D_{\sigma(j)}/\epsilon$.
- **good**: either $D_{\sigma^*(j)} \leq \epsilon D_{\sigma(j)}$ or there is some $i' \in \mathcal{O}$ with $\delta(\sigma^*(j), i') \leq \epsilon \cdot D_{\sigma^*(j)}$ and $(\sigma(j), i') \in \mathcal{N}$ where both $\sigma(j)$ and i' lie in the same part.
- **bad**: j is not lucky, long, or good. Note, by Theorem 2.1, that $\Pr[j \text{ is bad}] \leq \epsilon$ over the random construction of π.

Finally, as a technicality for each $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ where $\sigma^*(j)$ lies in a part that is split by some group and j is either lucky or long, let $s(j)$ be any index such that group $G_{s(j)} \in \mathcal{G}$ contains the part with $\sigma^*(j)$. The idea is that we will reassign j to $\sigma^*(j)$ only when group $G_{s(j)}$ is processed.

Similarly, for any $j \in \mathcal{X}^o$, let $s(j)$ be any index such that $\sigma^*(j), \sigma(\kappa(j)) \in G_{s(j)}$. The following lemma shows this is always possible.

LEMMA 2.4. *For each $j \in \mathcal{X}^o$ there is at least one group $G_s \in \mathcal{G}$ where $\sigma^*(j), \sigma(\kappa(j)) \in G_s$.*

Proof. If $\sigma^*(j)$ and $\sigma(\kappa(j))$ lie in the same part P, this holds because each part is a subset of some group. Otherwise, j is paired with $\kappa(j)$ at some point in Algorithm 2 and an edge (P_a^+, P_b^-) is added to $\mathcal{P}$ where $\sigma^*(j) \in P_a^+, \sigma(\kappa(j)) \in P_b^-$. The centres in both endpoints of this super edge were added to some group G_s.

We now place a bound on the cost change in each swap. Recall $|G_s \cap \mathcal{S}|, |G_s \cap \mathcal{O}| \leq \rho'$ and $\mathcal{S}$ is a local optimum, so

$$0 \leq \text{cost}((\mathcal{S} \triangle \mathcal{S}_s) \cup \mathcal{O}_s) - \text{cost}(\mathcal{S}).$$

Algorithm 2 Pairing Unpaired Outliers

$\mathcal{P} \leftarrow \emptyset$ ▷ A set of superedges between π^+ and π^-.
$a \leftarrow 1, b \leftarrow 1$
while there are unpaired points **do**
  Arbitrarily pair up (via κ) unpaired points between P_a^+ and P_b^- until no longer possible.
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{(P_a^+, P_b^-)\}$.
  if P_a^+ has no unpaired point **then**
    $a \leftarrow a + 1$
  if P_b^- has no unpaired point **then**
    $b \leftarrow b + 1$
return $\mathcal{S}$

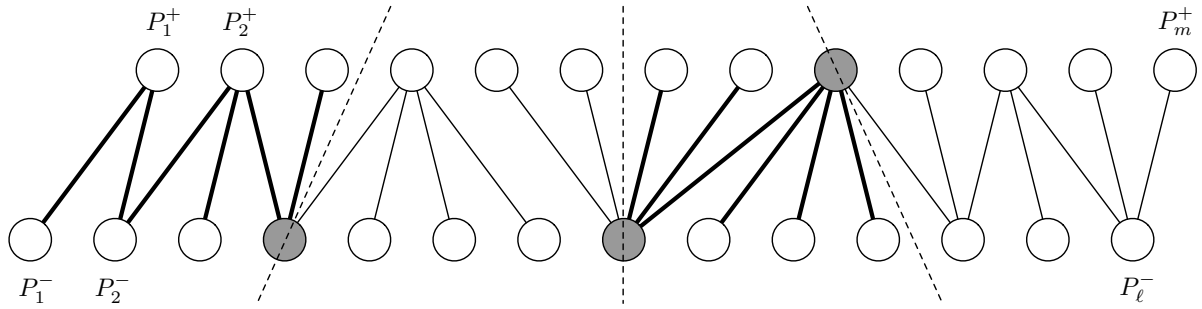

Figure 1: The bipartite graph with sides π^+, π^- and edges $\mathcal{P}$. The edges are ordered left-to-right in order of their creation time, so e_0 is the leftmost edge. A grouping with group size $\alpha = 6$ is depicted. The edges in E_1 and E_3 are bold. Note the last group has more than α edges. The parts that are split by some group are shaded. While not depicted, a part could be split by many groups (if it has very high degree in the bipartite graph).

We describe a feasible reassignment of points to upper bound the cost change. This may cause some points in $\mathcal{X}^a$ becoming outliers and other points in $\mathcal{X}^o$ now being assigned. We take care to ensure the number of points that are outliers in our reassignment is exactly z, as required.

Consider the following instructions describing one possible way to reassign a point $j \in \mathcal{X}$ when processing G_s. This may not describe an optimal reassignment, but it places an upper bound on the cost change. First we describe which points should be moved to $\sigma^*(j)$ if it becomes open. Note that the points in $\mathcal{X}^o \cup \mathcal{X}^{o^*}$ are paired via κ (and $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$). Below we specify what to do for each point $j \in \mathcal{X}^o$ and $\kappa(j)$ together.

- If $j \in \mathcal{X}^o$ and $s = s(j)$, then make $\kappa(j)$ an outlier and connect j to $\sigma^*(j)$, which is now open. The total assignment cost change for j and $\kappa(j)$ is $c_j^* - c_{\kappa(j)}$.

Note that so far this reassignment still uses z outliers because each new outlier j' has its paired point $\kappa^{-1}(j')$ that used to be an outlier become connected. The rest of the analysis will not create any more outliers. The rest of the cases are for when $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$.

- If j is lucky or long and $s = s(j)$, reassign j from $\sigma(j)$ to $\sigma^*(j)$. The assignment cost change for j is $c_j^* - c_j$.
- If j is long, move j to the open centre that is nearest to $\sigma(j)$. By Lemma 2.1 and because j is long, the assignment cost increase for j can be bounded as follows:

$$5 \cdot D_{\sigma(j)} \le 5\epsilon \cdot \delta(\sigma(j), \sigma^*(j)) \le 5\epsilon \cdot (c_j^* + c_j).$$

- If j is good and $D_{\sigma^*(j)} \le \epsilon \cdot D_{\sigma(j)}$, move j to the open centre that is nearest to $\sigma^*(j)$. By (2.1) in §2.1 and Lemma 2.1, the assignment cost increase for j can be bounded as follows:

$$\begin{aligned} & c_j^* + 5 \cdot D_{\sigma^*(j)} - c_j \\ \le\ & c_j^* + 5\epsilon \cdot D_{\sigma(j)} - c_j \\ \le\ & c_j^* + 5\epsilon \cdot (c_j^* + c_j) - c_j \\ =\ & (1 + 5\epsilon) \cdot c_j^* - (1 - 5\epsilon) \cdot c_j. \end{aligned}$$

- If j is good but $D_{\sigma^*(j)} > \epsilon \cdot D_{\sigma(j)}$, then let i' be such that $\sigma(j), i'$ both lie in G_s and $\delta(\sigma^*(j), i') \le$

$\epsilon \cdot D_{\sigma^*(j)}$. Reassigning j from to i' bounds its assignment cost change by

$$\begin{aligned} & c_j^* + \delta(\sigma^*(j), i') - c_j \\ \leq\ & c_j^* + \epsilon \cdot D_{\sigma^*(j)} - c_j \\ \leq\ & (1+\epsilon) \cdot c_j^* - (1-\epsilon) \cdot c_j. \end{aligned}$$

- Finally, if j is bad then simply reassign j to the open centre that is nearest to $\sigma(j)$. By (2.1) and Lemma 2.1, the assignment cost for j increases by at most $5 \cdot D_{\sigma(j)} \leq 5 \cdot (c_j^* + c_j)$.

  This looks large, but its overall contribution to the final analysis will be scaled by an ϵ-factor because a point is bad only with probability at most ϵ over the random sampling of π.

Note this accounts for all points j where $\sigma(j)$ is closed. Every other point j may stay assigned to $\sigma(j)$ to bound its assignment cost change by 0.

For $j \in \mathcal{X}$, let Δ_j denote the total reassignment cost change over all swaps when moving j as described above. This should not be confused with the previously-used notation Δ_P for a part $P \in \pi$. We bound Δ_j on a case-by-case basis below.

- If $j \in \mathcal{X}^o$ then the only time j is moved is for the swap involving $G_{s(j)}$. So $\Delta_j = c_j^*$.
- If $j \in \mathcal{X}^{o^*}$ then the only time j is moved is for the swap involving $G_{s(\kappa^{-1}(j))}$. So $\Delta_j = -c_j$.
- If j is lucky then it is only moved when $G_{s(j)}$ is processed so $\Delta_j = c_j^* - c_j$.
- If j is long then it is moved to $\sigma^*(j)$ when $G_{s(j)}$ is processed and it is moved near $\sigma(j)$ when $\sigma(j)$ is closed, so $\Delta_j \leq c_j^* - c_j + 5\epsilon \cdot (c_j^* + c_j) = (1+5\epsilon) \cdot c_j^* - (1-5\epsilon) \cdot c_j$.
- If j is good then it is only moved when $\sigma(j)$ is closed so $\Delta_j \leq (1+5\epsilon) \cdot c_j^* - (1-5\epsilon) \cdot c_j$.
- If j is bad then it is only moved when $\sigma(j)$ is closed so $\Delta_j \leq 5 \cdot (c_j^* + c_j)$.

To handle the centre opening cost change, we use the following fact.

LEMMA 2.5. $\sum_{G_s \in \mathcal{G}} |\mathcal{O}_s| - |\mathcal{S}_s| \leq (1+2\epsilon) \cdot |\mathcal{O}| - (1-2\epsilon) \cdot |\mathcal{S}|$

Proof. For each $G_s \in \mathcal{G}$, let P_s be the union of all parts used to form G_s that are not split by G_s. Let $\overline{P}_s = G_s - P_s$, these are centres in G_s that lie in a part split by G_s.

Now, $|\overline{P}_s| \leq 2\rho$ because at most two parts are split by G_s by Lemma 2.3. On the other hand, by Lemmas 2.2 and 2.3 there are at least $\alpha - 3$ parts that were used to form G_s that were not split by G_s. As each part contains at least one centre, then $|P_s| \geq \alpha - 3$. Thus, for small enough ϵ we have

$$|\overline{P}_s| \leq 2\rho \leq \epsilon \cdot (\alpha - 3) \leq \epsilon |P_s|.$$

Note $\sum_{G_s \in \mathcal{G}} |P_s| \leq |\mathcal{S}| + |\mathcal{O}|$ because no centre appears in more than one set of the form P_s. Also note $|\mathcal{O}_s| \leq |P_s \cap \mathcal{O}| + |\overline{P}_s|$ and $|\mathcal{S}_s| \geq |G_s \cap \mathcal{S}| - |\overline{P}_s|$,

$$\begin{aligned} \sum_{G_s \in \mathcal{G}} |\mathcal{O}_s| - |\mathcal{S}_s| &\leq \sum_{G_s \in \mathcal{G}} |P_s \cap \mathcal{O}| - |G_s \cap \mathcal{S}| + 2\epsilon |P_s| \\ &\leq |\mathcal{O}| - |\mathcal{S}| + 2\epsilon \cdot (|\mathcal{O}| + |\mathcal{S}|). \end{aligned}$$

Putting this all together,

$$\begin{aligned} 0 \leq\ & \sum_{G_s \in \mathcal{G}} \text{cost}((\mathcal{S} - \mathcal{S}_s) \cup \mathcal{O}_s) - \text{cost}(\mathcal{S}) \\ \leq\ & \sum_{j \in \mathcal{X}} \Delta_j + \sum_{G_s \in \mathcal{G}} |\mathcal{O}_s| - |\mathcal{S}_s| \\ \leq\ & \sum_{\substack{j \in \mathcal{X}^a \cap \mathcal{X}^{a^*} \\ j \text{ is not bad}}} \left[(1+5\epsilon) \cdot c_j^* - (1-5\epsilon) c_j\right] + \\ & \sum_{j \text{ bad}} 5(c_j^* + c_j) + \sum_{j \in \mathcal{X}^o} c_j^* - \sum_{j \in \mathcal{X}^{o^*}} c_j + \\ & (1+2\epsilon) \cdot |\mathcal{O}| - (1-2\epsilon) \cdot |\mathcal{S}|. \end{aligned}$$

This holds for any π supported by the partitioning scheme described in Theorem 2.1. Taking expectations over the random choice of π and recalling $\Pr[j \text{is bad}] \leq \epsilon$ for any $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$,

$$\begin{aligned} 0 \leq\ & \sum_{j \in \mathcal{X}} \left[(1+10\epsilon) \cdot c_j^* - (1-10\epsilon) \cdot c_j\right] \\ & + (1+2\epsilon) \cdot |\mathcal{O}| - (1-2\epsilon) \cdot |\mathcal{S}|. \end{aligned}$$

Rearranging and relaxing slightly further shows

$$(1-10\epsilon) \cdot \text{cost}(\mathcal{S}) \leq (1+10\epsilon) \cdot \text{cost}(\mathcal{O}).$$

Ultimately, $\text{cost}(\mathcal{S}) \leq (1+30\epsilon) \cdot OPT$.

3 k-Median and k-Means with Outliers

In this section we show how the results of the previous section can be extended to get a bicriteria approximation scheme for k-MEDIAN-OUT and k-MEANS-OUT with outliers in doubling metrics (Theorem 1.2). For ease of exposition we present the result for k-MEDIAN-OUT. More specifically, given a set $\mathcal{X}$ of points, set $\mathcal{C}$ of possible centers, positive integers k as the number of clusters

and z as the number of outliers for k-MEDIAN, we show that a ρ'-swap local search (for some $\rho' = \rho(\epsilon, d)$ to be specified) returns a solution of cost at most $(1+\epsilon)OPT$ using at most $(1+\epsilon)k$ centres (clusters) and has at most z outliers. Note that a local optimum $\mathcal{S}$ satisfies $|\mathcal{S}| = (1+\epsilon)\cdot k$ unless $\text{cost}(\mathcal{S}) = 0$ already, in which case our analysis is already done. Extension of the result ot k-MEANS or in general to a clustering where the distance metric is the ℓ_q^q-norm is fairly easy and is discussed in the next section where we also show how we can prove the same result for shortest path metric of minor-closed families of graphs.

The proof uses ideas from both [19] for the PTAS for k-MEANS as well as the results of the previous section for UNIFORM-UFL-OUT. Let $\mathcal{S}$ be a local optimum solution returned by Algorithm 3 that uses at most $(1+\epsilon)k$ clusters and has at most z outliers and let $\mathcal{O}$ be a global optimum solution to k-MEDIAN-OUT with k clusters and z outliers. Again, we use $\sigma : \mathcal{X} \to \mathcal{S} \cup \{\perp\}$ and $\sigma^* : \mathcal{X} \to \mathcal{O} \cup \{\perp\}$ as the assignment of points to cluster centres, where $\sigma(j) = \perp$ (or $\sigma^*(j) = \perp$) indicates an outlier. For $j \in \mathcal{X}$ we use $c_j = \delta(j, \mathcal{S})$ and $c_j^* = \delta(j, \mathcal{O})$. The special pairing $\mathcal{T} \subseteq \mathcal{S} \times \mathcal{O}$ and Lemma 2.1 as well as the notion of nets $\mathcal{N} \subseteq \mathcal{S} \cup \mathcal{O}$ are still used in our setting. A key component of our proof is the following slightly modified version of Theorem 4 in [19]:

THEOREM 3.1. *For any $\epsilon > 0$, there is a constant $\rho = \rho(\epsilon, d)$ and a randomized algorithm that samples a partitioning π of $\mathcal{S} \cup \mathcal{O}$ such that:*

- *For each part $P \in \pi$, $|P \cap \mathcal{O}| < |P \cap \mathcal{S}| \le \rho$*
- *For each part $P \in \pi$, $S \Delta P$ contains at least one centre from every pair in $\mathcal{T}$*
- *For each $(i^*, i) \in \mathcal{N}$, $\Pr[i, i^* \text{ lie in different parts of } \pi] \le \epsilon$.*

The only difference of this version and the one in [19] is that in Theorem 4 in [19], for the first condition we have $|P \cap \mathcal{O}| = |P \cap \mathcal{S}| \le \rho$. The theorem was proved by showing a randomized partitioning that satisfies conditions 2 and 3. To satisfy the 1st condition $|P \cap \mathcal{O}| = |P \cap \mathcal{S}| \le \rho$ a balancing step was performed at the end of the proof that would merge several (but constant) number of parts to get parts with equal number of centres from $\mathcal{S}$ and $\mathcal{O}$ satisfying the two other conditions. Here, since $|\mathcal{S}| = (1+\epsilon)k$ and $|\mathcal{O}| = k$, we can modify the balancing step to make sure that each part has at least one more centre from $\mathcal{S}$ than from $\mathcal{O}$, details of this procedure appear in the full version.

We define $\mathcal{X}^a, \mathcal{X}^o, \mathcal{X}^{a^*}$, and $\mathcal{X}^{o^*}$ as in the previoius section for UNIFORM-UFL-OUT. Note that $|\mathcal{X}^o| = |\mathcal{X}^{o^*}| = z$. As before, we assume $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$ and $\mathcal{S} \cap \mathcal{O} = \emptyset$. Let π be a partitioning as in Theorem 3.1. Recall that for each $P \in \pi$, $\Delta_P := |\{j \in \mathcal{X}^o : \sigma^*(j) \in P\}| - |\{j \in \mathcal{X}^{o^*} : \sigma(j) \in P\}|$; we define π^+, π^-, π^0 to be the parts with positive, negative, and zero Δ_P values. We define the bijection $\kappa : \mathcal{X}^o \to \mathcal{X}^{o^*}$ as before: for each $P \in \pi$, we pair up points (via κ) in $\{j \in \mathcal{X}^{o^*} : \sigma(j) \in P\}$ and $\{j \in \mathcal{X}^o : \sigma^*(j)\} \in P$ arbitrarily. Then (after this is done for each P) we begin pairing unpaired points in $\mathcal{X}^o \cup \mathcal{X}^{o^*}$ between groups using Algorithm 2. The only slight change w.r.t. the previous section is in the grouping the super-edges of the bipartite graph defined over $\pi^+ \cup \pi^-$: we use parameter $\alpha = 2\rho + 3$ instead. Lemmas 2.2 and 2.3 still hold.

Suppose we run Algorithm 3 with $\rho' = \alpha\rho$. As in the case of UNIFORM-UFL-OUT, we add each $P \in \pi^0$ as a separate group to $\mathcal{G}$ and so each $P \in \pi$ is now contained in at least one group $G_s \in \mathcal{G}$ and $|G_s \cap \mathcal{S}|, |G_s \cap \mathcal{O}| \le \rho'$. For the points $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ (i.e. those that are not outliner in neither $\mathcal{S}$ nor $\mathcal{O}$) the analysis is pretty much the same as the PTAS for standard k-MEANS (without outliers) in [19]. We need extra care to handle the points j that are outliers in one of $\mathcal{S}$ or $\mathcal{O}$ (recall that $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$.

As in the analysis of UNIFORM-UFL-OUT, for $G_s \in \mathcal{G}$, let $\mathcal{S}_s$ be the centers in $G_s \cap \mathcal{S}$ that are not in a part P that is split, and $\mathcal{O}_s = G_s \cap \mathcal{O}$; consider the test swap that replaces $\mathcal{S}$ with $(\mathcal{S} - \mathcal{S}_s) \cup \mathcal{O}_s$. To see this is a valid swap considered by our algorithm, note that the size of a test swap is at most ρ'. Furthremore, there are at least $\alpha - 3$ unsplit parts P and at most two split parts in each G_s, and each unsplit part has at least one more centre from $\mathcal{S}$ than $\mathcal{O}$ (condition 1 of Theorem 3.1); therefore, there are at least $\alpha - 3$ more centres that are swapped out in $\mathcal{S}_s$ and these can account for the at most 2ρ centres of $\mathcal{S} \cup \mathcal{O}$ in the split parts that are not swapped out (note that $\alpha - 3 = 2\rho$). Thus, the total number of centres of $\mathcal{S}$ and $\mathcal{O}$ after the test swap is still at most $(1+\epsilon)k$ and k, respectively.

We classify each $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ to **lucky**, **long**, **good**, and **bad** in the same way as in the case of UNIFORM-UFL-OUT. Furthermore, $s(j)$ is defined in the same manner: for each $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ where $\sigma^*(j)$ lies a part that is split by some group and j is either lucky or long, let $s(j)$ be any index such that group $G_{s(j)} \in \mathcal{G}$ contains the part with $\sigma^*(j)$. Similarly, for any $j \in \mathcal{X}^o$ let $s(j)$ be any index such that $\sigma^*(j), \sigma(\kappa(j)) \in G_{s(j)}$ (note that since $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$, if $j \in \mathcal{X}^o$ then $j \in \mathcal{X}^{a^*}$). Lemma 2.4 still holds. For each $G_s \in \mathcal{G}$, since $|G_s \cap \mathcal{S}|, |G_s \cap \mathcal{O}| \le \rho'$ and $\mathcal{S}$ is a local optimum, any test swap based on a group G_s is not improving, hence $0 \le \text{cost}((\mathcal{S} \triangle \mathcal{S}_s) \cup \mathcal{O}_s) - \text{cost}(\mathcal{S})$.

For each test swap G_s, we describe how we could

Algorithm 3 k-MEDIAN ρ'-Swap Local Search

Let $\mathcal{S}$ be an arbitrary set of $(1+\epsilon)k$ centres from $\mathcal{C}$
while $\exists$ sets $P \subseteq \mathcal{C} - \mathcal{S}$, $Q \subseteq \mathcal{S}$ with $|P|, |Q| \leq \rho'$ s.t. $\text{cost}((\mathcal{S} - Q) \cup P) < \text{cost}(\mathcal{S})$ and $|(\mathcal{S} - Q) \cup P| \leq (1+\epsilon)k$ **do**
  $\mathcal{S} \leftarrow (\mathcal{S} - Q) \cup P$
return $\mathcal{S}$

re-assign the each point j for which $\sigma(j)$ becomes closed or and bound the cost of each re-assignment depending on the type of j. This case analysis is essentially the same as the one we had for UNIFORM-UFL-OUT and is deferred to the full version. Note that the points in $\mathcal{X}^o \cup \mathcal{X}^{o^*}$ are paired via κ.

3.1 Minor-Closed Families of Graphs We consider the problem k-MEDIAN-OUT in families of graphs that exclude a fixed minor H. Recall that a family of graphs is closed under minors if and only if all graphs in that family exclude some fixed minor.

Let $G = (V, E)$ be an edge-weighted graph excluding H as a minor where $\mathcal{X}, \mathcal{C} \subseteq V$ and let δ denote the shortest-path metric of G. We will argue Algorithm 3 for some appropriate constant $\rho' := \rho'(\epsilon, H)$ returns a set $\mathcal{S} \subseteq \mathcal{C}$ with $|\mathcal{S}| = (1+\epsilon) \cdot k$ where $\text{cost}(\mathcal{S}) \leq (1+\epsilon) \cdot OPT$. This can be readily adapted to k-MEANS-OUT using ideas from Section **??**. We focus on k-MEDIAN-OUT for simplicity. This will also complete the proof of Theorem 1.2 for minor-closed metrics. The proof of Theorem 1.1 for minor-closed metrics is proven similarly and is slightly simpler.

We use the same notation as our analysis for k-MEDIAN-OUT in doubling metrics. Namely, $\mathcal{S} \subseteq \mathcal{C}$ is a local optimum solution, $\mathcal{O} \subseteq \mathcal{C}$ is a global optimum solution, χ^a are the points assigned in the local optimum solution, χ^o are the outliers in the local optimum solution, etc. We assume $\mathcal{S} \cap \mathcal{O} = \emptyset$ (one can "duplicate" a vertex by adding a copy with a 0-cost edge to the original and preserve the property of excluding H as a minor) and $\chi^o \cap \chi^{o^*} = \emptyset$.

A key difference is that we do not start with a perfect partitioning of $\mathcal{S} \cup \mathcal{O}$, as we did with doubling metrics. Rather, we start with the r-divisions described in [14] which provides "regions" which consist of subsets of $\mathcal{S} \cup \mathcal{O}$ with limited overlap. We present a brief summary, without proof, of their partitioning scheme and how it is used to analyze the multiswap heuristic for UNCAPACITATED FACILITY LOCATION. Note that their setting is slightly different in that they show local search provides a true PTAS for k-MEDIAN and k-MEANS, whereas we are demonstrating a bicriteria PTAS for k-MEDIAN-OUT and k-MEANS-OUT. It is much easier to describe a PTAS using their framework if $(1+\epsilon) \cdot k$ centres are opened in the algorithm. Also, as the locality gap examples in the next section show, Algorithm 3 may not be a good approximation when using solutions $\mathcal{S}$ of size exactly k.

First, the nodes in V are partitioned according to their nearest centre in $\mathcal{S} \cup \mathcal{O}$, breaking ties in a consistent manner. Each part (i.e. Voronoi cell) is then a connected component so each can be contracted to get a graph G' with vertices $\mathcal{S} \cup \mathcal{O}$. Note G' also excludes H as a minor. Then for $r = d_H/\epsilon^2$ where d_H is a constant depending only on H, they consider an r-division of G'. Namely, they consider "regions" $R_1, \ldots, R_m \subseteq \mathcal{S} \cup \mathcal{O}$ with the following properties (Definition III.1.1 in [14]) First, define the boundary $\partial(R_a)$ for each region to be all centres $i \in R_a$ incident to an edge (i, i') of G' with $i' \notin R_a$.

- Each edge of G' has both endpoints in exactly one region.
- There are at most $c_H/r \cdot (|\mathcal{S}| + |\mathcal{O}|)$ regions where c_H is a constant depending only on H.
- Each region has at most r vertices.
- $\sum_{a=1}^{m} |\partial(R_a)| \leq \epsilon \cdot (|\mathcal{S}| + |\mathcal{O}|)$.

In general, the regions are not vertex-disjoint.

For each region R_a, the test swap $\mathcal{S} \to \mathcal{S} - ((R_a - \partial(R_a)) \cap \mathcal{S}) \cup (R_a \cap \mathcal{O})$ is considered. Each j with $\sigma(j) \in R_a - \partial(R_a)$ is moved in one of two ways:

- If $\sigma^*(j) \in R_a$, move j to $\sigma^*(j)$ for a cost change of $c_j^* - c_j$.
- Otherwise, if point j is in the Voronoi cell for some $i \in R_a$ then $\delta(j, \partial(R_a)) \leq c_j^*$ because the shortest path from j to $\sigma^*(j)$ must include a vertex v in the Voronoi cell of some $i \in \partial(R_a)$. By definition of the Voronoi partitioning, i is closer to v than $\sigma^*(j)$. So the cost change for j is at most $c_j^* - c_j$ again.
- Finally, if point j does not lie in the Voronoi cell for any $i \in R_a \cup \partial(R_a)$ then $\delta(j, \partial(R_a)) \leq c_j$ because the shortest path from j to $\sigma(j)$ again crosses the boundary of R_a. So the cost change for j is at most 0.

Lastly, for each $j \in \mathcal{X}$ if no bound of the form $c_j^* - c_j$ is generated for j according to the above rules, then j should move from $\sigma(j)$ to $\sigma^*(j)$ in some swap that opens $\sigma^*(j)$.

We use this approach as our starting point for k-MEDIAN-OUT. Let $\epsilon' > 0$ be a constant such that we run Algorithm 3 using $(1+\epsilon') \cdot k$ centres in $\mathcal{S}$. We will fix the constant ρ' dictating the size of the neighbourhood to be searched momentarily.

Form the same graph G' obtained by contracting the Voronoi diagram of G', let $R_1, \ldots, R_m$ be the subsets with the same properties as listed above for $\epsilon := \epsilon'/10$. The following can be proven in a similar manner to Theorem 3.1. Its proof appears in the full version.

LEMMA 3.1. *We can find regions $R'_1, \ldots, R'_\ell$ such that the following hold.*

- *Each edge lies in exactly one region.*
- $|R'_a| \leq O(r)$ *for each* $1 \leq a \leq \ell$.
- $|(R'_a - \partial(R'_a)) \cap \mathcal{S}| > |R'_a \cap \mathcal{O}|$.

The rest of the proof proceeds as with our analysis of k-MEDIAN-OUT in doubling metrics. For each $j \in \mathcal{X}^o$ let $\tau(j)$ be any index such that $\sigma^*(j) \in R'_a$ and for each $j \in \mathcal{X}^{o^*}$ let $\tau^*(j)$ be any index such that $\sigma(j) \in R_a$. For each R'_b, define the imbalance of outliers Δ_b to be $|\{j \in \mathcal{X}^o : \tau(j) = b\}| - |\{j \in \mathcal{X}^{o^*} : \tau^*(j) = b\}|$. Note $\sum_{b=1}^{\ell} \Delta_b = 0$.

Everything now proceeds as before so we only briefly summarize: we find groups G_s each consisting of $\Theta(r^2/\epsilon')$ regions of the form R'_b along with similar a similar bijection $\kappa : \mathcal{X}^o \to \mathcal{X}^{o^*}$ such that for each $j \in \mathcal{X}^o$ we have $R'_{\tau(j)}$ and $R'_{\tau^*(\kappa(j))}$ appearing in a common group. Finally, at most two regions are split by this grouping. At this point, we determine that swaps of size $\rho' = O(r^3/\epsilon') = O(d_H^3/{\epsilon'}^7)$ in Algorithm 3 suffice for the analysis.

For each group G_s, consider the test swap that opens all global optimum centres lying in some region R'_a used to form G_s and closes all local optimum centres in G_s that are neither in the boundary $\partial(R'_a)$ of any region forming G_s or in a split region. Outliers are reassigned exactly as with k-MEDIAN-OUT, and points in $\mathcal{X}^a \cap \mathcal{X}^{a^*}$ are moved as they would be in the analysis for UNCAPACITATED FACILITY LOCATION described above.

For each $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$, the analysis above shows the total cost change for j can be bounded by $c_j^* - c_j$. Similarly, for each $j \in \mathcal{X}^o$ we bound the total cost change of both j and $\kappa(j)$ together by $c_j^* - c_{\kappa(j)}$. So in fact $\text{cost}(\mathcal{S}) \leq \text{cost}(\mathcal{O})$.

4 General Metrics

Here we prove Theorem 1.3. That is, we show how to apply our framework for k-MEDIAN-OUT and k-MEANS-OUT to the local search analysis in [22] for k-MEDIAN and k-MEANS in general metrics where no assumptions are made about the distance function δ apart from the metric properties. The algorithm and the redirections of the clients are the same with both k-MEDIAN-OUT and k-MEANS-OUT. We describe the analysis and summarize the different bounds at the end to handle outliers.

Let $\epsilon > 0$ be a constant and suppose Algorithm 3 is run using solutions $\mathcal{S}$ with $|\mathcal{S}| \leq (1+\epsilon) \cdot k$ and neighbourhood size ρ' for some large constant $\rho' > 0$ to be determined. We use the same notation as before and, as always, assume $\mathcal{S} \cap \mathcal{O} = \emptyset$ and $\mathcal{X}^o \cap \mathcal{X}^{o^*} = \emptyset$.

Let $\phi : \mathcal{O} \to \mathcal{S}$ map each $i^* \in \mathcal{O}$ to its nearest centre in $\mathcal{S}$. Using a trivial adaptation of Algorithm 1 in [22] (the only difference being we have $|\mathcal{S}| = (1+\epsilon) \cdot k$ rather than $|\mathcal{S}| = k$), we find *blocks* $B_1, \ldots, B_m \subseteq \mathcal{S} \cup \mathcal{O}$ with the following properties.

- The blocks are disjoint and each $i^* \in \mathcal{O}$ lies in some block.
- For each block B_a, $|B_a \cap \mathcal{S}| = |B_a \cap \mathcal{O}|$.
- For each block B_a, there is exactly one $i \in B_a \cap \mathcal{S}$ with $\phi^{-1}(i) \neq \emptyset$. For this i we have $B_a \cap \mathcal{O} = \phi^{-1}(i)$.

Call a block **small** if $|B_a \cap \mathcal{S}| \leq 2/\epsilon$, otherwise it is **large**. Note there are at most $\frac{\epsilon}{2} \cdot k$ large blocks. On the other hand, there are $\epsilon \cdot k$ centres in $\mathcal{S}$ that do not appear in any block. Assign one such unused centre to each large block, note these centres i satisfy $\phi^{-1}(i) = \emptyset$ and there are still at least $\frac{\epsilon}{2} \cdot k$ centres in $\mathcal{S}$ not appearing in any block.

Now we create parts P_s. For each large block B_a, consider any paring between $B_a \cap \mathcal{O}$ and $\{i \in B_a \cap \mathcal{S} : \phi^{-1}(i) = \emptyset\}$. Each pair forms a part on its own. Finally, each small block is a part on its own. This is illustrated in Figure 2.

Note each part P_s satisfies $|P_s \cap \mathcal{S}| = |P_s \cap \mathcal{O}| \leq 2/\epsilon$. Then perform the following procedure: while there are at least two parts with size at most $2/\epsilon$, merge them into a larger part. If there is one remaining part with size at most $2/\epsilon$, then merge it with any part created so far. Now all parts have size between $2/\epsilon$ and $6/\epsilon$, call these groups $P'_1, \ldots, P'_\ell$. Note $\ell \leq \frac{\epsilon}{2} \cdot k$ because the sets $P'_a \cap \mathcal{O}$ partition $\mathcal{O}$. To each part, add one more $i \in \mathcal{S}$ which does not yet appear in a part. Summarizing, the parts $P'_1, \ldots, P'_\ell$ have the following properties.

- Each $i^* \in \mathcal{O}$ appears in precisely one part. Each $i \in \mathcal{S}$ appears in at most one part.

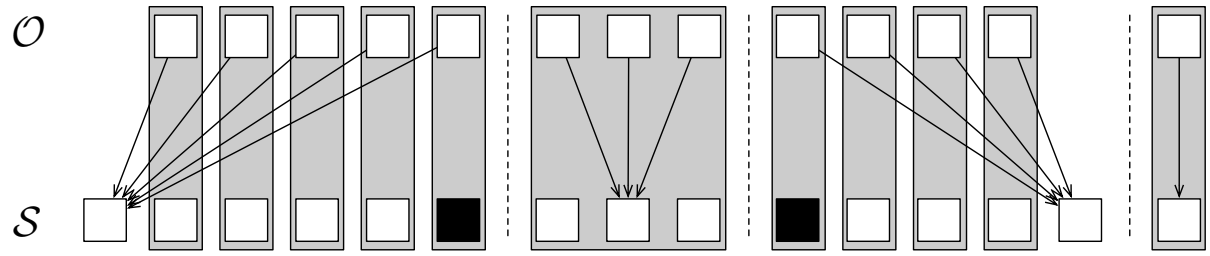

Figure 2: Depiction of the formation of the parts P_s where the arrows between them indicate the mapping ϕ. The white nodes are in the initial blocks and the blocks are separated by white dashed lines. The first and third block are large, so another centre (black) not yet in any block is added to them. The grey rectangles are the parts that are formed from the blocks.

- $|P'_a \cap \mathcal{O}| < |P'_a \cap \mathcal{S}| \leq \frac{6}{\epsilon} + 1 \leq \frac{7}{\epsilon}$.

Construct a pairing $\kappa : \mathcal{X}^o \to \mathcal{X}^{o^*}$ almost as before. First, pair up outliers within a group arbitrarily. Then arbitrarily pair up unpaired $j \in \mathcal{X}^o$ with points $j' \in \mathcal{X}^{o^*}$ such that $\sigma(j)$ does not appear in any part P'_a. Finally, pair up the remaining unpaired outliers using Algorithm 2 applied to these parts. Form groups G_s with these parts in the same way as before. Again, note some $i \in \mathcal{S}$ do not appear in any group, this is not important. What is important is that each $i^* \in \mathcal{O}$ appears in some group.

The swaps used in the analysis are of the following form. For each group G_s we swap in all global optimum centres appearing in G_s and swap out all local optimum centres appearing in G_s that are not in a part split by G_s. As each group is the union of $\Theta(1/\epsilon)$ parts, the number of centres swapped is $O(\epsilon^{-2})$. This determines $\rho' := O(\epsilon^{-2})$ in Algorithm 3 for the case of general metrics.

The clients are reassigned as follows. Note by construction that if $j \in \mathcal{X}^{o^*}$ has $\sigma(j)$ in a part that is not split then $\sigma^*(\kappa^{-1}(j))$ lies in the same group as $\sigma(j)$. Say that this is the group when $\kappa^{-1}(j)$ should be connected. Otherwise, pick any group containing $\sigma^*(\kappa^{-1}(j))$ and say this is when $\kappa^{-1}(j)$ should be connected.

For each group G_s,

- For each $j \in \mathcal{X}^o$, if j should be connected in this group then connect j to $\sigma^*(j)$ and make $\kappa(j)$ an outlier for a cost change of $c^*_j - c_{\kappa(j)}$.
- For any $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ where $\sigma(j)$ is closed, move j as follows:
  - If $\sigma^*(j)$ is now open, move j to $\sigma^*(j)$ for a cost change bound of $c^*_j - c_j$.
  - Otherwise, move j to $\phi(\sigma^*(j))$ which is guaranteed to be open by how we constructed the parts. The cost change here is summarized below for the different cases of k-MEDIAN and k-MEANS.

Finally, for any $j \in \mathcal{X}^a \cap \mathcal{X}^{a^*}$ that has not had its $c^*_j - c_j$ bound generated yet, move j to $\sigma^*(j)$ in any one swap where $\sigma^*(j)$ is opened to get a bound of $c^*_j - c_j$ on its movement cost.

The analysis of the cost changes follows directly from the analysis in [22], so we simply summarize.

- For k-MEDIAN-OUT, $\text{cost}(\mathcal{S}) \leq (3 + O(\epsilon)) \cdot OPT$.
- For k-MEANS-OUT, $\text{cost}(\mathcal{S}) \leq (25 + O(\epsilon)) \cdot OPT$.
- For k-clustering with ℓ_q^q-norms of the distances, $\text{cost}(\mathcal{S}) \leq ((3 + O(\epsilon)) \cdot q)^q$.

These are slightly different than the values reported in [22] because they are considering the ℓ_q norm whereas we are considering ℓ_q^q. This concludes the proof of Theorem 1.3.

5 The Locality Gap

In this section, we show that the natural local search heuristic has unbounded gap for UFL-OUT (with non-uniform opening costs). We also show that local search multiswaps that do not violate the number of clusters and outliers can have arbitrarily large locality gap for k-MEDIAN-OUT and k-MEANS, even in the Euclidean metrics. We further strengthen our example by showing a similar large locality gap even if a small violation of the number of outliers is permitted. Locality gap here refers to the ratio of any local optimum solution produced by the local search heuristics to the global optimum solution.

5.1 UFL-out with Non-uniform Opening Costs

First, we consider a multi-swap local search heuristic for the UFL-OUT problem with non-uniform centre opening costs. We show this for any local search that does

ρ-swaps and does not discard more than z points as outliers, where ρ is a constant and z is part of the input has unbounded ratio. Assume the set of $\mathcal{X} \cup \mathcal{C}$ is partitioned into disjoint sets A, B_1, B_2, ..., B_z, where:

- The points in different sets are at a large distance from one another.
- A has one centre i with the cost of ρ and z points which are colocated at i.
- For each of $\ell = 1, 2, \ldots, z$, the set B_ℓ has one centre i_ℓ with the cost of 1 and one point located at i_ℓ.

The set of centres $\mathcal{S} = \{i_1, i_2, \ldots, i_z\}$ is a local optimum for the ρ-swap local search heuristic; any ρ-swap between B_ℓ's will not reduce the cost, and any ρ-swaps that opens i will incur an opening cost of ρ which is already as expensive as the potential savings from closing ρ of the B_ℓ's. Note that the z points of A are discarded as outliers in $\mathcal{S}$. It is straightforward to verify that the global optimum in this case is $\mathcal{O} = \{i\}$, which discards the points in $B_1, \ldots, B_z$'s as outliers. The locality gap for this instance is $\text{cost}(\mathcal{S})/\text{cost}(\mathcal{O}) = \frac{z}{\rho}$, which can be arbitrarily large for any fixed ρ.

Note this can also be viewed as a planar metric, showing local search has an unbounded locality gap for UFL-OUT in planar graphs.

5.2 k-median-out and k-means-out Chen presents in [12] a bad gap example for local search for k-MEDIAN-OUT in general metrics. The example shows an unbounded locality gap for the multiswap local search heuristic that does not violate the number of clusters and outliers. We adapt this example to Euclidean metrics and prove Claim 1 below. The same example shows standard multiswap local search for k-MEANS-OUT has an unbounded locality gap.

CLAIM 1. *The ρ-swap local search heuristic that generates no more than k clusters and discards at most z outliers has an unbounded locality gap in Euclidean metrics, if $\rho < k - 1$.*

Proof. Consider an input in which $n \gg z \gg k > 1$. The set of points $\mathcal{X}$ is partitioned into disjoint sets B, $C_1, C_2, \ldots, C_{k-1}$, $D_1, D_2, \ldots, D_{k-2}$, and E:

- The distance between every pair of points from different sets is large.
- B has $n - 2z$ colocated points.
- For each $i = 1, 2, \ldots, k-1$, C_i has one point in the centre and $u - 1$ points evenly distributed on the perimeter of a circle with radius β from the centre.
- For each $j = 1, 2, \ldots, k-2$, D_j has $u-1$ colocated points.
- E has one point at the centre and $u + k - 3$ points evenly distributed on the perimeter of the circle with radius γ,

where $u = z/(k-1)$, and β and γ are chosen such that $\gamma < (u-1)\beta < 2\gamma$ (see Figure 3). Let $f(.)$ denote the centre point of a set (in the case of colocated sets, any point from the set). Then, the set $\mathcal{S} = \{f(B), f(D_1), f(D_2), \ldots, f(D_{k-2}), f(E)\}$ is a local optimum for the ρ-swap local search if $\rho < k - 1$ with cost $(u+k-3)\gamma$. The reason is that since the distance between the sets is large, we would incur a large cost by closing $f(B)$ or $f(E)$. Therefore, we need to close some points in the sets $D_1, \ldots, D_{k-2}$, and open some points in $C_1, \ldots, C_{k-1}$ to ensure we do not violate the number of outliers. Since $z \gg k$, we can assume $u > k - 1$. We can show via some straightforward algebra that if we close ρ points from D_j's, then we need to open points from exactly ρ different C_i's to keep the number of outliers below z. Since the points on the perimeter are distributed evenly, we incur the minimum cost by opening $f(C_i)$'s. So, we only need to show that swapping at most ρ points from $f(D_j)$'s with ρ points in $f(C_i)$'s does not reduce the cost. Assume w.l.o.g. that we swap $f(D_1), f(D_2), \ldots, f(D_\rho)$ with $f(C_1), f(C_2), \ldots, f(C_\rho)$. The new cost is $\rho(u-1)\beta + (u + k - \rho - 3)\gamma$ since as a result of the swap, we can reclaim ρ points from the set E as outliers. Notice that $\text{cost}(\mathcal{S}) = (u+k-3)\gamma$. Therefore, the cost has, in fact, increased as a result of the ρ-swap since $(u-1)\beta > \gamma$. Now, we can show the claim for k-MEDIAN-OUT.

Consider the solution $\mathcal{O} = \{f(B), f(C_1), f(C_2), \ldots, f(C_{k-1})\}$ which costs $(k-1)(u-1)\beta$. This is indeed the global optimum for the given instance. The locality gap then would be

$$\frac{(u+k-3)\gamma}{(k-1)(u-1)\beta} > \frac{u+k-2}{2(k-1)},$$

since $(u-1)\beta < 2\gamma$. This ratio can be arbitrarily large as z (and consequently u) grows. A slight modification of this example to planar metrics also shows local search has an unbounded locality gap for k-MEDIAN-OUT and k-MEANS-OUT in planar graphs. In particular, the sets of collocated points can be viewed as stars with 0-cost edges and the sets E and all C_i can be viewed as stars where the leaves are in $\mathcal{X}$ and have distance γ or β (respectively) to the middle of the star, which lies in $\mathcal{C}$.

Small Violation on the Number of Outliers: Now, we consider a setting where we are allowed to violate the number of outliers by $z^{1-\delta}$, for a given small constant δ.

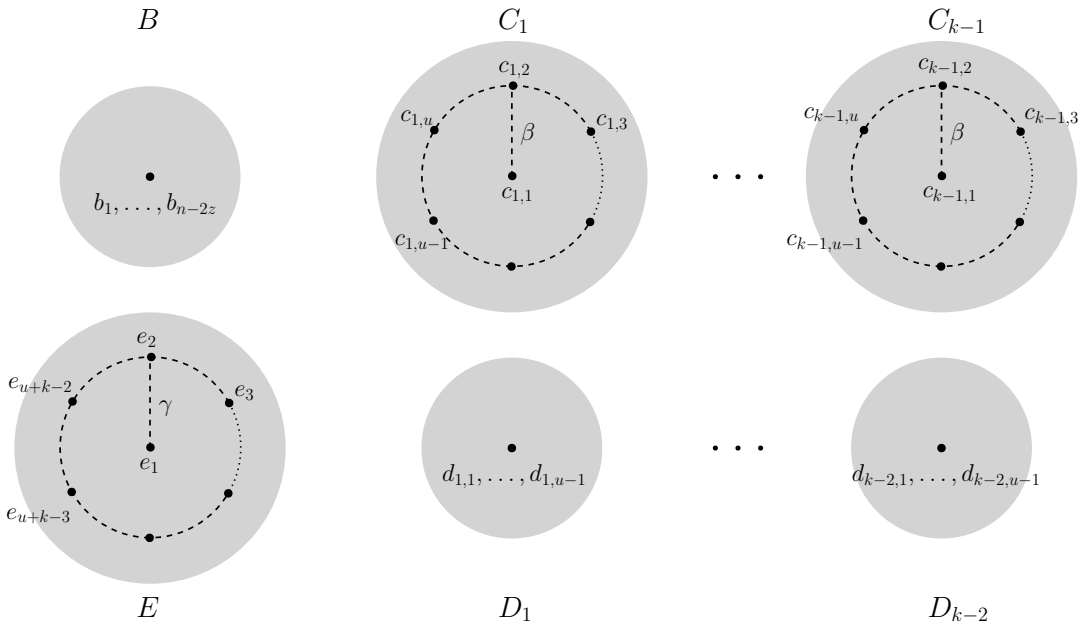

Figure 3: The locality gap counter example for k-MEDIAN in Euclidean metrics.

CLAIM 2. *The ρ-swap local search heuristic that generates no more than k clusters and discards at most $z + z^{1-\delta}$ outliers for any δ , $0 < \delta \leq 1$, has an unbounded locality gap in Euclidean metrics, if $\rho < k - 1$.*

Proof. We again consider the solution $\mathcal{S} = \{f(B), f(E), F(D_1), f(D_2), \ldots, f(D_{k-2})\}$, this time with $z^{1-\delta}$ outliers, namely the sets $C_1, C_2, \ldots, C_{k-1}$, and any $z^{-\delta}$ distinct points from the perimeter of the set E. First, note that $\mathcal{S}$ is locally optimal. For the same reasons as before, we do not swap out $f(B)$ and $f(E)$. Therefore, any other solutoin obtained by a ρ-swap will be of the form $\mathcal{S}' = \{f(B), f(E), f(C_1), f(C_2), \ldots, f(C_\rho), f(D_{\rho+1}), \ldots, f(D_{k-2})\}$, i.e., it closes some points from the D_j's and opens the same number from C_i's. The set of outliers consists of the sets $D_1, D_2, \ldots, D_\rho, C_{\rho+1}, \ldots, C_{k-1}$, and any $z^{-\delta} + \rho$ distinct perimeter points from E. Comparing the cost of $\mathcal{S}$ and $\mathcal{S}'$, we have:

$$\begin{aligned} &cost(\mathcal{S}) - cost(\mathcal{S}') \\ &= (u + k - 3 - z^{1-\delta})\gamma \\ &\quad - \big(\rho(u-1)\beta + (u + k - 3 - z^{1-\delta} - \rho)\gamma\big) \\ &= (\gamma - (u-1)\beta)\,\rho \quad < 0, \end{aligned}$$

where the inequality is due to our choice of β and γ. Finally, comparing the cost of $\mathcal{S}$ against the global optimal $\mathcal{O}$, we get

$$\begin{aligned} \frac{cost(\mathcal{S})}{cost(\mathcal{O})} &= \frac{(u + k - 3 - z^{1-\delta})\gamma}{(k-1)(u-1)\beta} \\ &> \frac{u + k - 3 - z^{1-\delta}}{2(k-1)} \\ &= \frac{\dfrac{z}{k-1} - z^{1-\delta}}{2(k-1)} + \frac{k-3}{2(k-1)} \\ &> \frac{(z^\delta - k + 1)}{2(k-1)^2} \cdot z^{1-\delta} \end{aligned}$$

where the first inequality is due to the choice of the parameters β and γ. Since $z \gg k$, this ratio grows with z for arbitrarily small values of δ. We conclude that the cost of a local optimal that violates the number of outliers by $z^{1-\delta}$ can be unbounded compared to the cost of a global optimum.

References

[1] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for K-means and euclidean K-median by primal-dual algorithms. *arXiv preprint arXiv:1612.07925*, 2016.

[2] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Mach. Learn.*, 75(2):245–248, 2009.

[3] Sanjeev Arora. Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems. *J. ACM*, 45(5):753–782, 1998.

[4] Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for Euclidean K-medians and related problems. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC '98)*, pages 106–113. ACM, 1998.

[5] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pages 1027–1035. SIAM, 2007.

[6] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristic for K-median and facility location problems. In *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing (STOC '01)*, pages 21–29. ACM, 2001.

[7] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for K-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

[8] Sayan Bandyapadhyay and Kasturi Varadarajan. On variants of K-means clustering. In *Proceedings of the 32nd International Symposium on Computational Geometry (SoCG '16)*, Leibniz International Proceedings in Informatics (LIPIcs), 2016.

[9] Pavel Berkhin et al. A survey of clustering data mining techniques. *Grouping multidimensional data*, 25:71, 2006.

[10] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for K-median, and positive correlation in budgeted optimization. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 737–756. SIAM, 2014.

[11] Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 642–651. Society for Industrial and Applied Mathematics, 2001.

[12] Ke Chen. *ALGORITHMS ON CLUSTERING, ORIENTEERING, AND CONFLICT-FREE COLORING*. PhD thesis, University of Illinois at Urbana-Champaign, 2007.

[13] Ke Chen. A constant factor approximation algorithm for K-median clustering with outliers. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 826–835, 2008.

[14] Vincent Cohen-Addad, Philip N. Klein, and Claire Mathieu. Local search yields approximation schemes for k-means and k-median in euclidean and minor-free metrics. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 353–364, 2016.

[15] W. Fernandez de la Vega, Marek Karpinski, Claire Kenyon, and Yuval Rabani. Approximation schemes for clustering problems. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing (STOC '03)*, pages 50–58. ACM, 2003.

[16] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33, 2004.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[18] Dan Feldman, Morteza Monemizadeh, and Christian Sohler. A PTAS for K-means clustering based on weak Coresets. In *Proceedings of the Twenty-third Annual Symposium on Computational Geometry (SoCG '07)*, SoCG '07, pages 11–18. ACM, 2007.

[19] Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a ptas for k-means in doubling metrics. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 365–374, 2016.

[20] Zachary Friggstad, Mohsen Rezapour, and Mohammad R. Salavatipour. Local search yields a PTAS for k-means in doubling metrics. *CoRR*, abs/1603.08976, 2016.

[21] Sudipto Guha and Samir Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of algorithms*, 31(1):228–248, 1999.

[22] A. Gupta and T. Tangwongsan. Simpler analyses of local search algorithms for facility location. *CoRR, abs/0809.2554*, 2008.

[23] Shalmoli Gupta, Ravi Kumar, Kefu Lu, Benjamin Moseley, and Sergei Vassilvitskii. Local search methods for K-means with outliers. *Proceedings of the VLDB Endowment*, 10(7):757–768, 2017.

[24] Sariel Har-Peled and Akash Kushal. Smaller Coresets for K-median and K-means clustering. In *Proceedings of the Twenty-first Annual Symposium on Computational Geometry (SoCG '05)*, pages 126–134. ACM, 2005.

[25] Sariel Har-Peled and Soham Mazumdar. On Coresets for K-means and K-median clustering. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing (STOC '04)*, pages 291–300. ACM, 2004.

[26] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[27] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted Voronoi diagrams and randomization to variance-based K-clustering. In *Proceedings of the tenth annual Symposium on Computational Geometry (SoCG '94)*, pages 332–339. ACM, 1994.

[28] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.*, 31(8):651–666, 2010.

[29] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems.

In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 731–740. ACM, 2002.

[30] Kamal Jain and Vijay V Vazirani. Approximation algorithms for metric facility location and K-median problems using the primal-dual schema and Lagrangian relaxation. *Journal of the ACM (JACM)*, 48(2):274–296, 2001.

[31] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. A local search approximation algorithm for K-means clustering. *Comput. Geom. Theory Appl.*, 28(2-3):89–112, 2004.

[32] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\epsilon)$ -approximation algorithm for K-means clustering in any dimensions. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS '04)*, pages 454–462. IEEE Computer Society, 2004.

[33] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. Linear-time approximation schemes for clustering problems in any dimensions. *J. ACM*, 57(2):5:1–5:32, 2010.

[34] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for K-means. *Information Processing Letters*, 120:40–43, 2017.

[35] Shi Li. A 1.488-approximation for the uncapacitated facility location problem. In *Proceedings of the 38th Annual International Colloquium on Automata, Languages and Programming (ICALP '11)*, pages 45–58, 2011.

[36] Shi Li and Ola Svensson. Approximating K-median via pseudo-approximation. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing (STOC '13)*, pages 901–910. ACM, 2013.

[37] S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 2006.

[38] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar K-means problem is NP-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation (WALCOM '09)*, pages 274–285. Springer-Verlag, 2009.

[39] Jirı Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[40] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the K-means problem. *J. ACM*, 59(6):28:1–28:22, 2013.

[41] Thrasyvoulos N Pappas. An adaptive clustering algorithm for image segmentation. *IEEE Transactions on signal processing*, 40(4):901–914, 1992.

[42] Andrea Vattani. The hardness of K-means clustering in the plane. *Manuscript*, 2009.