**Clustering What Matters: Optimal Approximation for Clustering with Outliers** \*

**Akanksha Agrawal**1 **, Tanmay Inamdar**2 **, Saket Saurabh**2, 3 **, Jie Xue**4

1 Indian2Icy, s,,ai,India.

3 The Institsig, iennai, India

akanksha@cse.iitm.ac.in, Tanmay.Inamdar@uib.no, saket@imsc.res.in, jiexue@nyu.edu

**Abstract**

Clustering with outliers is one of the most fundamental prob- lems in Computer Science. Given a set X of n points and two integers k and m, the clustering with outliers aims to ex- clude m points from X , and partition the remaining points into k clusters that minimizes a certain cost function. In this paper, we give a general approach for solving clustering with outliers,which results in a fxed-parameter tractable (FPT) al- gorithmin k and m (i.e., an algorithm with running time of the form f(k, m) · nO(1) for some function f),that almost matches the approximation ratio for its outlier-free counter- part.

As a corollary, we obtain FPT approximation algorithms with optimal approximation ratios for k-MEDIAN and k-MEANS with outliers in general metrics. We also exhibit more appli- cations of our approach to other variants of the problem that impose additional constraints on the clustering, such as fair- ness or matroid constraints.

**Introduction**

Clustering is a family of problems that aims to group a given set of objects in a meaningful way—the exact “meaning” may vary based on the application. These are fundamen- tal problems in Computer Science with applications rang- ing across multiple felds like pattern recognition, machine learning, computational biology, bioinformatics and social science. Thus, these problems have been a subject of exten- sive studies in the feld of Algorithm Design (and its sub- felds), see for instance, the surveys on this topic (and ref- erences therein) (Xu and Tian 2015; Rokach 2009; Blmer et al. 2016).

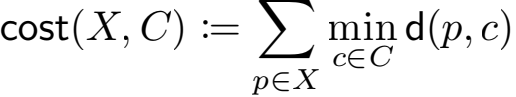
Two of the central clustering problems are k-MEDIAN and k-MEANS. In the standard k-MEDIAN problem, we are given a set X of n points, and an integer k, and the goal is

\*The authors are listed in the alphabetical order of the last names.

A full version of the paper containing missing proofs and other de- tails can be found is available on arXiv:2212.00696 (Agrawal et al. 2022).

Copyright © 2023, Association for the Advancement of Artifcial Intelligence (www.aaai.org). All rights reserved.

to fnd a set C\* ⊆ X of at most k *centers*, such that the fol- lowing cost function is minimized over all subsets C of size at most k.



In k-MEANS, the objective function instead contains the sum of squares of distances.

Often real world data are contaminated with a small amount of noise and these noises can substantially change the clusters that we obtain using the underlying algorithm. To circumvent the issue created by such noises, there are several studies of clustering problems with outliers, see for instance, (Chen 2008; Krishnaswamy, Li, and Sandeep 2018; Goyal, Jaiswal, and Kumar 2020; Feng et al. 2019; Friggstad et al. 2019; Almanza et al. 2022).

In outlier extension of the k-MEDIAN problem, which we call k-MEDIANOUT, we are also given an additional inte- ger m ≥ 0 that denotes the number of *outliers* that we are allowed to drop. We want to fnd a set C of at most k cen- ters, and a set Y ⊆ X of at most m outliers, such that cost(X \ Y, C) := Σp∈X\Y minc∈C d(p,c) is minimized over all (Y, C) satisfying the requirements. Observe that the cost of clustering fork-MEDIANOUT equals the sum of dis- tances of each point to its nearest center, after excluding a set of mpoints from consideration 1 . We remark that in a simi- lar spirit we can defne the outlier version of the k-MEANS problem,which we call k-MEANSOUT.

In this paper, we will focus on approximation algorithms. An algorithm is said to have an approximation ratio of α, if it is guaranteed to return a solution of cost no greater than α times the optimal cost, while satisfying all other conditions. That is, the solution must contain at most k centers, and drop m outliers. If the algorithm is randomized, then it must re- turn such a solution with *high probability*, i.e., probability at least 1 − n −c for some c ≥ 1.

For a fxed set C of centers, the set of m outliers is auto- matically defned, namely the set of m points that are far- thest from C (breaking ties arbitrarily). Thus, an optimal

1In the technical section, we consider a more general formu- lation of k-MEDIAN, where the set of candidate centers may be different from the set X of points to be clustered.

clustering fork-MEDIANOUT, just like k-MEDIAN, can be found in nO(k) time by enumerating all center sets. On the other hand, we can enumerate all nO(m) subsets of outliers, and reduce the problem directly to k-MEDIAN. Other than these straightforward observations, there are several non- trivial approximations known for k-MEDIANOUT, which we discuss in a subsequent paragraph.

**Our Results.** In this work, we describe a general frame- work that reduces a *clustering with outliers* problem (such as k-MEDIANOUT or k-MEANSOUT) to its outlier-free coun- terpart in an approximation-preserving fashion. More specif- ically, given an instance I of k-MEDIANOUT, our reduction runs in time f(k,m,ϵ) · nO(1), and produces multiple in- stances of k-MEDIAN, such that a β-approximation for *at least* one of the produced instances of k-MEDIAN implies a (β + ϵ)-approximation for the original instance I of k- MEDIANOUT. This is the main result of our paper.

Our framework does not rely on the specifc properties of the underlying metric space. Thus, for special metrics, such as Euclidean spaces, or shortest-path metrics induced by sparse graph classes, for which FPT (1 + ϵ)-approximations are known for k-MEDIAN, our framework implies match- ing approximation for k-MEDIANOUT. Finally, our frame- work is quite versatile in that one can extend it to obtain approximation-preserving FPT reductions for related *clus- tering with outliers* problems, such as k-MEANSOUT, and clustering problems with *fair outliers* (such as (Bandyapad- hyay et al. 2019; Jia, Sheth, and Svensson 2020)), and MA - TROID MEDIAN WITH OUTLIERS. We conclude by giving a partial list of the corollaries of our reduction framework. The running time of each algorithm is f(k,m,ϵ) · nO(1) for some function f that depends on the problem and the set- ting. Next to each result, we also cite the result that we use as a blackbox to solve the outlier-free clustering problem.

• (1 + 2/e + ϵ) ≈ (1.74 + ϵ)-approximation (resp. 1 + 8/e + ϵ)-approximation) for k-MEDIANOUT (resp. k- MEANSOUT) in general metrics (Cohen-Addad, Saulpic, and Schwiegelshohn 2021). These approximations are tight even for m = 0, under a reasonable complexity theoretic hypothesis, as shown in the same paper.

• (1 + ϵ)-approximation for k-MEDIANOUT and k- MEANSOUT in (i) metric spaces of constant doubling di- mensions, which includes Euclidean spaces of constant dimension, (ii) metrics induced by graphs of bounded treewidth, and (iii) metrics induced by graphs that ex- clude a fxed graph as a minor (such as planar graphs). (Cohen-Addad, Saulpic, and Schwiegelshohn 2021).

• (2 + ϵ)-approximation for MATROID MEDIAN WITH OUTLIERS in general metrics, where k refers to the *rank* of thematroid. (Cohen-Addad et al. 2019)

• (1+2/e+ϵ)-approximation for COLORFUL k-MEDIAN in general metrics, where m denotes the total number of outliers across all color classes (Cohen-Addad et al. 2019). The preceding two problems are orthogonal gen- eralizations of k-MEDIANOUT, and are formally defned in Section .

**Our Techniques.** Our reduction is inspired from the following seemingly simple observation that relates k- MEDIANOUT and k-MEDIAN. Let I be an instance of k- MEDIANOUT, where we want to fnd a set C of k centers, such that the sum of distances of all except at most m points to the nearest center in C is minimized. By treating the out- liersin an optimal solution for I as *virtual centers*, one ob- tains a solution for (k+m)-MEDIAN *without outliers* whose cost is at most the optimal cost of I. In other words, the op- timal cost of an appropriately defned instance I- of (k+m)- MEDIAN is a *lower bound* on the optimal cost of I. Since k- MEDIAN is a well-studied problem, at this point, one would hope that it is suffcient to restrict the attention to I-. That i-s, if we obtain a solution (i.e., a set of k + m centers) for I, can then be modifed to obtain a solution (i.e., a set of k centers and m outliers) for I. However, it is unclear whether one can do such a modifcation without blowing up the cost for I. Nevertheless, this connection between I- and I turns out to be useful, but we need several new ideas to exploit it. As in before, we start with a constant approximation for I-, and perform a sampling similar to (Chen 2009) to obtain a weighted set of points. This set is obtained by dividing the set of points connected to each center in the approx- imate solution into *concentric rings*, such that the “error” introduced in the cost by treating all points in the ring as identical is negligible. Then, we sample O((k+m)log n/ϵ) points from each ring, and give each point an appropriate weight. We then prove a crucial concentration bound (cf.

Lemma 1), which informally speaking relates the connec- tion cost of original set of points in a ring, and the corre- sponding weighted sample. In particular, for *any set of* k *centers*, with good probability, the difference between the original and the weighted costs is “small”, *even after exclud- ing at most* m*outliers from both sets*. Intuitively speaking, this concentration bound holds because the sample size is large enough compared to both k and m. Then, by taking the union of all such samples, we obtain a weighted set S of O(((k + m)log n/ϵ)2 ) points that preserves the connection cost to any set of k centers, even after excluding m outliers with at least a constant probability. Then, we enumerate all sets Y of size m from S, and solve the resulting k-MEDIAN instance induced on S \ Y. Finally, we argue that at least one of the resulting instances I′ will have the property that, a β-approximation for I′ implies a (β + ϵ)-approximation for I.

**Related Work.** The frst constant approximation for k- MEDIANOUT was given by (Chen 2008) for some large constant. More recently, (Krishnaswamy, Li, and Sandeep 2018; Gupta, Moseley, and Zhou 2021) gave constant ap- proximations based on iterative LP rounding technique, and the 6.994-approximation by the latter is currently the best known approximation. These approximation algorithms run in polynomial time in n. (Krishnaswamy, Li, and Sandeep 2018) also give the best known polynomial approximations for related problems of k-MEANSOUT and MATROID ME - DIAN.

Now we turn to FPT approximations, which is also the

setting for our results. To the best of our knowledge, there

are three works in this setting, (Feng et al. 2019; Goyal,

Jaiswal, and Kumar 2020; Statman, Rozenberg, and Feld-

man 2020). The idea of relating k-MEDIAN WITH m OUT-

LIERS to (k + m)-MEDIAN that we discuss above is also

present in these works. Even though it is not stated explic-

itly, the approach of Statman et al. (Statman, Rozenberg,

and Feldman 2020) can be used to obtain FPT approxi-

mations in general metrics; albeit with a worse approxima-

tion ratio. However, by using additional properties of Eu-

clidean k-MEDIANOUT/k-MEANSOUT (where one is al-

lowed to place centers anywhere in Rd ) their approach yields

a (1+ϵ)-approximation in FPT time. The best FPT approxi-

mations in general metrics, to the best of our knowledge, are

3 + ϵ for k-MEDIANOUT by Goyal et al. (Goyal, Jaiswal,

and Kumar 2020), and 6 + ϵ for k-MEANSOUT by Feng

et al. (Feng et al. 2019). Thus, our FPT approximation algo-

rithms with ratio 1+ , and 1+  + ϵ

for k-MEANSOUT improve on these results. Furthermore,

our result is essentially an *approximation-preserving reduc-*

*tion* from k-MEDIANOUT to k-MEDIAN in the same kind

of metric, which automatically yields improved approxima-

tions in some special metrics as discussed earlier.

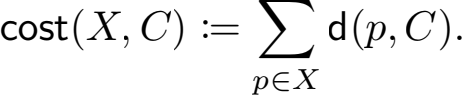
On the lower bound side, (Guha and Khuller 1999) showed it is NP-hard to approximate k-MEDIAN (and thus k-MEDIANOUT) within a factor 1 + 2/e − ϵ for any ϵ > 0. Recently, (Cohen-Addad et al. 2019) strengthened this re- sult under a reasonable complexity-theory hypothesis, and showed that an (1+2/e−ϵ)-approximation algorithm must take at least nkg(ϵ) time for some function g().

*Bicriteria approximations* relax the strict requirement of using at most k centers, or dropping at most m outliers, in order to give improved approximation ratios, or effciency (or both). For k-MEDIANOUT, (Charikar et al. 2001) gave a 4(1+1/ϵ)-approximation, while dropping m(1+ϵ) outliers. (Gupta et al. 2017) gave a constant approximation based on local search for k-MEANSOUT that drops O(kmlog(n∆)) outliers, where ∆ is the diameter of the set of points. (Frig- gstad, Rezapour, and Salavatipour 2019) gave a (25 + ϵ)- approximation that uses k(1 + ϵ) centers but only drops m outliers. In Euclidean spaces, they also give a (1 + ϵ)- approximation that returns a solution with k(1 + ϵ) centers.

**Preliminaries**

**Basic notions.** Let (Γ , d) be a metric space, where Γ is a fnite set of points, and d : Γ × Γ → R is a dis- tance function satisfying symmetry and triangle inequal- ity. For any fnite set S ⊆ Γ and a point p ∈ Γ, we let d(p, S) := mins∈S d(p, S), and let diam(S) := maxx,y∈S d(x,y). For two non-empty sets S, C ⊆ Γ, let d(S, C) = minp∈S d(p, S) = minp∈S minc∈C d(p,c). For a point p ∈ Γ, r ≥ 0, and a set C ⊆ Γ, let BC (p,r) = {q ∈ C : d(p,c) ≤ r}. Let T be a fnite (multi)set of n real numbers,for some positive integer n, and let 1 ≤ m ≤ n. Then, we use the notation sum~m(T) to denote the sum of n − m smallest valuesin T (including repetitions in case of a multi-set).

**The** k**-median problem.** In the k-MEDIAN problem, an instance is a triple I = (X,F, k), where X and F are fnite sets of points in some metric space (Γ , d), and k ≥ 1 is an integer. The points in X are called *clients*, and the points in F are called *facilities* or *centers*. The task is to fnd a subset C ⊆ F of size at most k that minimizes the cost function



The *size* of an instance I = (X,F, k) is defned as |I| = |X ∪ F|, which we denote by n.

k**-median with outliers.** The input to k-MEDIANOUT contains an additional integer 0 ≤ m ≤ n, and thus an instance is given by a 4-tuple I = (X,F,k, m). Let C ⊆ F be a set of facilities. We defne costm (X, C) := sum~m{cost(p, C) : p ∈ X}, i.e., the sum ofn−m smallest distances of points in X to the set of centers C. The goal is to fnd a set of centers C minimizing costm (X, C) over all sets C ⊆ F of size at most k. Given a set C ⊆ F of centers, we denote the corresponding solution by (Y, C), where Y ⊆ X is a set of m outlier points in X with largest distances real- izing costm (X, C). Given an instance I of k-MEDIANOUT, we use OPT(I) to denote the value of an optimal solution to I.

**Weighted sets and random samples.** During the course of the algorithm, we will often deal with *weighted sets* of points. Here, S ⊆ X is a weighted set, with each point p ∈ S having integer weight w(p) ≥ 0. For any set C ⊆ F and 1 ≤ m ≤ |S|, defnewcostm (S, C) := sum~m{d(p, C) · w(p) : p ∈ S}. A *random sample* of a fnite set S refers to a random subset of S. Throughout this paper, random sam- ples are always generated by picking points *uniformly* and *independently*.

k**-Median with Outliers**

In this section, we give our FPT reduction from k- MEDIANOUT to the standard k-MEDIAN problem. For- mally, we shall prove the following theorem.

**Theorem 1.** *Suppose there exists a* β*-approximation al- gorithm for* k-MEDIAN *with running time* T(n,k)*, and a* τ*-approximation algorithm for* k + m*-*MEDIAN *with polynomial running time, where* β *and* τ *are constants. Then there exists a* (β + ϵ)*-approximation algorithm for*

k-MEDIANOUT *with running time* O(m) · T(n,k)

·

nO(1)*, where* n *is the instance size and* m*is the number of outliers.*

Combining the above theorem with the known (1 + 2/e + ϵ)-approximation k-median algorithm (Cohen-Addad et al. 2019) that runs in (k/ϵ)O(k) · nO(1) time, we directly have the following result.

**Corollary 1.** There exists a (1+2/e+ϵ)-approximation al-

gorithm fork-MEDIANOUT O(m)

·



number of outliers.

The rest of this section is dedicated to proving Theorem 1. Let 工 = (X,F,k, m) be an instance of k-MEDIANOUT. We defne a (k+m)-MEDIAN instance 工′ = (X, FU X, k+m), where in addition to the original set of facilities, there is a facility co-located with each client. We have the following observation.

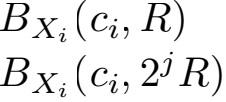
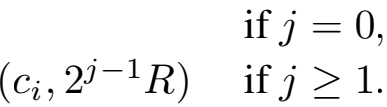
**Observation 1.** OPT(工′ ) ≤ OPT(工), i.e., the value of an optimal solution to 工′ is a lower bound on the value of an optimal solution to 工.

*Proof.* Let (Y∗ , C ∗ ) be an optimal solution to 工 realizing the value OPT(工). We defne a solution (Y′ , C′ ) for 工′ as follows: let Y′ = X , and C′ = C ∗ U Y ∗ . That is, the set of centers C′ is obtained by adding a facility co-located with each outlier point from Y ∗ . Now we argue about the costs. Since C∗ ≤ C′ , for each point p ∈ Y ∗ , d(p, C′ ) ≤ d(p, C∗ ). On the other hand, for each q ∈ X \ Y ∗ , d(q, C′ ) = 0, since there is a co-located center in C ∗ . This implies that cost0 (X, C′ ) ≤ costm (X, C). Since the solution (Y′ , C′ ) is feasible for the instance 工′ , it follows that OPT(工′ ) is no larger than the cost cost0 (X, C′ ). 

Now, we use τ-approximation algorithm guaranteed by the theorem, for the instance 工′ , and obtain a set of at most k′ ≤ k + m centers A such that cost0 (X, A) ≤ τ · OPT(工′ ) ≤ τ · OPT(工). By assumption, running this

algorithm takes polynomial time. Let R = 

lower bound on average radius, and ϕ = 「log(τn)l. For each center ci ∈ A, let Xi ≤ X denote the set of points whose closest center in A is ci. By arbitrarily breaking ties, we can assume that the sets Xi are disjoint,i.e., {Xi }1≤i≤k′ forms a partition of X. Now we further partition each Xi into smaller groups such that the points in each group have similar distances to ci. Specifcally, we defne

Xi,j := { \ BXi 

Let s = 

enough constant c. We defne a weighted set of points

Si,j ≤ Xi,j as follows. If |Xi,j | ≤ s, then we say Xi,j is

*small*. In this case, defne Si,j := Xi,j and let the weight

wi,j of each point p ∈ Si,j be 1. Otherwise, |Xi,j | > s

and we say that Xi,j is *large*. In this case, we take a ran-

dom sample Si,j ≤ Xi,j of size s. We set the weight of

every point in Si,j to be wi,j = |Xi,j |/|Si,j |. For conve-

nience, assume the weights wi,j to be integers 2 . Finally, let

S = ui,j Si,j . The set S can be thought of as an ϵ-coreset

for the k-MEDIANOUT instance 工. Even though we do not

defne this notion formally, the key properties of S will be

proven in Lemma 2 and 3. Thus, we will often informally

refer to S as a *coreset*.

**Proposition 1.** *We have* |S| = O(((k + m)log n/ϵ)2 ) *if* λ *is a constant.*

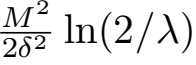
2For a large Xi,j, the quantity  may not be an integer. We

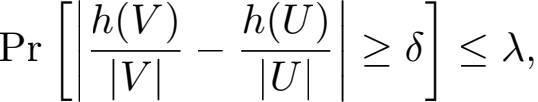
ignore this technicality for now, and discuss in the full version how to modify the construction slightly to ensure that the weights are integral.

*Proof.* For any p ∈ X , d(p, A) ≤ cost0 (X, A) = τn · R ≤ 2ϕ R. Therefore, for any ci ∈ A and j > ϕ, Xi,j′ = 0, and

Xi = u=0 Xi,j . It follows that the number of non-empty

sets Xi,j is at most |A|·(1+log(τn)) = O((k +m)log n), since |A| ≤ k + m and τ is a constant. For each non-empty Xi,j , |Si,j | ≤ 2s = O((m + klog n)/ϵ2 ), if λ is a constant. Since S = ui,j Si,j , the claimed bound follows. ii

**Proposition 2.** *(Chen 2009; Haussler 1992) Let* M ≥ 0 *and* η *be fxed constants, and let*h(·) *be afunction defned on a set* V *such that* η ≤ h(p) ≤ η + M *for all* p ∈ V*. Let* U ≤ V *be a random sample of size* s*, and* δ > 0 *be a parameter. If* s ≥ *, then*



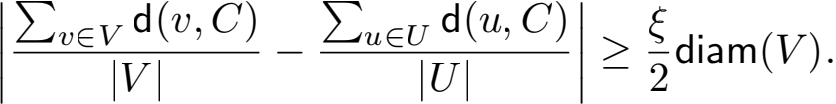


**Lemma 1.** Let (Γ , d) be a metric space, V ≤ Γ be a fnite set of points, λ′ ,ξ > 0, q ≥ 0 be parameters, and defne s′ =  sizes′ . Then for any fxedfnite set C ≤ F with probability at least 1 — λ′ it holds that for any 0 ≤ t ≤ q,

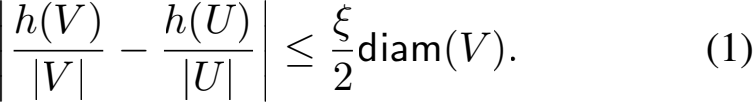
|costt (V, C) — wcostt′ (U, C)| ≤ ξ|V |(diam(V)+d(V, C)), where t′ = lt|U|/|V |」and w(u) = |V |/|U| for all u ∈ U.

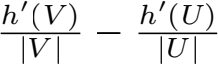
*Proof.* Throughout the proof, we fx the set C ≤ F and 0 ≤ t ≤ q as in the statement of the lemma. Next, we defne the following notation. For each v ∈ V, let h(v) := d(v, C), and leth(V) := Σv∈V h(v), and h(U) := Σu∈U h(u). Analogously, let h′ (V) := costt (V, C), and h′ (U) := costt′ (U, C). Let η(V) := minv∈V d(v, C), and η(U) := minu∈U d(u, C).

By applying Proposition 2 with η = η(V), M = diam(V) and δ = ξM/2, we know with probability at most λ′ , the following event happens.



As h(V) = Σv∈V d(v, C) and h(U) = Σu∈U d(u, C), with probability at least 1 — λ′ , we have that

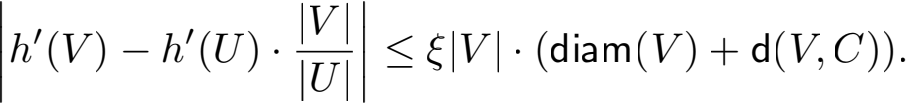


**Claim 1.** *Let* ∆ =  *. If Equation 1 holds, then*

*we have*

—ξ(diam(V) + d(V, C)) ≤ ∆ ≤ ξdiam(V).

The proof of the above claim involves a lot of calculations, so we defer it to the full version. Assuming its correctness, the inequality in Claim 1 implies



Since Equation 1 holds with probability at least 1 — λ′ , the above inequality also holds with probability at least 1 — λ′ .

The preceding inequality is equivalent to the one in the

emma, because h′ (V) = costt (V, C), and h′ (U) ·  =

 costt′ (U, C) = wcostt′ (U, C). Finally, notice that

Claim 1 holds when the h′ function is defned with respect to any choice of t ∈ {0, 1,..., q}. Therefore, with probability at least 1 − λ′ , the inequality in the lemma holds for every t ∈ {0, 1,..., q}, which completes the proof. 

Next, we show the following observation, whose proof is identical to an analogous proof in (Chen 2008).

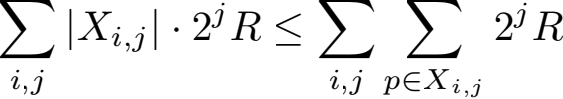
**Observation 2.** The following inequalities hold.

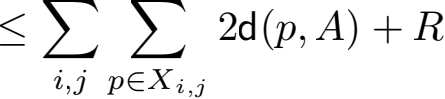
• Σi,j |Xi,j |2jR ≤ 3 · cost0 (X, A) ≤ 3τ · OPT(I).

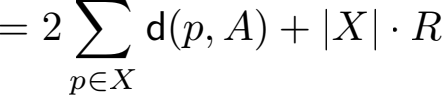
• Σi,j |Xi,j |diam(Xi,j ) ≤ 6·cost0 (X, A) ≤ 6τ·OPT(I).

*Proof.* For any p ∈ Xi,j , it holds that 2jR ≤

max{2d(p, A), R} ≤ 2d(p, A) + R. Therefore,







≤ 2 · cost0 (X, A) + nR

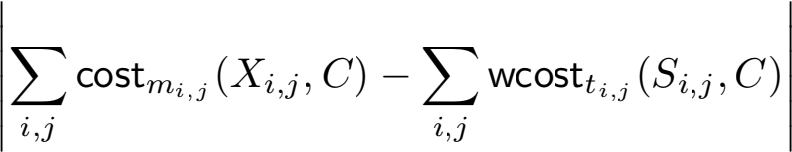
≤ 3 · cost0 (X, A) (By defnition of R) ≤ 3τOPT(I′ ) ≤ 3τOPT(I).

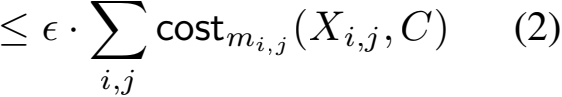
We also obtain the second item by observing that diam(Xi,j ) ≤ 2 · 2j · R. 

Next, we show that the following lemma, which infor- mally states that the union of the sets of sampled points ap- proximately preserve the cost of clustering w.r.t. *any* set of at most k centers, *even after* excluding at most m outliers overall.

**Lemma 2.** The following event happens with probability at least 1 − λ/2:

For all sets C ⊆ F of size at most k, and for all sets of non-negative integers {mi,j }i,j such that Σi,j mi,j ≤ m,





where ti,j = ⌊mi,j /wi,j ⌋ .

*Proof.* Fix an arbitrary set C ⊆ F of at most k centers,

and the integers {mi,j }i,j such that Σi,j mi,j ≤ m as in

the statement of the lemma. For each i = 1, . . . , |A|, and

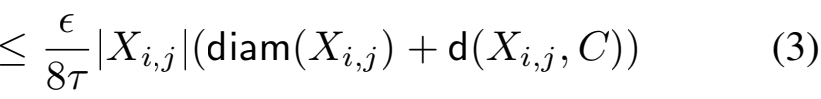
0 ≤ j ≤ ϕ, we invoke Lemma 1 by setting V = Xi,j ,

and U = Si,j , ξ =  , λ′ = n −kλ/(4(k + m)(1 + ϕ)),

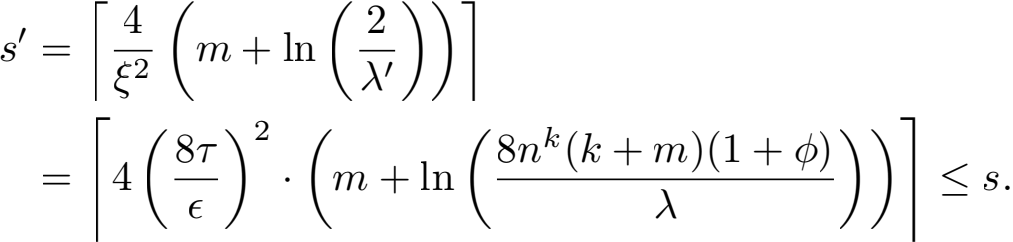
and q = m. This implies that, the following inequality holds

with probability at least 1 − λ′ for each set Xi,j , and the corresponding mi,j ≤ m,





Note that the sample size required in order for this in- equality to hold is



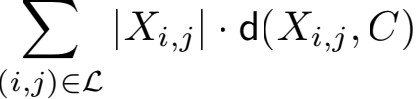
For any i,j, if Xi,j < s (i.e., Xi,j is *small*), then the sample Si,j is equal to Xi,j , and each point in Si,j has weight equal to 1. This implies that costmi,j (Xi,j , C) = wcostti,j (Si,j , C) for all such Xi,j , and their contribution to the right hand side of inequality (2) is zero. Thus, it suf- fces to restrict the sum on the right hand side of (2) over *large* sets Xi,j ’s. Let L consist of all pairs (i,j) such that Xi,j is large. We have the following claim.

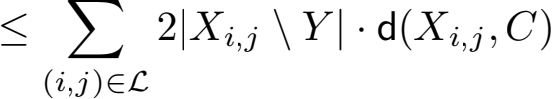
**Claim 2.** Σ (i,j)∈L |Xi,j |d(Xi,j , C) ≤ 2costm (X, C).

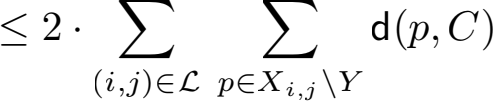
*Proof.* Let Y denote the farthest m points in X from the set of centers C. Now, fx (i,j) ∈ L and let qi,j := |Xi,j ∩Y | ≤ m denote the number of outliers in Xi,j . Since |Xi,j | ≥ 2m ≥ 2qi,j , the set Xi,j \ Y is non-empty, and all

points Xi,j \ Y contribute towards costm (X, C). That is, Σ Σ d(p, C) ≤ costm (X, C) (4) (i,j)∈Lp∈Xi,j \Y

For any p ∈ Xi,j \ Y, d(Xi,j , C) ≤ d(p, C) from the defnition. Therefore,





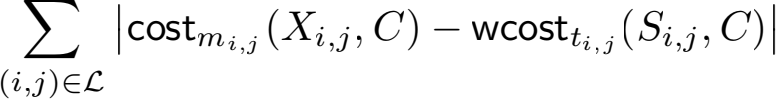


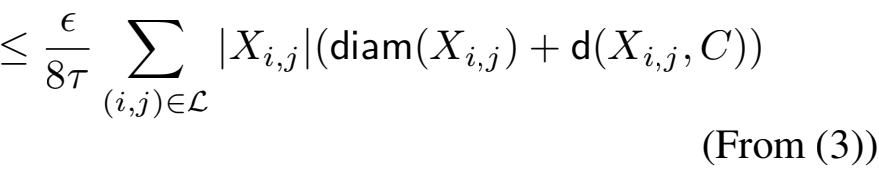
≤ 2 · costm (X, C)

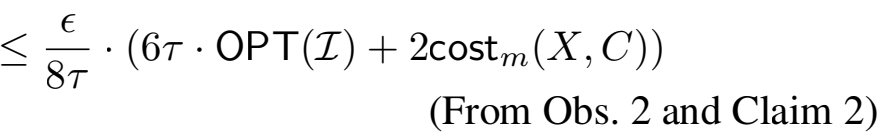
Here, to see the second inequality, see that |Xi,j | ≥ 2qi,j , which implies that |Xi,j | − qi,j ≤ 2(|Xi,j | − qi,j ). The last

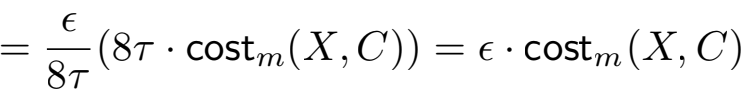
inequality follows from (4).

Thus, by revisiting (3) and (2), we get:







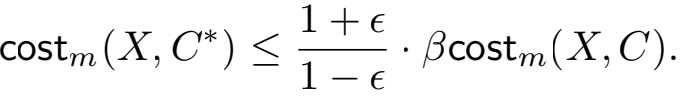


Where, in the last inequality, since C is an arbitrary set of at most k centers, OPT(I) ≤ costm (X, C). Note that the preceding inequality holds for a fxed set C of centers with probability at least 1−|A|·(1+ϕ)λ′ = 1−n−kλ/2, which follows from taking the union bound over all sets Xi,j , 1 ≤ i ≤ |A| ≤ k + m, and 0 ≤ j ≤ ϕ .

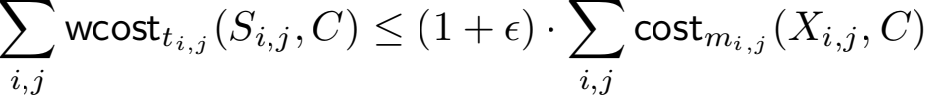
Since F has at most nk subsets of size at most k, the state- ment of the lemma follows from taking a union bound. i

Now we are ready to prove Theorem 1. We enumerate ev- ery subset T ⊆ S of size at most m. For each T, we compute a β-approximation solution for the (weighted) k-median in- stance (S\T,F, k). Theorem 1 only assumes the existence of a β-approximation algorithm for unweighted k-median, which cannot be applied to weighted point sets. However, we can transform S\T to an equivalent unweighted sets R, which contains, for each x ∈ S\T, w(x) copies of (un- weighted) x, where w(x) is the weight of x in S\T. It is clear that wcost(S\T, C) = cost(R, C) for all C ⊆ F. Thus, we can apply the β-approximation k-MEDIAN algo- rithm on (R,F, k) to compute a center set C ⊆ F of size k such that wcost(T, C) ≤ β · wcost(T, C′ ) for any C′ ⊆ F of size k. We do this for all T ⊆ S of size at most m. Let C denote the set of all center sets C computed. We pick a center set C\* ⊆ C that minimizes costm (X, C\*), and return (Y\* , C\*) as the solution where Y\* ⊆ X consists of them points in X farthest to the center set C\*.

**Lemma 3.** With probability at least 1 −  size k we have



*Proof.* The statement in Lemma 2 holds with probability at least 1 − λ/2. Thus, it suffces to assume the state- ment in Lemma 2, and show costm (X, C\*) ≤ (1 + ϵ)2 β · costm (X, C) for any C ⊆ F of size k. Fix a subset C ⊆ F of size k. Let Y ⊆ X consist of them points in X farthest to C, and defnemi,j = |Y ∩ Xi,j | . Set ti,j = ⌊mi,j /wi,j ⌋ . Note that costm (X, C) = Σi,j costmi,j (Xi,j , C). Further- more, by Lemma 2, we have

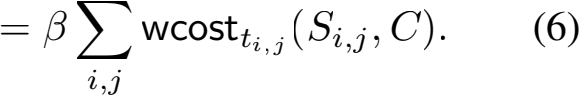


= (1 + ϵ) · costm (X, C). (5)

Now let Ti,j ⊆ Si,j consist of the ti,j points in Si,j far- thest to C, and defne T = ui,j Ti,j . Since |T| ≤ m, T is considered by our algorithm and thus there exists a center set C′ ∈ C that is a β-approximation solution for the (weighted)

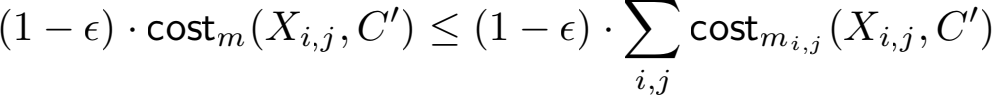
k-median instance (S\T,F, k). We have

wcost(S\T, C′ ) ≤ β · wcost(S\T, C)



Note that wcost(S\T, C′ ) ≥ Σi,j wcostti,j (Si,j , C′ ). Furthermore, by applying Lemma 2 again, we have Σi,j wcostti,j (Si,j , C′ ) ≥ (1−ϵ)·Σi,j costmi,j (Xi,j , C′ ).

It then follows that



≤ wcost(S\T, C′ ). (7) Finally, we have costm (X, C\*) ≤ costm (X, C′ ) by the con- struction of C\*. Combining this with (5), (6), and (7), we have costm (X, C\*) ≤  · βcostm (X, C), which com-

pletes the proof.

By choosing λ > 0 to be a suffciently small con- stant, and by appropriately rescaling ϵ 3 , the above lemma shows that our algorithm outputs a (β + ϵ)-approximation solution with a constant probability. By repeating the al- gorithm a logarithmic number of rounds, we can guaran- tee the algorithm succeeds with high probability. The num- ber of subsets T ⊆ S of size at most m is bounded by |S|O(m), which is ((k + m)log n/ϵ)O(m) by Proposition 1. Note that (log n)O(m) ≤ max{mO(m), nO(1)}. Thus, the number of subsets T ⊆ S of size at most m is bounded

by f(k,m,ϵ) · nO(1), where f(k,m,ϵ) = O(m) .

Thus, we need to call the β-approximation k-MEDIAN algo- rithmf(k,m,ϵ)·nO(1) times, which takes f(k,m,ϵ)nO(1)

·

T(n,k) time overall. The frst call of the algorithm for ob- taining a τ-approximation to the (k +m)-MEDIAN instance takes polynomial time. Besides this, the other parts of our algorithm can all be implemented in polynomial time. This completes the proof of Theorem 1.

**Extensions**

k**-Means with Outliers**

This is similar to k-MEDIANOUT, except that the cost func- tion is the sum of *squares* of distances of all except m out- lier points to a set of k facilities. This generalizes the well- known k-MEANS problem. Here, the main obstacle is that the squares of distances do not satisfy triangle inequality, and thus it does not form a metric. However, they satisfy a *relaxed* version of triangle inequality (i.e., d(p,q)2 ≤ 2(d(p,r)2 + d(r, q)2 )). This technicality makes the argu- ments tedious, nevertheless, we can follow the same ap- proach as for k-MEDIANOUT, to obtain optimal FPT ap- proximation schemes. Our technique implies an optimal (1+ 8/e + ϵ)-approximation fork-MEANSOUT (using the result of (Cohen-Addad et al. 2019) as a black-box), improving upon polynomial-time 53.002-approximation from (Krish- naswamy, Li, and Sandeep 2018), and (9+ϵ)-approximation from (Goyal,Jaiswal, and Kumar 2020) in time FPT ink, m and ϵ .

In fact, using our technique, we can get improved approx- imation guarantees for (k, z)-CLUSTERING WITH OUT- LIERS, where the cost function involves z-th power of dis- tances, where z ≥ 1 is fxed for a problem. Note that the cases z = 1 and z = 2 correspond to k-MEDIANOUT and k-MEANSOUT respectively. We give the details for (k, z)- CLUSTERING WITH OUTLIERS in the full version.

3 Since Lemma 3 implies a β(1 + O(ϵ))-approximation, and β is a constant, it suffces to redefne ϵ = ϵ/c for some large enough constant c to get the desired result.

**Matroid Median with Outliers**

A *matroid* is a pair M = (F, S), where F is a ground set, and S is a collection of subsets ofF with the following prop- erties: (i) ∅ ∈ S, (ii) If A ∈ S, then for every subset B ⊆ A, B ∈ S, and (iii) For any A, B ∈ S with |B| < |A|, there exists an b ∈ B \ A such that B ∪ {b} ∈ S. The rank of a matroid M is the size of the largest independent set in S.

An instance of MATROID MEDIAN WITH OUTLIERS is given by (X,F,M, m), where M = (F, S) is a matroid with rank k defned over a fnite ground set F, and X, F are sets of clients and facilities, belonging to a fnite metric space (Γ , d). The objective is to fnd a set C ⊆ F of facil- ities that minimizes costm (X, C), and C ∈ S, i.e., C is an independent set in the given matroid. Note that an explicit description of a matroid of rank k may be as large as nk. Therefore, we assume that we are given an effcient *oracle access* to thematroid M. That is, we are provided with an algorithm A that, given a candidate set S ⊆ F, returns in time T(A) (which is assumed to be polynomial in |F|), re- turns whether S ∈ I.

We can adapt our approach to MATROID MEDIAN WITH OUTLIERS in a relatively straightforward manner. Recall that our algorithm needs to start with an instance of outlier- free problem (i.e., MATROID MEDIAN) that provides a lower bound on the optimal cost of the given instance. To this end, given an instance I = (X,F, M = (F, S), m) of MATROID MEDIAN WITH OUTLIERS, we defne an in- stance I′ = (X,F, M′ ) MATROID MEDIAN, where M′ = (F ∪ X, S′ ) is defned as follows. S′ = {Y ∪ C : Y ⊆ X with |Y | ≤ mand C ⊆ F with C ∈ S}. That is, each in- dependent set of M′ is obtained by taking the union of an in- dependent set of facilities from M, and a subset of X of size at most m. It is straightforward to show that M′ is a matroid over the ground set F ∪ X. In particular, it is the direct sum of M and a uniform matroid over X of rankm (i.e., where any subset of X of size at most mis independent). Using the oracle algorithm A, we can simulate an oracle algorithm to test whether a candidate set C ⊆ F ∪ X is independent in M′ . Therefore, using a (2+ϵ)-approximation for MATROID MEDIAN (Cohen-Addad et al. 2019) in time FPT ink and ϵ, we can fnd a set A ⊆ F ∪ X of size at most k + mthat we can use to construct a coreset. The details about enumeration are similar to that fork-MEDIANOUT, and are thus omitted.

**Colorful** k**-Median**

This is an orthogonal generalization of k-MEDIANOUT to ensure a certain notion of *fairness* in the solution (see (Jia, Sheth, and Svensson 2020)). Suppose the set of points X is partitioned into ℓ different colors X1 ⊎ X2 ⊎ ... ⊎ Xℓ . We are also given the corresponding number of outliers m1 , m2 , . . . , mℓ . The goal is to fnd a set of at most facil- ities C to minimize the connection cost of all except at most mt outliers from each color class Xt , i.e., we want to min-

imize the cost function: Σ=1 costmt (Xt , C). This follows

a generalizations of the well-known k-CENTER problem in- troduced in (Bandyapadhyay et al. 2019) and (Anegg et al. 2020; Jia, Sheth, and Svensson 2020) , called COLORFUL k-CENTER. Similar generalization of FACILITY LOCATION

has also been studied in (Chekuri et al. 2022).

Using our ideas, we can fnd an FPT approximation pa-

rameterized by k, m = Σ=1 mt , and ϵ . To this end, we

sample suffciently many points from each color class Xt separately, and argue that it preserves the cost appropriately. The technical details follow the same outline as that for k- Median with m Outliers. In particular, during the enumer- ation phase—just like that for k-MEDIANOUT—we obtain several instances of k-MEDIAN. That is, our algorithm is *color-agnostic* after constructing the coreset. Thus, we ob- tain a tight (1 + 2/e + ϵ)-approximation for this problem. This is the frst non-trivial true approximation for this prob- lem – previous work (Gupta, Moseley, and Zhou 2021) only gives a *pseudo-approximation*, i.e., a solution with cost at most a constant times that of an optimal cost, but using slightly more thank facilities.

**A Combination of Above Generalizations**

Our technique also works for a combination of the afore- mentioned generalizations that are orthogonal to each other. To consider an extreme example, consider COLORFUL MA - TROID MEDIAN with ℓ different color classes (asimilar ver- sion for k-CENTER objective has been recently studied by (Anegg, Koch, and Zenklusen 2022)), where we want to fnd a set of facilities that is independent in the given matroid, in order to minimize the sum of distances of all except mt out- lier points for each color class Xt. By using a combination of the ideas mentioned above, one can get FPT approxima- tions for such generalizations.

**Concluding Remarks**

In this paper, we give a reduction from k-MEDIANOUT to k-MEDIAN that runs in time FPT in k,m, and ϵ, and pre- serves the approximation ratio up to an additive ϵ factor. As a consequence, we obtain improved FPT approximations for k-MEDIANOUT in general as well as special kinds of met- rics, and these approximation guarantees are known to be tight. Furthermore, our technique is versatile in that it also gives improved approximations for related clustering prob- lems, such as k-MEANSOUT, MATROID MEDIAN WITH OUTLIERS, and COLORFUL k-MEDIAN, among others.

The most natural direction is to improve the FPT run- ning time while obtaining the tight approximation ratios. More fundamentally, perhaps, is the question whether we need an FPT dependence on the number of outliers, m; or whether it is possible to obtain approximation guarantees for k-MEDIANOUT matching that for k-MEDIAN, with a run- ning time that is FPT ink and ϵ .

**Acknowledgments**

T. Inamdar is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819416). S. Saurabh is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 819416) and Swarnajayanti Fellowship (No. DST/SJF/MSA01/2017- 18).

**References**

Agrawal, A.; Inamdar, T.; Saurabh, S.; and Xue, J. 2022. Clustering What Matters: Optimal Approximation for Clus- tering with Outliers. arXiv:2212.00696.

Almanza, M.; Epasto, A.; Panconesi, A.; and Re, G. 2022. k-Clustering with Fair Outliers. In Candan, K. S.; Liu, H.; Akoglu, L.; Dong, X. L.; and Tang, J., eds., *WSDM ’22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, 5–15. ACM.

Anegg, G.; Angelidakis, H.; Kurpisz, A.; and Zenklusen, R. 2020. A Technique for Obtaining True Approximations for k-Center with Covering Constraints. In Bienstock, D.; and Zambelli,G., eds., *Integer Programming and Combinatorial Optimization - 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings*, volume 12125 of *Lecture Notes in Computer Science*, 52–65. Springer.

Anegg, G.; Koch, L. V.; and Zenklusen, R. 2022. Tech- niques for Generalized Colorful k-Center Problems. In Chechik, S.; Navarro, G.; Rotenberg, E.; and Herman, G., eds., *30th Annual European Symposium on Algorithms, ESA 2022, September 5-9, 2022, Berlin/Potsdam, Germany*, vol- ume 244 of *LIPIcs*, 7:1–7:14. Schloss Dagstuhl - Leibniz-

Zentrum fr Informatik.

Bandyapadhyay, S.; Inamdar, T.; Pai, S.; and Varadarajan, K. R. 2019. A Constant Approximation for Colorful k- Center. In Bender, M. A.; Svensson, O.; and Herman, G., eds., *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Ger- many*, volume 144 of *LIPIcs*, 12:1–12:14. Schloss Dagstuhl

- Leibniz-Zentrum fr Informatik.

Blmer, J.; Lammersen, C.; Schmidt, M.; and Sohler, C. 2016. Theoretical analysis of the k-means algorithm–a sur- vey. In *Algorithm Engineering*, 81–116. Springer.

Charikar, M.; Khuller, S.; Mount, D. M.; and Narasimhan, G. 2001. Algorithms for facility location problems with out- liers. In Kosaraju, S. R., ed., *Proceedings of the Twelfth An- nual Symposium on Discrete Algorithms, January 7-9, 2001, Washington, DC, USA*, 642–651. ACM/SIAM.

Chekuri, C.; Inamdar, T.; Quanrud, K.; Varadarajan, K.; and Zhang, Z. 2022. Algorithms for covering multiple submod- ular constraints and applications. *Journal of Combinatorial Optimization*, 1–32.

Chen, K. 2008. A constant factor approximation algorithm for*k*-median clustering with outliers. In Teng, S., ed., *Pro- ceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, Cali- fornia, USA, January 20-22, 2008*, 826–835. SIAM.

Chen, K. 2009. On coresets fork-median and k-means clus- tering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3): 923–947.

Cohen-Addad, V.; Gupta, A.; Kumar, A.; Lee, E.; and Li, J. 2019. Tight FPT Approximations for k-Median and k- Means. In Baier, C.; Chatzigiannakis, I.; Flocchini, P.; and Leonardi, S., eds., *46th International Colloquium on Au- tomata, Languages, and Programming, ICALP 2019, July*

*9-12, 2019, Patras, Greece*, volume 132 of *LIPIcs*, 42:1–

42:14. Schloss Dagstuhl - Leibniz-Zentrum fr Informatik.

Cohen-Addad, V.; Saulpic, D.; and Schwiegelshohn, C. 2021. A new coreset framework for clustering. In Khuller, S.; and Williams, V. V., eds., *STOC ’21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, 169–182. ACM.

Feng, Q.; Zhang, Z.; Huang, Z.; Xu, J.; and Wang, J. 2019. Improved Algorithms for Clustering with Outliers. In Lu, P.; and Zhang, G., eds., *30th International Sympo- sium on Algorithms and Computation, ISAAC 2019, De- cember 8-11, 2019, Shanghai University of Finance and Economics, Shanghai, China*, volume 149 of *LIPIcs*, 61:1–

61:12. Schloss Dagstuhl - Leibniz-Zentrum fr Informatik.

Friggstad, Z.; Khodamoradi, K.; Rezapour, M.; and Salavatipour, M. R. 2019. Approximation schemes for clustering with outliers. *ACM Transactions on Algorithms (TALG)*, 15(2): 1–26.

Friggstad, Z.; Rezapour, M.; and Salavatipour, M. R. 2019. Local Search Yields a PTAS for k-Means in Doubling Met- rics. *SIAM J. Comput.*, 48(2): 452–480.

Goyal, D.; Jaiswal, R.; and Kumar, A. 2020. FPT Approx- imation for Constrained Metric k-Median/Means. In Cao, Y.; and Pilipczuk, M., eds., *15th International Symposium on Parameterized and Exact Computation, IPEC 2020, De- cember 14-18, 2020, Hong Kong, China (Virtual Confer- ence)*, volume 180 of *LIPIcs*, 14:1–14:19. Schloss Dagstuhl

- Leibniz-Zentrum fr Informatik.

Guha, S.; and Khuller, S. 1999. Greedy Strikes Back: Im- proved Facility Location Algorithms. *J. Algorithms*, 31(1): 228–248.

Gupta, A.; Moseley, B.; and Zhou, R. 2021. Structural It- erative Rounding for Generalized k-Median Problems. In Bansal, N.; Merelli, E.; and Worrell, J., eds., *48th Interna- tional Colloquium on Automata, Languages, and Program- ming, ICALP 2021, July 12-16, 2021, Glasgow, Scotland (Virtual Conference)*, volume 198 of *LIPIcs*, 77:1–77:18.

Schloss Dagstuhl - Leibniz-Zentrum fr Informatik.

Gupta, S.; Kumar, R.; Lu, K.; Moseley, B.; and Vassilvitskii, S. 2017. Local Search Methods for k-Means with Outliers. *Proc. VLDB Endow.*, 10(7): 757–768.

Haussler, D. 1992. Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applica- tions. *Inf. Comput.*, 100(1): 78–150.

Jia, X.; Sheth, K.; and Svensson, O. 2020. Fair Colorful k-Center Clustering. In Bienstock, D.; and Zambelli, G., eds., *Integer Programming and Combinatorial Optimization - 21st International Conference, IPCO 2020, London, UK, June 8-10, 2020, Proceedings*, volume 12125 of *Lecture Notes in Computer Science*, 209–222. Springer.

Krishnaswamy, R.; Li, S.; and Sandeep, S. 2018. Constant approximation fork-median and k-means with outliers via iterative rounding. In Diakonikolas, I.; Kempe, D.; and Henzinger, M., eds., *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, 646–659. ACM.

Rokach, L. 2009. A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, 269–298. Springer.

Statman, A.; Rozenberg, L.; and Feldman, D. 2020. k- Means: Outliers-Resistant Clustering+++. *Algorithms*, 13(12): 311.

Xu, D.; and Tian, Y. 2015. A comprehensive survey of clus- tering algorithms. *Annals of Data Science*, 2(2): 165–193.