Article

k-Means+++: Outliers-Resistant Clustering

Adiel Statman [1], Liat Rozenberg [1,2] and Dan Feldman [1]

[1] Robotics & Big Data Lab, Computer Science Department, University of Haifa, Israel
[2] School of Information and Communication Technology, Griffith University, Australia

Abstract: The k-means problem is to compute a set of k centers (points) that minimizes the sum of squared distances to a given set of n points in a metric space. Arguably, the most common algorithm to solve it is k-means++ which is easy to implement, and provides a provably small approximation factor in time that is linear in n.
We generalize k-means++ to support: (i) non-metric spaces and any pseudo-distance function. In particular, it supports M-estimators functions that handle outliers, e.g. where the distance $\text{dist}(p, x)$ between a pair of points is replaced by $\min\{\text{dist}(p, x), 1\}$. (ii) k-means clustering with $m \geq 1$ outliers, i.e., where the m farthest points from the k centers are excluded from the total sum of distances. This is the first algorithm whose running time is linear in n and polynomial in k and m.

Keywords: Clustering, Approximation, Outliers

1. Introduction

In this section we introduce the notion of clustering, the special case of k-means++, the generalization to outliers and our contribution.

1.1. Clustering

For a given similarity measure, clustering is the problem of partitioning a given set of n objects into subsets, such that objects in the same group are more similar to each other, than to objects in the other sets. As mentioned in [1], clustering problems arise in many different applications, including data mining and knowledge discovery [2], data compression and vector quantization [3], pattern recognition and classification [4]. However, for most of its variants it is an NP-Hard problem when the number k of clusters is part of the input, as elaborated and proved in [5,6].

Hence, as in this paper, constant or near-constant (logarithmic in n or k) multiplicative factor approximations to the desired cost function were suggested over the years, whose running time is polynomial in both n and k. A more general approximations, called bi-criteria or (α, β) approximations, guarantee multiplicative factor α approximation but the number of used center (for approximating the optimal k centers) is βk. The factors α and β might be depended on k and n, and different methods give different dependencies and running times. For example, [7] and [8] provide an approximation algorithm of $\alpha \in O(\log k)$ and $\beta = 1$. A constant $\alpha = O(1)$ approximation was suggested by [8] under the assumption that the data is well separated in some formal sense. Consequently, [9] proved that sampling $O(k)$ centers using the seeding method from [7] yields a constant-factor approximation and [10] showed that sampling $O(k \log k)$ different randomized centers yields an $O(1)$-approximation, and leverage it to support streaming data. Recently, [11] provided a constant approximation for k-means, i.e $\alpha \in O(1)$ in time complexity of $O(dnk^2 \log\log k)$ where d is the dimension of the Euclidean data. The analysis of [12] limits α as a function of β.

Approximations for the k median problem, where sum instead of sum of squared distances is used, were suggested in [13–15]. However, the k-means++ algorithm supports any distance to the power of $z \geq 1$ as explained in [7].

1.2. Lloyd's k-means

As explained in [8], Lloyd [16–18] suggested a simple iterative heuristic that aims to minimize the clustering cost, assuming a solution to the case $k = 1$ is known. It is a special case of the EM (Expected Maximization) heuristic for computing a local minimum. The algorithm is initialized with k random points (seeds, centroids). At each iteration, each of the input points is classified to its closest centroid. A new set of k centroids is constructed by taking the mean (or solving the problem for $k = 1$, in the general case) of each of the current k clusters. This method is repeated until convergence or any given stopping condition.

Due to its simplicity, and the convergence to a local minimum [19], this method is very common; see [3,20–25] and references therein. The method has further improved in [1,26–28].

The drawback of this approach is that it converges to a local minimum - the one which is closest to the initial centers that had been chosen and may be arbitrarily far from the global minimum. There is also no upper bound for the convergence rate and number of iterations. Therefore, a lot of research has been done to choose good initial points, called "seeds" [29–35]. However, very few analytical guarantees were found to prove convergence.

*1.3. k-*MEANS++

As mentioned above, the initialization of the k-means algorithm depends on its initial seeds, which might be far from the optimal centers. The k-MEANS++ algorithm, proposed by [7,8], aims to handle this problem. It suggests a provable approximation to the k-means problem. It then calls Lloyd's algorithm as a heuristic that hopefully improves the approximation in practice. Since Lloyd's algorithm can only improve the initial solution, the provable upper bound on the approximation factor is still preserved. The original papers for k-means++ [7,8] suggested $O(\log k)$-approximations, but recent analysis [11] suggest variants that give $O(1)$-approximations. The k-Means++ algorithm is based on the intuition that the centroids should be well spread out. Hence, instead of starting with k random points, it samples k centers iteratively, via a distribution that is called D^2-sampling and is proportional to the distance of each input point to the centers that were already chosen. The first center is chosen uniformly at random from the input.

1.4. Clustering with outliers

In practice, data sets include some noise measurements which do not reflect a real part of the data. These are called outliers, and even a single outlier may completely change the optimal solution that is obtained without this outlier. One option to handle outliers is to change the distance function to a function that is more robust to outliers, such as M-estimators, e.g. where the distance $\text{dist}(p, x)$ between a pair of points is replaced by $\min\{\text{dist}(p, x), c\}$ for some fixed $c > 0$. Another option is to compute the set of k centers that minimizes the objective function, excluding the farthest m points from the candidate k centers. Here, $m \geq 1$ is a given parameter for the number of outliers. Of course, given the optimal k centers for this problem, the m outliers are simply the farthest m input points, and given these m outliers the optimal solution is the k-means for the rest of the points. However, the main challenge is to approximate the global optimum, i.e., compute the optimal centers and outliers simultaneously.

As explained in [36], detecting the ouliers themselves is also an NP-hard problem [37]. An intensive research has been done on this problem as explained in [38]) since it has numerous applications in many areas [39,40]). In the context of data mining, [41] proposed a definition of distance-based outlier, which is free of any distributional assumptions and it can be generalized to multidimensional datasets. Following

[41], further variations have been proposed [41–43]). Consequently, [44] introduced paradigm of local outlier factor (LOF). This paradigm has been extended in [39,45] in different directions.

As explained in [46], and following the discussion in [47], [48] provided an algorithm based on Lagrange-relaxation technique. Several algorithms [47,49,50] were also developed. The work of [51] gives a factor of $O(1)$ and a running time of $O(n^m)$. Other heuristic was developed by [52]. Finally, [46] provided an $O(1)$- approximation for the k-median problem (sum of distances in a metric space) in $O(k^2(k+m)^2n^3 \log n)$ time. In the context of k-means, [53] provided several algorithms of such constant factor approximation. However,the number of of the points which approximate the outliers is much greater than m, and is depend on the data, as well as the algorithm running time.

1.5. Our contribution

The k-MEANS++ algorithm was proved to hold for any distance to the power of $z \geq 1$. We generalize the algorithm to non-metric functions that only satisfy the weak triangle inequality (see e.g. [54]). This family of functions includes most of the M-estimators, including non-convex functions. In addition, we suggest a coreset for the k-median with m outliers, i.e., a subset of $O(k+m)$ input points that approximates the sum of distances from the original input to any given k centers, excluding their farthest m points, up to a factor guaranteed for clustering without outliers. We then apply the result of Ke-Chen [46] on this coreset to obtain a $O(\log(m+k))$ factor approximation to the k-median with outliers to obtain the first provable approximation to this problem that takes time that is linear in n, and also polynomial in both m and k. This coreset also provides a constant approximation for k-means with outliers that takes time that is linear in n for a set in $\mathbb{R}^d$. In addition, our version supports multiplicative weights for the input points.

The results are summarized in the following table. It contains the three problems which we improve or generalize in this paper: k-means seeding, k-median with m outliers and k-means with m outliers. We focus only on approximations that returns exactly k centres (and m outliers) and do not assume specific properties regarding the input. The first two rows show the state-of-the-art results, and the last three rows are the results of this paper. The parameter τ denotes the time it takes to compute the distance between a pair of points in the metric space as defined in the end of Section 2.

Problem name	Reference	Input space	Distance function	Approx. factor	Time complexity
k-means seeding	[7]	$\mathbb{R}^d$	$\|\cdot\|^z$ where $z \in \mathbb{N}, z \geq 2$	$O(\log k)$	$O(nkd)$
k-median m outliers	[46]	metric space	metric	$\mathbf{O(1)}$	$O(\tau k^2(k+m)^2n^3 \log n)$
k-means seeding	Theorem 7	**metric space**	**ρ-distance** See Definition 1	$O(\log k)$	$O(nk\tau)$
k-median m outliers	Corollary 10	metric space	metric	$O(\log k)$	$\boldsymbol{O(\tau n(k+m) + \tau k^2(k+m)^5 \log(k+m))}$
k-means m outliers	Corollary 12	$\mathbb{R}^d$	$\|\cdot\|^2$	$O(1)$	$O(dn(k+m)^2 \log\log(k+m) + d2^{k+m})$

2. Notation

Let P be a set of size n. We generalize the definition of distance function over P as follows.

Definition 1. *Let $\rho > 0$, and let $f : P^2 \to [0, \infty)$ be a symmetric function that satisfies $f(p,p) = 0$ for every $p \in P$. f is called a ρ-distance function over P iff the following "approximated" triangle inequality holds: for every $p, q, x \in P$*

$$f(q,x) \leq \rho\big(f(q,p) + f(p,x)\big). \tag{1}$$

For a point $p \in P$ and a set $X \subseteq P$, we define $f(p, X) = \min_{x \in X} f(p, x)$. The minimum or sum over an empty set is defined to be zero. Let $w : P \to (0, \infty)$ be called the *weight function* over P. For a non-empty set $Q \subseteq P$ we define

$$f(Q, X) = f_w(Q, X) = \sum_{p \in Q} w(p) f(p, X).$$

If Q is empty then $f(Q, X) = 0$. For brevity, we denote $f(Q, p) = f(Q, \{p\})$, and $f^2(\cdot) = (f(\cdot))^2$. Let $k \geq 1$ be an integer and denote $[k] = \{1, \cdots, k\}$. The *k-mean* of a set $Q \subseteq P$ is

$$f^*(Q, k) = f^*_w(Q, k) = \min_{X \subseteq Q, |X| = k} f(Q, X).$$

Let $\tau > 0$ be an upper bound on the time it takes to compute $f(p, q)$ between any two point $p, q \in P$.

3. Generality of ++ bounding

In this section we suggest a generalization of the k-means++ algorithm which for every ρ-distance function.

Lemma 1. *For every non-empty set $Q \subseteq P$,*

$$\sum_{x \in Q} w(x) f_w(Q, x) \leq 2\rho f^*_w(Q, 1) \sum_{x \in Q} w(x). \tag{2}$$

Proof. Let p^* be the weighted mean of Q, i.e., $f_w(Q, p^*) = f^*_w(Q, 1)$. By (1), for every $q, x \in Q$,

$$f(q, x) \leq \rho\big(f(q, p^*) + f(p^*, x)\big).$$

Summing over every weighted $q \in Q$ yields

$$f_w(Q, x) \leq \rho\big(f_w(Q, p^*) + f(p^*, x) \sum_{q \in Q} w(q)\big) = \rho\big(f^*_w(Q, 1) + f(x, p^*) \sum_{q \in Q} w(q)\big).$$

Summing again over the weighted points of Q yields

$$\sum_{x \in Q} w(x) f_w(Q, x) \leq \rho \sum_{x \in Q} w(x) \big(f^*_w(Q, 1) + f(x, p^*) \sum_{q \in Q} w(q)\big) = 2\rho f^*_w(Q, 1) \sum_{x \in Q} w(x).$$

□

Lemma 2 (See 6.1 in [54]). *For every pair of points $q, x \in P$ and a subset $X \subseteq P$ we have*

$$f(x, X) \leq \rho(f(q, x) + f(q, X)). \tag{3}$$

Corollary 3 (Approximated triangle inequality). *Let $x \in P$, and $Q, X \subseteq P$ be non-empty sets. Then*

$$f(x, X) \sum_{q \in Q} w(q) \leq \rho(f_w(Q, x) + f_w(Q, X)).$$

Proof. Summing (3) over every weighted $q \in Q$ yields

$$\begin{aligned} f(x, X) \sum_{q \in Q} w(q) &\leq \rho \left(\sum_{q \in Q} w(q) f(q, x) + \sum_{q \in Q} w(q) f(q, X) \right) \\ &= \rho(f_w(Q, x) + f_w(Q, X)). \end{aligned}$$

□

Lemma 4. *Let $Q, X \subseteq P$ such that $f_w(Q, X) > 0$. Then*

$$\frac{1}{f_w(Q,X)} \sum_{x \in Q} w(x) f(x,X) \cdot f_w(Q, X \cup \{x\}) \sum_{q \in Q} w(q) \leq 2\rho \sum_{x \in Q} w(x) f_w(Q,x).$$

Proof. By Corollary 3, for every $x \in Q$,

$$f(x,X) \sum_{q \in Q} w(q) \leq \rho(f_w(Q,x) + f_w(Q,X)).$$

Multiplying this by $\frac{f_w(Q, X \cup \{x\})}{f_w(Q,X)}$ yields

$$\begin{aligned} f(x,X) \cdot \frac{f_w(Q, X \cup \{x\})}{f_w(Q,X)} \sum_{q \in Q} w(q) &\leq \rho\big(f_w(Q,x) \cdot \frac{f_w(Q, X \cup \{x\})}{f_w(Q,X)} + f_w(Q, X \cup \{x\})\big) \\ &\leq \rho\big(f_w(Q,x) + f_w(Q, X \cup \{x\})\big) \leq 2\rho f_w(Q,x). \end{aligned}$$

After summing over every weighted point $x \in Q$ we obtain

$$\sum_{x \in Q} w(x) f(x,X) \cdot \frac{f_w(Q, X \cup \{x\})}{f_w(Q,X)} \sum_{q \in Q} w(q) \leq 2\rho \sum_{x \in Q} w(x) f_w(Q,x).$$

□

Algorithm 1: CLUSTERING++(P, w, f, X, t); see Theorem 7.

```
Input  : A finite set $P \subseteq \mathbb{R}^d$, a function $w : P \to [0, \infty)$, a subset $X \subseteq P$, an integer
         $t \in [0, |P| - |X|]$ and a $\rho$-distance $f$ over $P$.
Output: $Y \subseteq P$.
$Y := X$
for $i := 1$ to $t$ // If $t = 0$ then skip this "for" loop
do
    For every $p \in P$, $\Pr_i(p) = \frac{w(p) f(p,Y)}{\sum_{q \in P} w(Y) f(q,Y)}$ // $f(p, \emptyset) := 1$.
    Pick a random point $y_i$ from $P$, where $y_i = p$ with probability $\Pr_i(p)$ for every $p \in P$.
    $Y := X \cup \{y_1, \cdots, y_i\}$
return Y
```

Lemma 5. *Let $t, u \geq 0$ be a pair of integers. If $u \in [k]$ and $t \in \{0, \cdots, u\}$ then the following hold. Let $\{P_1, \cdots, P_k\}$ be a partition of P such that $\sum_{i=1}^{k} f^*(P_i, 1) = f^*(P, k)$. Let $U = \bigcup_{i=1}^{u} P_i$ denote the union of the first u sets. Let $X \subseteq P$ be a set that covers $P \setminus U$, i.e., $X \cap P_i \neq \emptyset$ for every integer $i \in [k] \setminus [u]$, and $X \cap U = \emptyset$. Let Y be the output of a call to* CLUSTERING++(P, w, f, X, t)*; see Algorithm 1. Then*

$$E[f(P,Y)] \leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U,u)\big) H_t + \frac{u-t}{u} \cdot f(U,X), \tag{4}$$

where $H_t = \begin{cases} \sum_{i=1}^{t} \frac{1}{i} & \text{if } t \geq 1 \\ 1 & \text{if } t = 0 \end{cases}$, and the randomness is over the t sampled points in $Y \setminus X$.

Proof. The proof is by the following induction on $t \geq 0$: (i) the base case $t = 0$ (and any $u \geq 0$), and (ii) the inductive step $t \geq 1$ (and any $u \geq 0$).

(i) Base Case : $t = 0$. We assume $u \in [k]$, otherwise the lemma trivially holds. We then have

$$\begin{aligned} E[f(P,Y)] &= E[f(P,X)] = f(P,X) = f(P \setminus U, X) + f(U,X) = f(P \setminus U, X) + \frac{u-t}{u} \cdot f(U,X) \\ &\leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U,u)\big)H_t + \frac{u-t}{u} \cdot f(U,X), \end{aligned}$$

where the first two equalities hold since $Y = X$ is not random, and the inequality holds since $t = 0$, $H_0 = 1$ and $f^*(U,u) \geq 0$. Hence, the lemma holds for $t = 0$ and any $u \geq 0$.

(ii) Inductive step: $t \geq 1$. We assume $u \in [k]$ and $t \in [u]$, otherwise the lemma trivially holds. Let $y \in P$ denote the first sampled point that was inserted to $Y \setminus X$ during the execution of Line 5 in Algorithm 1. Let $X' = X \cup \{y\}$. Let $j \in [k]$ such that $y \in P_j$, and $U' = U \setminus P_j$ denote the remaining "uncovered" $u' = |\{P_1, \cdots, P_u\} \setminus P_j|$ clusters, i.e, $u' \in \{u, u-1\}$. The distribution of Y conditioned on the known sample y is the same as the output of a call to CLUSTERING++(P, w, f, X', t') where $t' = t - 1$. Hence, we need to bound

$$E[f(P,Y)] = \Pr(y \in P \setminus U)E[f(P,Y) \mid y \in P \setminus U] + \Pr(y \in U)E[f(P,Y) \mid y \in U]. \tag{5}$$

We will bound each of the last two terms by expressions that are independent of X' or U'.

Bound on $E[f(P,Y) \mid y \in P \setminus U]$. Here $u' = u \in [k]$, $U' = U$ (independent of j), and by the inductive assumption that the lemma holds after replacing t with $t' = t - 1$, we obtain

$$\begin{aligned} E[f(P,Y) \mid y \in P \setminus U] &\leq \big(f(P \setminus U', X') + 4\rho^2 f^*(U', u')\big)H_{t'} + \frac{u'-t'}{u'} \cdot f(U', X') \\ &= \big(f(P \setminus U, X') + 4\rho^2 f^*(U, u)\big)H_{t-1} + \frac{u-t+1}{u} \cdot f(U, X') \\ &\leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U, u)\big)H_{t-1} + \frac{u-t+1}{u} \cdot f(U, X). \end{aligned} \tag{6}$$

Bound on $E[f(P,Y) \mid y \in U]$. In this case, $U' = U \setminus P_j$ and $u' = u - 1 \in \{0, \cdots, k-1\}$. Hence,

$$\begin{aligned} E[f(P,Y) \mid y \in U] &= \sum_{m=1}^{u} \Pr(j = m)E[f(P,Y) \mid j = m] \\ &= \sum_{m=1}^{u} \Pr(j = m) \sum_{x \in P_m} \Pr(y = x)E[f(P,Y) \mid x = y]. \end{aligned} \tag{7}$$

Put $m \in [u]$ and $x \in P_m$. We remove the above dependency of $E[f(P,Y) \mid y \in U]$ upon x and then m.

We have

$$\begin{aligned} E[f(P,Y) \mid y = x] &\leq \big(f(P \setminus U', X') + 4\rho^2 f^*(U', u')\big)H_{t-1} + \frac{u'-t'}{\max\{u', 1\}} \cdot f(U', X') \\ &= \big(f(P \setminus U, X') + f(P_m, X') + 4\rho^2 f^*(U, u) - 4\rho^2 f^*(P_m, 1)\big)H_{t-1} + \frac{u-t}{\max\{u', 1\}} \cdot f(U', X') \\ &\leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U, u)\big)H_{t-1} + \frac{u-t}{\max\{u', 1\}} \cdot f(U \setminus P_m, X) \\ &\quad + \big(f(P_m, X') - 4\rho^2 f^*(P_m, 1)\big)H_{t-1}, \end{aligned} \tag{8}$$

where the first inequality holds by the inductive assumption if $u' \geq 1$, or since $U' = U \setminus P_j = \emptyset$ if $u' = 0$. The second inequality holds since $X \subseteq X' = X \cup \{x\}$, and since $f^*(U \setminus P_m, u - 1) = f^*(U, u) - f^*(P_m, 1)$.

Only the term $f(P_m, X') = f(P_m, X \cup \{x\})$ depends on x, and not only on m. Summing it over every possible $x \in P_m$ yields

$$\begin{aligned}\sum_{x\in P_m} \Pr(y = x) f_w(P_m, X') &= \frac{1}{f_w(P_m, X)} \sum_{x\in P_m} w(x) f(x, X) f_w(P_m, X \cup \{x\}) \\ &\leq \frac{2\rho}{\sum_{q\in P_m} w(q)} \sum_{x\in P_m} w(x) f_w(P_m, x) \\ &\leq \frac{2\rho}{\sum_{q\in P_m} w(q)} \cdot 2\rho f^*_w(P_m, 1) \sum_{x\in P_m} w(x) \leq 4\rho^2 f^*(P_m, 1),\end{aligned}$$

where the inequalities follow by substituting $Q = P_m$ in Lemma 4 and Lemma 1, respectively. Hence, the expected value of (8) over x is non-positive as

$$\sum_{x\in P_m} \Pr(y = x)\big(f(P_m, X') - 4\rho^2 f^*(P_m, 1)\big) = -4\rho^2 f^*(P_m, 1) + \sum_{x\in P_m} \Pr(y = x) f_w(P_m, X') \leq 0.$$

Combining this with (7) and then (8) yields a bound on $E[f(P, Y) \mid y = x]$ that is independent upon x,

$$\begin{aligned}E[f(P, Y) \mid y \in U] &= \sum_{m=1}^{u} \Pr(j = m) \sum_{x\in P_m} \Pr(y = x) E[f(P, Y) \mid y = x] \\ &\leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U, u)\big) H_{t-1} + \frac{u - t}{\max\{u', 1\}} \sum_{m=1}^{u} \Pr(j = m) f(U \setminus P_m, X),\end{aligned} \tag{9}$$

It is left to remove the dependency on m, which occurs in the last term $f(U \setminus P_m, X)$ of (9). We have

$$\begin{aligned}\sum_{m=1}^{u} \Pr(j = m) f(U \setminus P_m, X) &= \sum_{m=1}^{u} \frac{f(P_m, X)}{f(P, X)} \big(f(U, X) - f(P_m, X)\big) \\ &= \frac{1}{f(P, X)} \left(f^2(U, X) - \sum_{m=1}^{u} f^2(P_m, X)\right).\end{aligned} \tag{10}$$

By Jensen's (or power-mean) inequality, for every convex function $g : \mathbb{R} \to \mathbb{R}$, and a real vector $v = (v_1, \cdots, v_u)$ we have $(1/u)\sum_{m=1}^{u} g(v_m) \geq g((1/u)\sum_{m=1}^{u} v_m)$. Specifically, for $g(z) := z^2$ and $v = \big(f(P_1, X), \cdots, f(P_u, X)\big)$,

$$\sum_{m=1}^{u} \frac{1}{u} f^2(P_m, X) \geq \left(\frac{1}{u} \sum_{m=1}^{u} f(P_m, X)\right)^2.$$

Multiplying by u yields

$$\sum_{m=1}^{u} f^2(P_m, X) \geq \frac{f^2(U, X)}{u}.$$

Plugging this in (10) gives the desired bound on the term $f(U', X)$,

$$\sum_{m=1}^{u} \Pr(j = m) f(U \setminus P_m, X) \leq \left(1 - \frac{1}{u}\right) \frac{f^2(U, X)}{f(P, X)} = \frac{u'}{u} \frac{f^2(U, X)}{f(P, X)} \leq \frac{\max\{u', 1\}}{u} f(U, X),$$

where the last inequality holds since $f(U, X) \leq f(P, X)$. Plugging the last inequality in (9) bounds $E[f(P, Y) \mid y \in U]$ by

$$E[f(P, Y) \mid y \in U] \leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U, u)\big) H_{t-1} + \frac{u - t}{u} f(U, X). \tag{11}$$

Bound on $E[f(P,Y)]$. Plugging (11) and (6) in (5) yields

$$E[f(P,Y)] \leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U,u)\big)H_{t-1} + f(U,X)\left(\Pr(y \in P \setminus U) \cdot \frac{u-t+1}{u} + \Pr(y \in U) \cdot \frac{u-t}{u}\right). \tag{12}$$

Firstly, we have

$$\Pr(y \in P \setminus U) \cdot \frac{u-t+1}{u} + \Pr(y \in U) \cdot \frac{u-t}{u} = \frac{u-t}{u} + \Pr(y \in P \setminus U) \cdot \frac{1}{u} \leq \frac{u-t}{u} + \Pr(y \in P \setminus U) \cdot \frac{1}{t},$$

where the last inequality holds as $u \geq t$. Secondly, since $U \subseteq P$,

$$f(U,X)\Pr(y \in P \setminus U) = f(U,X)\frac{f(P \setminus U, X)}{f(P,X)} \leq f(P \setminus U, X).$$

Hence, we can bound (12) by

$$\begin{aligned} f(U,X)\left(\Pr(y \in P \setminus U)\frac{u-t+1}{u} + \Pr(y \in U) \cdot \frac{u-t}{u}\right) &\leq \frac{u-t}{u} \cdot f(U,X) + f(U,X)\Pr(y \in P \setminus U) \cdot \frac{1}{t} \\ &\leq \frac{u-t}{u} \cdot f(U,X) + \big(f(P \setminus U, X) + 4\rho^2 f^*(U,u)\big) \cdot \frac{1}{t}. \end{aligned}$$

This proves the inductive step and bounds (12) by

$$E[f(P,Y)] \leq \big(f(P \setminus U, X) + 4\rho^2 f^*(U,u)\big)H_t + \frac{u-t}{u} \cdot f(U,X).$$

□

Corollary 6. *Let $\delta \in (0,1]$ and let q_0 be a point that is sampled at random from a non-empty set $Q \subseteq P$ such that $q_0 = q$ with probability $\frac{w(q)}{\sum_{q' \in Q} w(q')}$. Then with probability at least $1 - \delta$,*

$$f_w(Q, \{q_0\}) \leq \frac{2}{\delta}\rho f_w^*(Q,1).$$

Proof. By Markov's inequality, for every non-negative random variable G and $\delta > 0$ we have

$$\Pr\{G < \frac{1}{\delta}E[G]\} \geq 1 - \delta. \tag{13}$$

Substituting $G = f_w(Q, \{q_0\})$ yields

$$\Pr\left\{f_w(Q,\{q_0\}) < \frac{1}{\delta}\sum_{q_0 \in Q}\frac{w(q_0)}{\sum_{q \in Q} w(q)} f_w(Q,\{q_0\})\right\} \geq 1 - \delta. \tag{14}$$

By Lemma 1 we have

$$\sum_{q_0 \in Q}\frac{w(q_0)}{\sum_{q \in Q} w(q)} f_w(Q,\{q_0\}) \leq 2\rho \cdot f_w^*(Q,1). \tag{15}$$

Plugging (15) in (14) yields,

$$\Pr\Big\{f_w(Q,\{q_0\}) < \frac{2}{\delta}\rho \cdot f_w^*(Q,1)\Big\} \geq \Pr\Big\{f_w(Q,\{q_0\}) < \frac{1}{\delta}\sum_{q_0\in Q}\frac{w(q_0)}{\sum_{q\in Q} w(q)} f_w(Q,\{q_0\})\Big\}$$
$$\geq 1-\delta.$$

□

The following theorem is a particular case of Lemma 5. It proves that the output of CLUSTERING++; see Algorithm 1, is a $O(\log k)$-approximation of its optimum.

Theorem 7. *Let P be a set of points, $k \geq 2$ be an integer, $\rho > 0$, and let $f : P^2 \to [0,\infty)$ be a ρ-distance function over P; See Definition 1. Let $w : P \to (0,\infty)$, $\delta \in (0,1)$ and let Y be the output of a call to* CLUSTERING++$(P,w,f,\emptyset,k)$*; See Algorithm 1. Then $|Y| = k$, and with probability at least $1-\delta$,*

$$f(P,Y) \leq \frac{8\rho^2}{\delta^2}(1+\ln k) f^*(P,k).$$

Proof. Let $\delta' = \delta/2$ and let $\{P_1,\cdots,P_k\}$ be an optimal partition of P, i.e., $\sum_{i=1}^k f^*(P_i,1) = f^*(P,k)$. Let p_0 be a point that is sampled at random from P_k such that $p_0 = p$ with probability $\dfrac{w(p)}{\sum_{p'\in P_k} w(p')}$. Applying Lemma 5 with $u = t = k-1$ and $X = \{p_0\}$ yields,

$$E[f(P,Y)] \leq \big(f(P_k,\{p_0\}) + 4\rho^2 f^*(P\setminus P_k,k-1)\big)\cdot\sum_{i=1}^{k-1}\frac{1}{i} \tag{16}$$
$$\leq \big(f(P_k,\{p_0\}) + \frac{2\rho^2}{\delta'} f^*(P\setminus P_k,k-1)\big)\cdot\sum_{i=1}^{k-1}\frac{1}{i} \tag{17}$$
$$= \big(f(P_k,\{p_0\}) + \frac{2\rho^2}{\delta'} f^*(P,k) - \frac{2\rho^2}{\delta'} f^*(P_k,1)\big)\cdot\sum_{i=1}^{k-1}\frac{1}{i}, \tag{18}$$

where (18) holds by the definition of f^* and P_k. By plugging $Q = P_k$ in Corollary 6, with probability at least $1-\delta'$ over the randomness of p_0, we have

$$f(P_k,\{p_0\}) - \frac{2\rho}{\delta'} f^*(P_k,1) \leq 0, \tag{19}$$

and since $\rho \geq 1$, with probability at least $1-\delta'$ we also have

$$f(P_k,\{p_0\}) - \frac{2\rho^2}{\delta'} f^*(P_k,1) \leq 0. \tag{20}$$

Plugging (20) in (18) yields that with probability at least $1-\delta'$ over the randomness of p_0,

$$E[f(P,Y)] \leq \frac{2\rho^2}{\delta'} f^*(P,k)\cdot\sum_{i=1}^{k-1}\frac{1}{i} \leq \frac{2\rho^2}{\delta'} f^*(P,k)\cdot(1+\ln k). \tag{21}$$

Relating to the randomness of Y, by Markov's inequality we have

$$\Pr\{f(P,Y) < \frac{1}{\delta'}E[f(P,Y)]\} \geq 1-\delta'. \tag{22}$$

By (21) we have,

$$\Pr\left\{\frac{1}{\delta'}E[f(P,Y)] \leq \frac{2\rho^2}{\delta'^2}(1+\ln k)f^*(P,k)\right\} \geq 1-\delta'. \tag{23}$$

Using the union bound on (22) and (23) we obtain

$$\Pr\left\{f(P,Y) < \frac{2\rho^2}{\delta'^2}(1+\ln k)f^*(P,k)\right\} \geq 1-2\delta',$$

and thus

$$\Pr\left\{f(P,Y) < \frac{8\rho^2}{\delta^2}(1+\ln k)f^*(P,k)\right\} \geq 1-\delta.$$

□

4. Clustering with Outliers

In this section we prove that for every ρ-distance function f, any approximation algorithm for k-means clustering can be used to obtain the same approximation for clustering with outliers. We also provide application for classical k-medians and k-means (the distance functions will be defined below), exploiting its state-of-the-art approximation algorithm.

The following lemma is a framework that uses general clustering approximation and assumes a determination of m outliers out of a small subset of the data, in order to get a linear (in n) time approximation of clustering with m outliers.

Lemma 8. *Let $\delta \in [0,1)$ and let $\alpha, T_1, T_2, \rho > 0$. Let $m \geq 1$ be an integer such that $k+m < n$. Let f be a ρ-distance function over P. Suppose that $Y \subseteq P$ such that $|Y| = k+m$ and $f(P,Y) \leq \alpha f^*(P,k+m)$ with probability at least $1-\delta$, can be computed in T_1 time. In addition, Suppose that determining k-means of $k+m$ points can be computed in T_2 time. Let $M \in \operatorname{argmin}_{Q\subseteq P,|Q|=n-m} f^*(Q,k)$. Then, a set M_Y of $n-m$ points in P can be computed in $O(\max\{nk\tau, T_1\} + T_2)$ time such that $f(M_Y,k) \leq 3\rho^4\alpha \cdot f^*(M,k)$, with probability at least $1-\delta$.*

Proof. Let $Y^* \in \operatorname{argmin}_{Y''\subseteq Y,|Y''|=k} f(Y,Y'')$ denote the k-median of Y,

and let $M^* \in \operatorname{argmin}_{M''\subseteq M,|M''|=k} f(M,M'')$ denote the k-median of M. We have that, with probability at least $1-\delta$ over the randomness of Y,

$$f(P,Y) \leq \alpha f^*(P,k+m). \tag{24}$$

Assume that (24) holds, which happens with probability at least $1-\delta$, then we have

$$\begin{aligned} f(M,Y) &\leq f(P,Y) \qquad (25)\\ &\leq \alpha f^*(P,k+m)\\ &\leq \alpha(f^*(M,k) + f^*(P\setminus M,m))\\ &= \alpha f^*(M,k), \qquad (26)\end{aligned}$$

where (25) holds since $M \subset P$ and (26) holds since $f^*(P\setminus M,m) = 0$. For every $p' \in M$, let $y_{p'} \in \operatorname{argmin}_{y\in Y} f(p',y)$. Let $p \in M$. Substituting $x = y_p$, $q = p$ and $X = M^*$ in Lemma 2 yields

$$f(y_p,M^*) \leq \rho f(p,y_p) + \rho f(p,M^*) = \rho f(p,Y) + \rho f(p,M^*).$$

Multiplying by $w(p)$ and summing over every $p \in M$ yields

$$\begin{aligned}
\sum_{p\in M} w(p) f(y_p, M^*) &\leq \rho \sum_{p\in M} w(p) f(p, Y) + \rho \sum_{p\in M} w(p) f(p, M^*) \\
&= \rho f(M, Y) + \rho f(M, M^*) \\
&\leq \rho^3 \alpha f^*(M, k) + \rho f^*(M, k) && (27) \\
&= (\rho^3 \alpha + \rho) f^*(M, k), && (28)
\end{aligned}$$

where (27) holds by (26) and the definition of M^*. Since Y^* is the k-median of Y, we have

$$\begin{aligned}
\sum_{p\in M} w(p) f(y_p, Y^*) &\leq \sum_{p\in M} w(p) f(y_p, M^*) \\
&\leq (\rho^3 \alpha + \rho) f^*(M, k), && (29)
\end{aligned}$$

where (29) holds by (28). Substituting $x = p, q = y_p$ and $X = Y^*$ in Lemma 2 yields

$$f(p, Y^*) \leq \rho(f(p, y_p) + f(y_p, Y^*)) = \rho(f(p, Y) + f(y_p, Y^*)).$$

Multiplying by $w(p)$ and summing over every $p \in M$ yields

$$\begin{aligned}
f(M, Y^*) &= \sum_{p\in M} w(p) f(p, Y^*) \\
&\leq \rho \sum_{p\in M} w(p) \cdot (f(p, Y) + f(y_p, Y^*)) \\
&\leq \rho f(M, Y) + \rho \sum_{p\in M} w(p) f(y_p, Y^*) \\
&\leq \rho^3 \alpha f^*(M, k) + \rho \sum_{p\in M} w(p) f(y_p, Y^*) && (30) \\
&\leq \rho^3 \alpha f^*(M, k) + \rho(\rho^3 \alpha + \rho) f^*(M, k). && (31)
\end{aligned}$$

where (30) holds by (26) and (31) holds by (29). Let $M_Y \in \text{argmin}_{M' \subset P, |M'| = n-m} f(M', Y^*)$ denote the $n - m$ closest points of P from Y^*. We get that

$$f(M_Y, Y^*) \leq f(M, Y^*) \leq 3\rho^4 \alpha f^*(M, k).$$

The time complexity for computing Y is given by T_1. The time complexity of determining Y^* from Y is T_2. The time complexity of determining M_Y, i.e. finding the m points in P with the maximal distance from any point in Y^* is $O(nk\tau)$. The overall running time it takes to compute M_Y is then $O(\max\{nk\tau, T_1\} + T_2)$. □

4.1. *k-median with m outliers*

The following theorem is the state-of-the-art result for approximation of k-median with m outliers.

Theorem 9 (Theorem 3.1 in [46]). *Let $\{P, f\}$ be a metric space. Let $m \geq 1$ be an integer such that $k + m < n$. A set M of $n - m$ points can be computed in $O(\tau k^2 (k+m)^2 n^3 \log n)$ time such that $f(M, k) \in O(1) \cdot \min_{Q \subseteq P, |Q| = n-m} f^*(Q, k)$.*

The following corollary applies the framework of Lemma 8 with Theorem 7 and Theorem 9 in order to get a linear time approximation for k-median with m outliers.

Corollary 10. *Let $\{P, f\}$ be a metric space. Let $\delta \in (0,1]$, and let $m \geq 1$ be integers such that $k+m < n$. A set M of $n-m$ points can be computed in $O(\tau n(k+m) + \tau \cdot k^2(k+m)^5 \log(k+m))$ time such that, with probability at least $1-\delta$, $f(M,k) \in O(\log(k+m)) \cdot \min_{Q \subset P, |Q|=n-m} f^*(Q,k)$.*

Proof. Since $\{P, f\}$ is a metric space, f holds the triangle inequality over P, thus f is a ρ-distance function over P with $\rho = 1$. We apply Lemma 8 with $\alpha = \frac{8\rho^2}{\delta^2}(1+\ln(k+m))$ given by Theorem 7, and $T_1 = O(\tau n(k+m))$, running time of CLUSTERING++$(P, w, f, \emptyset, k+m)$ which results Theorem 7. In addition, from Theorem 9 one can determine m outliers out of $k+m$ points, which remains the set's k means, in time $T_2 = O(\tau \cdot k^2(k+m)^5 \log(k+m))$, which proves the corollary, since $3\alpha\rho^4$ is $O(\log(k+m))$. □

4.2. *k-means with m outliers*

The following theorem is $O(1)$-factor (state-of-the-art) result of k-means approximation.

Theorem 11 (Theorem 1 in [11]). *Let $d \geq 1$ be an integer and suppose that $P \subseteq \mathbb{R}^d$ and $f(p,q) = \|p-q\|^2$ for every $p, q \in P$. Then, there exists an algorithm whose running time is $O(dnk^2 \log\log k)$ that gets an input (P, w, k) and outputs Y of size $|Y| = k$ such that*

$$E[f(P,Y)] \leq 509 \cdot f^*(P,k).$$

The following corollary applies the framework of Lemma 8 with Theorem 11 and with a trivial determination of m outliers out of subset of $m+k$ points, in order to get a $O(1)$-factor approximation for k-means with m outliers.

Corollary 12. *Let $d \geq 1$ be an integer and suppose that $P \subseteq \mathbb{R}^d$ and $f(p,q) = \|p-q\|^2$ for every $p, q \in P$. Let $\delta \in (0,1]$, and let $m \geq 1$ be integers such that $k+m < n$. A set M of $n-m$ points can be computed in $O(dn(k+m)^2 \log\log(k+m) + d \cdot 2^{k+m})$ time such that, with probability at least $1-\delta$, $f(M,k) \in O(1) \cdot \min_{Q \subset P, |Q|=n-m} f^*(Q,k)$.*

Proof. By Markov's inequality, we have that

$$\Pr\{f(P,Y) < \frac{1}{\delta} E[f(P,Y)]\} \geq 1-\delta.$$

Since by Theorem 11, $E[f(P,Y)] \leq 509 \cdot f^*(P,k)$ we get that

$$\Pr\{f(P,Y) < \frac{1}{\delta} 509 \cdot f^*(P,k)\} \geq 1-\delta.$$

By Jensen's (or power-mean) inequality, for every convex function $g : \mathbb{R} \to \mathbb{R}$, and a real vector $v = (v_1, \cdots, v_u)$ we have $(1/u)\sum_{m=1}^{u} g(v_m) \geq g\left((1/u)\sum_{m=1}^{u} v_m\right)$. Specifically, for $g(z) := z^2$ and $v = \left(\|p-p'\|, \|p'-q\|\right)$ where $p, p', q \in P$, one may get

$$\begin{aligned} \|p-q\|^2 &\leq (\|p-p'\| + \|p'-q\|)^2 \qquad (32) \\ &\leq 2(\|p-p'\|^2 + \|p'-q\|^2), \end{aligned}$$

where (32) holds by the triangle inequality. Thus f is a ρ-distance function over P with $\rho = 2$. Appling Lemma 8 with $\alpha = \frac{1}{\delta}509$, $T_1 = O(dn(k+m)^2 \log\log(k+m))$ given by Theorem 11, and taking the trivial time $T_2 = d \cdot 2^{k+m}$ for determining k-means of $k+m$ points, proves the corollary, since $3\alpha\rho^4$ is $O(1)$. □

5. Conclusion

We proved that the k-Means++ algorithm can be generalized to clustering based on any ρ-distance function. We used this result to suggest a constant or near-constant factor approximations that is robust to m outliers and takes time linear in n.

Open problems include generalizations of k-Means++ for other shapes, such as k lines or k multi-dimensional subspaces, and using these approximations for developing coreset algorithms for these problems. Other directions include improving or generalizing the constant factor approximations for the original k-Means++ and its variants in this paper.

Author Contributions: Conceptualization, A.S. and D.F.; methodology, A.S. and D.F.; formal analysis, A.S. and D.F.; writing–original draft preparation, A.S and D.F; writing–review and editing, L.R and D.F; supervision, D.F.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. A local search approximation algorithm for k-means clustering. Proceedings of the eighteenth annual symposium on Computational geometry, 2002, pp. 10–18.
2. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; others. Knowledge Discovery and Data Mining: Towards a Unifying Framework. KDD, 1996, Vol. 96, pp. 82–88.
3. Gersho, A.; Gray, R.M. *Vector quantization and signal compression*; Vol. 159, Springer Science & Business Media, 2012.
4. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern classification and scene analysis*; Vol. 3, Wiley New York, 1973.
5. Aloise, D.; Deshpande, A.; Hansen, P.; Popat, P. NP-hardness of Euclidean sum-of-squares clustering. *Machine learning* **2009**, *75*, 245–248.
6. Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S.; Vinay, V. Clustering large graphs via the singular value decomposition. *Machine learning* **2004**, *56*, 9–33.
7. Arthur, D.; Vassilvitskii, S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
8. Ostrovsky, R.; Rabani, Y.; Schulman, L.J.; Swamy, C. The effectiveness of Lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)* **2013**, *59*, 1–22.
9. Aggarwal, A.; Deshpande, A.; Kannan, R. Adaptive sampling for k-means clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*; Springer, 2009; pp. 15–28.
10. Ailon, N.; Jaiswal, R.; Monteleoni, C. Streaming k-means approximation. Advances in neural information processing systems, 2009, pp. 10–18.
11. Lattanzi, S.; Sohler, C. A better k-means++ Algorithm via Local Search. International Conference on Machine Learning, 2019, pp. 3662–3671.
12. Makarychev, K.; Makarychev, Y.; Sviridenko, M.; Ward, J. A bi-criteria approximation algorithm for k Means. *arXiv preprint arXiv:1507.04227* **2015**.
13. Bartal, Y.; Charikar, M.; Raz, D. Approximating min-sum k-clustering in metric spaces. Proceedings of the thirty-third annual ACM symposium on Theory of computing, 2001, pp. 11–20.
14. Charikar, M.; Guha, S.; Tardos, É.; Shmoys, D.B. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences* **2002**, *65*, 129–149.
15. Chuzhoy, J.; Rabani, Y. Approximating k-median with non-uniform capacities. Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2005, pp. 952–958.
16. Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory* **1982**, *28*, 129–137.
17. Cox, D.R. Note on grouping. *Journal of the American Statistical Association* **1957**, *52*, 543–547.

18. Bailey, D.E.; Tryon, R.C. *Cluster analysis and the BC TRY system*; Tryon-Bailey Assoc. and the University of Colorado, 1970.
19. Milligan, G.W.; Sokol, L.M. A two-stage clustering algorithm with robust recovery characteristics. *Educational and psychological measurement* **1980**, *40*, 755–759.
20. Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory* **1960**, *6*, 7–12.
21. MacQueen, J.; others. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Oakland, CA, USA, 1967, Vol. 1, pp. 281–297.
22. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **1977**, *39*, 1–22.
23. Linde, Y.; Buzo, A.; Gray, R. An algorithm for vector quantizer design. *IEEE Transactions on communications* **1980**, *28*, 84–95.
24. Gray, R.M.; Neuhoff, D.L. Quantization. *IEEE transactions on information theory* **1998**, *44*, 2325–2383.
25. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: a review. *ACM computing surveys (CSUR)* **1999**, *31*, 264–323.
26. Alsabti, K.; Ranka, S.; Singh, V. An efficient k-means clustering algorithm **1997**.
27. Pelleg, D.; Moore, A. Accelerating exact k-means algorithms with geometric reasoning. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 277–281.
28. Phillips, S.J. Acceleration of k-means and related clustering problems. Proc. 4th ALENEX, 2002.
29. Forgey, E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classification. *Biometrics* **1965**, *21*, 768–769.
30. Rousseeuw, P.J.; Kaufman, L. Finding groups in data. *Hoboken: Wiley Online Library* **1990**, *1*.
31. Fisher, D. Iterative optimization and simplification of hierarchical clusterings. *Journal of artificial intelligence research* **1996**, *4*, 147–178.
32. Higgs, R.E.; Bemis, K.G.; Watson, I.A.; Wikel, J.H. Experimental designs for selecting molecules from large chemical databases. *Journal of chemical information and computer sciences* **1997**, *37*, 861–870.
33. Snarey, M.; Terrett, N.K.; Willett, P.; Wilton, D.J. Comparison of algorithms for dissimilarity-based compound selection. *Journal of Molecular Graphics and Modelling* **1997**, *15*, 372–385.
34. Bradley, P.S.; Fayyad, U.M. Refining initial points for k-means clustering. ICML. Citeseer, 1998, Vol. 98, pp. 91–99.
35. Meila, M.; Heckerman, D. An experimental comparison of several clustering and initialization methods. *arXiv preprint arXiv:1301.7401* **2013**.
36. Chawla, S.; Gionis, A. k-means–: A unified approach to clustering and outlier detection. Proceedings of the 2013 SIAM International Conference on Data Mining. SIAM, 2013, pp. 189–197.
37. Lin, J.H.; Vitter, J.S. e-Approximations with minimum packing constraint violation. Proceedings of the twenty-fourth annual ACM symposium on Theory of computing, 1992, pp. 771–782.
38. Hautamäki, V.; Cherednichenko, S.; Kärkkäinen, I.; Kinnunen, T.; Fränti, P. Improving k-means by outlier removal. Scandinavian Conference on Image Analysis. Springer, 2005, pp. 978–987.
39. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)* **2009**, *41*, 1–58.
40. Hawkins, D.M. *Identification of outliers*; Vol. 11, Springer, 1980.
41. Knox, E.M.; Ng, R.T. Algorithms for mining distancebased outliers in large datasets. Proceedings of the international conference on very large data bases. Citeseer, 1998, pp. 392–403.
42. Bay, S.D.; Schwabacher, M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 29–38.
43. Ramaswamy, S.; Rastogi, R.; Shim, K. Efficient algorithms for mining outliers from large data sets. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 427–438.
44. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 2000, pp. 93–104.

45. Papadimitriou, S.; Kitagawa, H.; Gibbons, P.B.; Faloutsos, C. Loci: Fast outlier detection using the local correlation integral. Proceedings 19th international conference on data engineering (Cat. No. 03CH37405). IEEE, 2003, pp. 315–326.
46. Chen, K. A constant factor approximation algorithm for k-median clustering with outliers. Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, 2008, pp. 826–835.
47. Charikar, M.; Khuller, S.; Mount, D.M.; Narasimhan, G. Algorithms for facility location problems with outliers. SODA, 2001, Vol. 1, pp. 642–651.
48. Jain, K.; Vazirani, V.V. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *Journal of the ACM (JACM)* **2001**, *48*, 274–296.
49. Jain, K.; Mahdian, M.; Markakis, E.; Saberi, A.; Vazirani, V.V. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. *Journal of the ACM (JACM)* **2003**, *50*, 795–824.
50. Mahdian, M. Facility location and the analysis of algorithms through factor-revealing programs. PhD thesis, Massachusetts Institute of Technology, 2004.
51. Arya, V.; Garg, N.; Khandekar, R.; Meyerson, A.; Munagala, K.; Pandit, V. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing* **2004**, *33*, 544–562.
52. Aboud, A.; Rabani, Y. Correlation clustering with penalties and approximating the reordering buffer management problem. PhD thesis, Computer Science Department, Technion, 2008.
53. Gupta, S.; Kumar, R.; Lu, K.; Moseley, B.; Vassilvitskii, S. Local search methods for k-means with outliers. *Proceedings of the VLDB Endowment* **2017**, *10*, 757–768.
54. Braverman, V.; Feldman, D.; Lang, H. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889* **2016**.