Clustering Oligarchies

Margareta Ackerman Caltech
Shai Ben-David University of Waterloo
David Loker University of Waterloo
Sivan Sabato Microsoft Research

Abstract

We investigate the extent to which clustering algorithms are robust to the addition of a small, potentially adversarial, set of points. Our analysis reveals radical differences in the robustness of popular clustering methods.

k-means and several related techniques are robust when data is clusterable, and we provide a quantitative analysis capturing the precise relationship between clusterability and robustness. In contrast, common linkage-based algorithms and several standard objective-function-based clustering methods can be highly sensitive to the addition of a small set of points even when the data is highly clusterable. We call such sets of points *oligarchies*.

Lastly, we show that the behavior with respect to oligarchies of the popular Lloyd's method changes radically with the initialization technique.

1 Introduction

Our investigation begins with the following question: can the output of an algorithm be radically altered by the addition of a small, possibly adversarial, set of points? We use the term *oligarchies* to describe such sets of "influential" points.

At first glance, it appears that all clustering methods are susceptible to oligarchies. Even k-means can substantially change its output upon the addition of a small set; if a data set has multiple structurally distinct solutions with near-optimal loss, then even a single point can radically alter the resulting partition. However, a more interesting picture emerges when considering how algorithms behave on well-clusterable data[1].

Examining their behavior on data that is well-clusterable, we find that some clustering methods exhibit a high degree of robustness to oligarchies; even small sets chosen in an adversarial manner have very limited influence on the output of these algorithms. These methods include k-means, k-medians, and k-medoids, as well as the popular Lloyd's method when initialized with random centers. We perform a quantitative analysis of these techniques, showing precisely how clusterability effects their robustness to small sets. Our results demonstrate that the more clusterable a data set, the greater its robustness to the influence of potential oligarchies.

In contrast, other well-known methods admit oligarchies even on data that is highly clusterable. We prove that common linkage-based algorithms, including the popular average-linkage, exhibit this behavior. Several well-known objective-function-based methods, as well as Lloyd's method initialized with pairwise distant centers, also fall within this category. More generally, we prove that all methods that detect clusterings satisfying a natural separability criteria, admit oligarchies even when the original data is well-clusterable.

Given the same well-clusterable input, algorithms that admit oligarchies can produce very different outputs from algorithms that prohibit them. For example, consider the data set displayed in Figure 1(a) and set the number of clusters, k, to 3. All algorithms that we considered, both those that admit and those that prohibit oligarchies, cluster this data as shown in Figure 1(a). As illustrated in Figure 1(b), when a small number of points is added, algorithms that prohibit oligarchies (eg. k-means) partition the original data in the same way as they did before the small set was introduced. In contrast, algorithms that admit oligarchies (eg. average-linkage) yield a radically different partition of the original data after the small set is

Preliminary work. Under review by AISTATS 2013. Do not distribute.

[1] Notice that the behavior of a clustering algorithm is often less important to the user when data is inherently un-clusterable.

added, as shown in Figure 1(c).

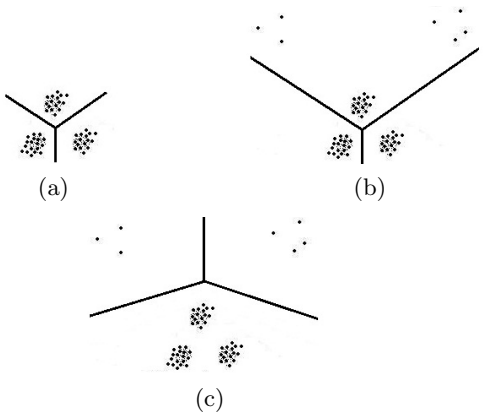

Figure 1: Contrasting the input-output behaviour of algorithms that prohibit oligarchies (b) with those that admit them (c).

Our work relates to a line of research by Ackerman, Ben-David, and colleagues [3, 16, 4, 2, 1] on a disciplined approach for selecting clustering algorithms. This approach involves identifying significant characteristics pertaining to the input-output behavior of different clustering paradigms. The characteristics should on the one hand distinguish between different clustering paradigms, and on the other hand be relevant to the domain knowledge that a user might have. Based on domain expertise, users could then choose which traits they want an algorithm to satisfy, and select an algorithm accordingly.

For some clustering applications, algorithms that prohibit oligarchies are preferred. This occurs, for example, when some of the data may be faulty. This may be the case in fields such as cognitive science and psychology, when analyzing subject-reported data. In such cases, an algorithm that is heavily influenced by a small number of elements is inappropriate since the resulting clustering may be an artifact of faulty data.

Algorithms that prohibit oligarchies may also be preferable when the data is entirely reliable, but clusters are expected to be roughly balanced (in terms of the number of points). Consider, for example, the use of clustering for identifying marketing target groups. Since target groups are typically large, no small set of individuals should have radical influence on how the data is partitioned.

However, there are applications that call for algorithms that admit oligarchies. Consider the task of positioning a predetermined number of fire stations within a new district. To ensure that the stations can quickly reach all households in the district, we may require that the maximum distance of any household to a station be minimized. It follows that a small number of houses can have significantly effect on the resulting clustering.

The paper is organized as follows. We begin with a summary of related previous work followed by an introduction of our formal framework. In Section 4, we present a summary of our main results, contrasting the manner in which different algorithms treat oligarchies. In Section 5 and Section 6 we provide a quantitative analysis of the extent to which some popular clustering methods are robust to potential oligarchies.

2 Previous Work

Work on the robustness of clustering methods to the addition of small set has so far focused on proposing new methods that are robust in this regard ([10], [11], [12], [13]). Previous measures of robustness to small sets did not lead to interesting differences among classical techniques ([13], [14]). On the other hand, we are able to obtain radical differences in the behaviour of classical algorithms. The main technical difference that makes this possible is that we bound the diameter of the data. Without bounding the diameter, all the standard clustering methods studied in this paper are sensitive to the addition of small sets: even a *single* outlier can radically modify the resulting clustering. That is, when outliers are placed sufficiently far away, clusters in the original data are forced to merge.

One notable algorithm designed to be robust to arbitrarily distant small sets is *trimmed k-means*, which discards a user-specified fraction of points that leads to an optimal k-means cost for the remaining data ([13], [14]). Similar to our work, [13] and [14] also rely on data clusterability to show robustness. In addition, just as in previous work, we also consider the setting where the number of clusters is fixed (see [13] for a detailed discussion on this condition). However, we are the first to obtain sharp distinctions between the behaviour of classical methods.

3 Notation and Definitions

We consider a space (E, d) where E is a set and d is a a distance function $d : E \times E \to \mathbb{R}^+$. It is assumed that d is symmetric and non-negative, and $d(x, x) = 0$ for all $x \in E$. The triangle inequality is assumed only if explicitly stated. Throughout this paper we consider only finite subsets of E.

The *diameter* of a set $X \subseteq E$ is $\max_{x,y \in X} d(x, y)$. We assume the diameter of E is at most 1. The *size* of a set X, denoted $|X|$, refers to the number of elements in X.

For a set $X \subseteq E$ and an integer $k \geq 1$, a k-clustering of X is a partition $C = \{C_1, \ldots, C_k\}$ of X into k disjoint sets, where $\cup_i C_i = X$. The diameter of a clustering C is the maximal diameter of a cluster in C.

For a clustering C of X and points $x, y \in X$, we write $x \sim_C y$ if x and y belong to the same cluster in C, and $x \nsim_C y$ otherwise.

The *Hamming distance* between clusterings C and C' of the same set X is defined by $\Delta(C_1, C_2) =$

$$|\{\{x, y\} \subset X \mid (x \sim_C y) \oplus (x \sim_{C'} y)\}| / \binom{|X|}{2},$$

where $\oplus$ denotes the logical XOR operation.

For sets $X, Z \subseteq E$ such that $X \subseteq Z$ and a clustering C of Z, $C|X$ denotes the restriction of C to X, thus if $C = \{C_1, \ldots, C_k\}$, then $C|X = \{C_1 \cap X, \ldots, C_k \cap X\}$.

A clustering algorithm $\mathcal{A}$ is a function that accepts a set $X \subseteq E$ and the space (E, d) and returns a clustering of X. $\mathcal{A}(X)$ denotes a clustering of X (since (E, d) will be clear from context, it is omitted from the notation of $\mathcal{A}$). Some algorithms accept the number of desired clusters as a parameter. In that case we denote the output clustering by $\mathcal{A}(X, k)$. k is sometimes omitted when it is clear from context.

In this paper we consider the robustness of sets to a small number of points. This is quantified by the following definition. Consider a data set X and a (typically large) subset Y, where the set $O = X \setminus Y$ is a potential oligarchy. The set Y is robust to the potential oligarchy O relative to an algorithm, if Y is clustered similarly with and without the points in O.

Definition 3.1 (δ-Robust)**.** *Given data sets X, Y and O, where $X = Y \cup O$, Y is* δ-robust *to O with respect to algorithm $\mathcal{A}$, if*

$$\Delta(\mathcal{A}(Y), \mathcal{A}(X)|Y) \leq \delta.$$

A small δ indicates a robust subset, meaning that the data within that subset determines how it is clustered (to a large extent). For example, if $\delta = 0$, then how the subset is clustered is entirely determined by the data within that subset. On the other hand, large values of δ represent a subset that is volatile to oligarchy O, where data outside of this subset have substantial influence on how data within this subset is partitioned. Note that δ ranges between 0 and 1.

For a randomized algorithm $\mathcal{A}$ we define probabilistic robustness as follows:

Definition 3.2 (Probabilistically δ-Robust)**.** *Let $\mathcal{A}$ be a randomized algorithm. Given data sets X, Y, and O where $X = Y \cup O$, Y is* δ-robust *to O with respect to algorithm $\mathcal{A}$ with probability $1 - \epsilon$, if with probability $1 - \epsilon$ over the randomization of $\mathcal{A}$,*

$$\Delta(\mathcal{A}(Y), \mathcal{A}(X)|Y) \leq \delta.$$

As our results will show, the robustness of a dataset is affected by whether it is well-clusterable, as captured in the following definition, based on a notion by Epter et al [8].

Definition 3.3 (α-Separable)**.** *A clustering C of X is* α-separable *for $\alpha \geq 0$ if for any $x_1, x_2, x_3, x_4 \in X$ such that $x_1 \sim_C x_2$ and $x_3 \nsim_C x_4$, $\alpha d(x_1, x_2) < d(x_3, x_4)$.*

If an algorithm contains an α-separable clustering for some large α (such as $\alpha \geq 1$), then it is well-clusterable. We define a balanced clustering based on the balance of cluster cardinalities.

Definition 3.4 (β-Balanced)**.** *A clustering $C = \{C_1, \ldots, C_k\}$ of X is* β-balanced *if $|C_i| \leq \beta |X|$ for all $1 \leq i \leq k$.*

Note that $\frac{1}{k} \leq \beta \leq 1$ and that $\beta = \frac{1}{k}$ for a perfectly balanced clustering.

4 Main Results

We demonstrate radical differences in the behaviour of clustering algorithms under the addition of a small number of elements. The k-means, k-medians and k-medoids objective functions are robust to the addition of small sets. Our first main result shows that the robustness of a set to potential oligarchies with respect to these objective functions is proportional to its size and degree of clusterability.

In the following theorem, we consider data set X, a typically large subset $Y \subset X$, and $O = X \setminus Y$ representing a potential oligarchy. The set Y is α-separable and β-balanced – this quantifies its degree of clusterability.

Theorem 4.1 bounds the robustness of Y in terms of its degree of clusterability and diameter, and the relationship between its size and the size of the potential oligarchy. The theorem shows that the larger and more clusterable a subset, the more robust it is to the influence of small sets.

Theorem 4.1. *Let $\mathcal{A}$ be one of k-means, k-medians or k-medoids. Let $p = 2$ if $\mathcal{A}$ is k-means and $p = 1$ otherwise. Consider data sets X, Y, and O where $X = Y \cup O$ and the set Y has an α-separable, β-balanced k-clustering of diameter s, for some $\alpha > 0$, $\beta \in [\frac{1}{k}, 1]$ and $s \in (0, 1]$. Then Y is δ-robust to O with respect to $\mathcal{A}$ for*

$$\delta \leq \tfrac{4p}{\alpha^p}(1 + \tfrac{|O|2^p}{|Y|s^p}) + 2k \cdot \beta^2.$$

Section 5.1 is devoted to proving this result.

To see the implications of this theorem, suppose $\beta = c/k$ where $c \geq 1$ is a small constant, so that the cluster sizes are fairly balanced in C. Fix s, d and α, and assume $\alpha \gg 4p$. In that case, if the size of the potential oligarchy is small, $|O| \ll |Y|$, then the robustness of Y is bounded by approximately $2c^2/k$.

It is important to note that Theorem 4.1 applies when some of the data in O is located within the convex hull of Y, which can be thought of as noise within Y. This effectively relaxes the clusterability condition on the region containing Y, allowing some data to lie between the well-separated clusters.

Note also that even if Y has a very small diameter, if it sufficiently large and clusterable, then it is robust to the influence of small sets.

In contrast to k-means and similar objective functions, we show that many clustering techniques do not have a property such as Theorem 4.1 in a strong sense. We show that algorithms that detect α-separable clusterings, for a large enough α, admit oligarchies.

Formally, we define this property of being α-*separability detecting* as follows.[2]

Definition 4.2 (α-Separability Detecting)**.** *An algorithm $\mathcal{A}$ is* α-separability-detecting *for $\alpha \geq 1$, if for all X and all $2 \leq k \leq |X|$, if there exists an α-separable k-clustering C of X, then $\mathcal{A}(X, k) = C$.*

In other words, whenever there is a clustering of the full data that consists of well-separated clusters, then this clustering is produced by the algorithm.

The above property is satisfied by many well-known clustering methods. In Section 6, we show that the linkage-based algorithms single-linkage, average-linkage, and complete-linkage, and the min-diameter objective functions, are all 1-separability detecting, and the k-center objective function is 2-separability-detecting.

The following Theorem demonstrates a sharp contrast between the behaviour of k-means (and similar objectives) as captured in Theorem 4.1 and algorithms that are α-separability detecting. It shows that for any desired level of clusterability, there exists a data set X with a subset $Y \subset X$ and $O = X \setminus Y$, such that Y is highly clusterable, the set O representing an oligarchy that contains as few as $k-1$ points, and yet Y is poorly robust to O with respect to these algorithms – thus Y is volatile to the influence of the oligarchy O.

Theorem 4.3. *Let $\mathcal{A}$ be an algorithm that is α-separability detecting for some $\alpha \geq 1$. Then for any $\beta \in [1/k, 1]$, $s \in [0, \frac{1}{(\alpha+1)^2})$ and any integer $m \geq k-1$, there exist data sets X, Y, and O where $X = Y \cup O$, the set O contains at most m elements, Y has an α-separable, β-balanced k-clustering with diameter s, and yet Y is not even $\beta(k-1)$-robust to O with respect to $\mathcal{A}$.*

Proof. Let Y be a set of points with diameter $s' < \frac{1}{\alpha+1}$ that contains all but $k-1$ elements of X, and let Y have an α-separable, β-balanced k-clustering with diameter $s < \frac{1}{(\alpha+1)^2}$. Let the data set O contain $k-1$ points at distance $\alpha s' + \epsilon$ from each other and from any point in Y. Then $\mathcal{A}(X, k)$ places all elements in Y within the same cluster because it is α-separability detecting, while $\mathcal{A}(Y, k)$ produces a β-balanced clustering of Y. □

Theorem 4.3 shows that even when Y is very large ($\frac{|Y|}{|X|}$ can be arbitrarily close to 1) and has an arbitrarily well-separable (α can be arbitrary large) and balanced ($\beta = \frac{1}{k}$) partition, the robustness score of Y to the oligarchy O can be bounded from below by $\beta(k-1)$, which approaches the worst possible score of robustness 1 as k grows.

This shows that α-separability detecting algorithms admit oligarchies of constant size (in particular, size $k-1$), even on data that is highly clusterable.

Lastly, we show that the behaviour of Lloyd's method changes radically with the method of initialization. The furthest-centroid initialization method deterministically selects a set of pairwise distant centers. We show that this algorithm is 1-separability detecting, implying that it admits oligarchies (see Section 6).

In contrast, in Section 5 below we show that Lloyd's method with random initialization behaves similarly to the k-means objective function, whereby well-clusterable sets are robust to the influence of a small number of elements.

5 Methods that Prohibit Oligarchies

In this section, we study clustering methods that are robust to the influence of a small number of elements when the data is well-clusterable. We distinguish between clustering objective functions and practical clustering algorithms, providing bounds for both popular objective functions, such as k-means, k-medians and k-medoids, and for Lloyd's method with random center initialization, a popular heuristic for finding clusterings with low k-means loss. For this section we assume that $(E, \|\cdot\|)$ is a normed space, with $d(x, y) = \|x - y\|$ for any $x, y \in E$.

[2] Note that for $\alpha \geq 1$, the α-separable k-clustering of any given data set is unique, if it exists.

5.1 k-means, k-medians and k-medoids objective functions

k-means and k-medians find the clustering $C = \{C_1, \ldots, C_k\}$ that minimizes the relevant cost denoted by $\text{COST}_p(C) = \sum_{i\in[k]} \min_{c_i \in E}\{\sum_{x\in C_i} \|x - c_i\|^p\}$, where the k-means cost is COST_2 and the k-medians cost is COST_1. The k-medoids cost relies on cluster centers selected from the input set, $\text{COST}_m(C) = \sum_{i\in[k]} \min_{c_i \in C_i}\{\sum_{x\in C_i} \|x - c_i\|\}$.

We work towards proving Theorem 4.1 by first showing that if the optimal clustering of a subset is relatively stable in terms of cost, then the subset is robust. Some stability assumption is necessary, since if there are two very different clusterings for the data set which have very similar costs, then even a single additional point might flip the balance between the two clusterings. We use the following notion of a cost-optimal clustering (which bears similarity to a notion by Balcan et al [7]).

Definition 5.1 ((δ, c)-cost-optimal)**.** *A clustering C of X is* (δ, c)-cost-optimal *with respect to a cost function* COST *if for all clusterings C' of X for which* $\text{COST}(C') \leq \text{COST}(C) + c$, $\Delta(C, C') \leq \delta$.

Lemma 5.2. *Let $\mathcal{A}$ be one of k-means, k-medians or k-medoids. Consider data sets X and $Y \subseteq X$. If there exists a $(\delta, 2^p|X \setminus Y|)$-cost-optimal clustering of Y relative to the cost associated with $\mathcal{A}$, then Y is 2δ-robust in X with respect to $\mathcal{A}$.*

Proof. Let $C = \{C_1, \ldots, C_k\}$ be the assumed cost-optimal clustering of Y. Let COST be the cost associated with $\mathcal{A}$. Let $p = 2$ if $\mathcal{A}$ is k-means and $p = 1$ otherwise. For $i \in [k]$, let $T_i = E$ if $\mathcal{A}$ is k-means or k-medians, and let $T_i = C_i$ if $\mathcal{A}$ is k-medoids.

Let $\bar{c}_i = argmin_{c_i \in T_i}\{\sum_{x\in C_i} \|x - c_i\|^p\}$. Then, the cost of the clustering $\mathcal{A}(X)$ is at most the cost of the clustering $C_1, \ldots, C_{k-1}, C_k \cup X \setminus Y$, since this is a possible clustering of X. Thus

$$\text{COST}(\mathcal{A}(X)) \leq \sum_{i\in[k]} \sum_{x\in C_i} \|x - \bar{c}_i\|^p + \sum_{z\in X\setminus Y} \|z - \bar{c}_k\|^p.$$

We now show that for all algorithms, for all $z \in X \setminus Y$, $\|z - \bar{c}_k\| \leq 2$. If $T_i = C_i$ then this is trivial, since X has diameter at most 1. If $T_i = E$, then let $\bar{x} = argmin_{x\in C_k} \|x - \bar{c}_k\|^p$. Clearly, $\|\bar{x} - \bar{c}_k\| \leq 1$, since otherwise $\sum_{x\in C_k} \|x - \bar{c}_k\|^p > |C_k|$, while $\sum_{x\in C_k} \|x - \bar{x}\|^p \leq |C_k| - 1$, contrary to the optimality of $\bar{c}_k$. It follows that for all $z \in X \setminus Y$,

$$\|z - \bar{c}_k\| \leq \|z - \bar{x}\| + \|\bar{x} - \bar{c}_k\| \leq 2.$$

Since $\text{COST}(\mathcal{A}(X)|Y) \leq \text{COST}(\mathcal{A}(X))$, it follows that

$$\begin{aligned}&\text{COST}(\mathcal{A}(X)|Y) \leq \\ &\sum_{i\in[k]} \sum_{x\in C_i} \|x - \bar{c}_i\|^p + 2^p|X \setminus Y| \\ &= \text{COST}(C) + 2^p|X \setminus Y|.\end{aligned}$$

Thus, by the cost-optimality property of C, $\Delta(\mathcal{A}(X)|Y, C) \leq \delta$. In addition, $\text{COST}(\mathcal{A}(Y)) \leq \text{COST}(C)$, thus $\Delta(\mathcal{A}(Y), C) \leq \delta$. It follows that $\Delta(\mathcal{A}(X)|Y, \mathcal{A}(Y)) \leq 2\delta$, thus the robustness of Y in X with respect to $\mathcal{A}$ is at most 2δ. □

The next lemma provides a useful connection between the Hamming distance of two clusterings, and the number of disjoint pairs that belong to the same cluster in one clustering, but to different clusters in the other.

Lemma 5.3. *Let C_1 and C_2 be two clusterings of Y, where C_1 is β-balanced and has k clusters. If $\Delta(C_1, C_2) \geq \delta$, then the number of disjoint pairs $\{x, y\} \subseteq Y$ such that $x \nsim_{C_1} y$ and $x \sim_{C_2} y$ is at least $\frac{1}{2}(\delta - k \cdot \beta^2)|Y|$.*

Proof. Let $A = \{\{x, y\} \mid x \nsim_{C_1} y, x \sim_{C_2} y\}$, and let $B = \{\{x, y\} \mid x \sim_{C_1} y, x \nsim_{C_2} y\}$. If $\Delta(C_1, C_2) \geq \delta$ then $|A \cup B| \geq \frac{1}{2}\delta|Y|(|Y| - 1)$. Since every cluster in C_1 is of size at most $\beta|Y|$,

$$|B| \leq |\{\{x, y\} \mid x \sim_{C_1} y\}| \leq \tfrac{1}{2}k \cdot \beta|Y|(\beta|Y| - 1).$$

It follows that

$$|A| \geq \tfrac{1}{2}\delta|Y|(|Y| - 1) - \tfrac{1}{2}k \cdot \beta|Y|(\beta|Y| - 1),$$

thus

$$|A| \geq \tfrac{1}{2}(\delta - k \cdot \beta^2)|Y|(|Y| - 1).$$

Now, for every x such that $\{x, y\} \in A$, there are at most $|Y| - 1$ pairs in A that include x. Thus the number of disjoint pairs in A is at least $|A|/(|Y| - 1)$. Therefore that are at least $\frac{1}{2}(\delta - k \cdot \beta^2)|Y|$ disjoint pairs in A. □

We now show that clusterings that are balanced and well-separable in a geometrical sense are also cost-optimal.

Lemma 5.4. *Suppose a k-clustering C of Y is α-separable, β-balanced and has diameter s. Let* COST *be one of* COST_1, COST_2 *or* COST_m. *Let $p = 2$ if* COST *is* COST_2 *and $p = 1$ otherwise. Then for any $\delta \in (0, 1)$, C is $(\delta, |Y|s^p(\frac{\alpha^p(\delta - k\cdot\beta^2)}{2p} - 1))$-cost-optimal with respect to* COST.

Proof. Let C' be a clustering of Y such that $\Delta(C, C') \geq \delta$. For $i \in [k]$, let $T_i = E$ if $\mathcal{A}$ is k-means or k-medians, and let $T_i = C_i$ if $\mathcal{A}$ is k-medoids. Let $c_i = argmin_{c_i \in T_i}\{\sum_{x \in C_i} \|x - c_i\|^p\}$, and $c'_i = argmin_{c'_i \in T_i}\{\sum_{x \in C'_i} \|x - c'_i\|^p\}$.

For every cluster C_i in C, and every $x \in C_i$, $\|x - c_i\|^p \leq s^p$. Thus $\text{COST}(C) \leq |Y|s^p$. On the other hand, for every pair $\{x, y\} \subseteq Y$, if $x \nsim_C y$ and $x \sim_{C'} y$, then for $p = \{1, 2\}$

$$\|x - c'_i\|^p + \|y - c'_i\|^p \geq \|x - y\|^p/p \geq (\alpha s)^p/p.$$

The first inequality is the triangle inequality for $p = 1$. For $p = 2$ the inequality can be derived by observing that the left hand side is minimized for $c'_i = (x + y)/2$. The last inequality follows from the properties of C and the fact that $x \nsim_C y$.

By Lemma 5.3, there are at least $|Y|\frac{1}{2}(\delta - k \cdot \beta^2)$ such $\{x, y\}$ pairs. Thus $\text{COST}(C') \geq |Y|\frac{1}{2p}(\alpha s)^p(\delta - k \cdot \beta^2)$. It follows that $\text{COST}(C') - \text{COST}(C) \geq |Y|(\frac{1}{2p}(\alpha s)^p(\delta - k \cdot \beta^2) - s^p)$. The lemma follows from the definition of cost-optimality. □

The proof of our first main result, Theorem 4.1, follows by letting $\delta' = \frac{2p}{\alpha^p}(1 + \frac{|O|2^p}{|Y|s^p}) + k \cdot \beta^2$. Then, by Lemma 5.4, C is $(\delta', 2^p|O|)$-cost-optimal. Thus by Lemma 5.2, the robustness of Y to O is at most $2\delta'$.

5.2 Lloyd's Method with Random Initial Centers

The results above pertain to algorithms that find the minimal-cost clustering. In practice, this task is often not tractable, and algorithms that search for a locally optimal clustering are used instead. For k-means, a popular algorithm is Lloyd's method. A common initialization for Lloyd's method is to select k random points from the input data set [9]. We call this algorithm Randomized Lloyd. It is also commonly referred to as "the k-means algorithm."

In order to find a solution with low k-means loss, it is common practice to run Randomized Lloyd multiple times and then select the minimal cost clustering. We show that clusterable data sets are immune to the influence of oligarchies when Randomizes Lloyd is repeated enough times. Specifically, we show that large clusterable subsets are robust with respect to this technique.

Formally, for a set $A \subseteq E$, define $\mu(A) = \sum_{x \in A} x/|A|$. Lloyd's method operates as follows:

1. Input: a dataset $Z \subseteq E$, an integer k, and an initial set of centers $\{p_1^0, \ldots, p_k^0\} \subseteq Z$.
2. $t \leftarrow 0$.
3. Repeat until $P^t = P^{t-1}$:
   (a) Let $P^t \leftarrow \{P_1^t, \ldots, P_k^t\}$ be the clustering of Z induced by:
   $x \in P_i^t \iff i = argmin_{i \in [k]} \|x - p_i^t\|$.
   (b) For all $i \in [k]$, $p_i^{t+1} \leftarrow \mu(P_i^t)$.
   (c) $t \leftarrow t + 1$.
4. Output: P^t.

We consider the following procedure: Run Randomized Lloyd n times with independent draws of initial centers, and output the final clustering with the least cost. We show that whenever there is a large subset that can be partitioned into a nice clustering $C = \{C_1, \ldots, C_k\}$, then with high probability over the randomization of the procedure, this subset is robust with respect to this procedure. The following two lemmas will allow us to show that if the initial centers are each in a different cluster of C, then the final clustering will be similar to C.

Lemma 5.5. *Suppose that the set $Y \subseteq E$ has an α-separable k-clustering $C = \{C_1, \ldots, C_k\}$ with diameter s, for $\alpha \geq 1$. Let P be the clustering of Y induced by $p_1, \ldots, p_k \in E$ as in step 3a. If there exists a permutation $\sigma : [k] \to [k]$ such that $\forall i \in [k], \|p_{\sigma(i)} - \mu(C_i)\| < \frac{(\alpha-1)s}{2}$, then $\forall i \in [k], P_{\sigma(i)}^t = C_i$.*

Proof. Without loss of generality, let σ be the identity permutation. For any $x \in C_i$, $\|x - p_i\| \leq \|x - \mu(C_i)\| + \|\mu(C_i) - p_i\| \leq s + \frac{(\alpha-1)s}{2} = \frac{(\alpha+1)s}{2}$. On the other hand, for every $j \neq i$, $\|x - p_j\| \geq \|x - \mu(C_j)\| - \|\mu(C_j) - p_j\| > \alpha s - \frac{(\alpha-1)s}{2} = \frac{(\alpha+1)s}{2}$. Therefore $\|x - p_i\| < \|x - p_j\|$, thus $x \in P_i$. □

Lemma 5.6. *Let $Y \subseteq X \subseteq E$. Let $P = \{P_1, \ldots, P_k\}$ be a k-clustering of X and let $C_i = P_i \cap Y$ for $i \in [k]$. Then $\forall i \in [k], \|\mu(P_i) - \mu(C_i)\| < \frac{|X \setminus Y|}{|C_i|}$.*

Proof. For any $i \in [k]$, let $Z_i = P_i \setminus C_i$. We have

$$\mu(P_i) = \frac{|C_i|}{|C_i| + |Z_i|}\mu(C_i) + \frac{|Z_i|}{|C_i| + |Z_i|}\mu(Z_i).$$

therefore $\mu(P_i) - \mu(C_i) =$

$$\left(\frac{|C_i|}{|C_i| + |Z_i|} - 1\right)\mu(C_i) + \frac{|Z_i|}{|C_i| + |Z_i|}\mu(Z_i)$$

$$= (\mu(Z_i) - \mu(C_i))\frac{|Z_i|}{|C_i| + |Z_i|}.$$

Since $\mu(C_i)$ and $\mu(Z_i)$ are both in the convex hull of E which has diameter at most 1, $\|\mu(Z_i) - \mu(C_i)\| \leq 1$.

In addition, $0 \leq |Z_i| \leq |X \setminus Y|$. Therefore

$$\|\mu(P_i)-\mu(C_i)\| = \|\mu(Z_i)-\mu(C_i)\|\frac{|Z_i|}{|C_i|+|Z_i|} \leq \frac{|X \setminus Y|}{|C_i|}.$$

□

We now show that if Randomized Lloyd is repeated enough times, then with high probability, at least one draw of initial centers has each of the centers in a distinct cluster of C.

Lemma 5.7. *Consider data sets X and $Y \subseteq X$, and a clustering C of Y, such that the smallest cluster in C is of size m. Then, for $\epsilon \in (0,1)$, if Randomized Lloyd's is run n times, where $n \geq \left(\frac{e|X|}{km}\right)^k \log(1/\epsilon)$, then with probability at least $1-\epsilon$, at least one draw of initial centers has each of the points in a distinct cluster of C.*

Proof. The probability that a single run of Randomized Lloyd has initial points in distinct clusters of C is $\theta = k! \prod_{i\in[k]} \frac{|C_i|}{|X|} \geq k! \left(\frac{m}{|X|}\right)^k \geq \left(\frac{km}{e|X|}\right)^k$, where the last inequality follows from Stirling's formula. If $n \geq \ln(1/\epsilon)/\theta$, then the probability that at least one draw has initial points in distinct clusters of C is at least $1-(1-\theta)^n \geq 1-\exp(-\theta n) \geq 1-\epsilon$. □

We can now prove that if Randomized Lloyd is run enough times with independent random initial centers then large clusterable sets are robust to oligarchies. In Section 6.2, we show that the type of initialization for Lloyd's method is a crucial factor in this outcome.

Theorem 5.8. *Consider data sets X, Y and O where $X = Y \cup X$ such that there exists an α-separable, β-balanced k-clustering C of Y with diameter $s > 0$, for some $\alpha \geq 3$. Let m be the size of the smallest cluster in C, and assume $m \geq \frac{2|O|}{(\alpha-1)s}$. Then with probability at least $1-\epsilon$, Y is δ-robust to O with respect to n runs of Randomized Lloyd, for*

$$n \geq \left(\frac{e|X|}{km}\right)^k \log(2/\epsilon),$$

and

$$\delta \leq \frac{8}{\alpha^2}\left(1+\frac{4|O|}{s^2|Y|}\right) + 2\beta^2 k.$$

Proof. First, we show that if Lloyd's method is executed with initial centers each belonging to a distinct cluster of C, and P is the output of this run of Lloyd's method, then $P|Y = C$. Assume without loss of generality that $\forall i \in [k], p_i^0 \in C_i$.

We prove by induction that for any $t \geq 0$, $P^t|Y = C$.
Induction basis: For all $i \in [k]$, $p_i^0 \in C_i$, thus $\|p_i^0 - \mu(C_i)\| \leq s \leq \frac{(\alpha-1)s}{2}$. By Lemma 5.5, $P^0|Y = C$.

Inductive step: Assume that $P^t|Y = C$. By Lemma 5.6, there is a numbering of the clusters in P^t such that $\|\mu(P_i^t) - \mu(C_i)\| \leq \frac{|O|}{|C_i|} \leq \frac{(\alpha-1)s}{2}$. By Lemma 5.5, $P^{t+1}|Y = C$.

Now, by Lemma 5.7, with probability $1-\epsilon/2$, Randomized Lloyd's with n repeats has at least one run with initial clusters belonging to distinct clusters of C for each of the inputs X and Y. Thus the probability of both runs to have a least one such run is at least $1-\epsilon$. For the input X, this run results in a clustering $\bar{P}$ such that $\bar{P}|Y = C$. Thus the clustering $P = \mathcal{A}(X)$ chosen out of all the runs satisfies $\textsc{cost}(P) \leq \textsc{cost}(\bar{P})$. It follows (similarly to the derivation in Lemma 5.2), that

$$\textsc{cost}(P|Y) \leq \textsc{cost}(\bar{P}|Y) + 4|O| = \textsc{cost}(C) + 4|O|.$$

Let $\delta = \frac{4}{\alpha^2}(1+\frac{4|O|}{s^2|Y|}) + k\beta^2$. By Lemma 5.4, C is $(\delta, |O|)$-cost-optimal. Thus $d(\mathcal{A}(X)|Y, C) \leq \delta$. For the run with input Y, the same lemma can be applied with $X = Y$, and it implies $d(\mathcal{A}(Y), C) \leq \delta$. Thus $\Delta(\mathcal{A}(Y), \mathcal{A}(X)|Y) \leq 2\delta$. □

6 Methods that Admit Oligarchies

We now turn to algorithms that admit oligarchies. In Section 4, we proved Theorem 4.3, showing that all algorithms that detect α-separable clusterings admit oligarchies even on data that is highly clusterable.

Theorem 4.3 demonstrates a sharp contrast between the behaviour of α-separability detecting algorithms and the behaviour captured in Theorem 4.1 for k-means and similar objective functions. We will now show that many well-known clustering methods are α-separability-detecting, resulting in the immediate conclusion that Theorem 4.3 holds for them.

6.1 Separability-detecting algorithms

In this section, we show that several common algorithms are α-seperability detecting. First, we consider linkage-based clustering, one of the most commonly-used clustering paradigms. Linkage-based algorithms use a greedy approach; at first every element is in its own cluster. Then the algorithm repeatedly merges the "closest" pair of clusters until some stopping criterion is met (see, for example, [3]).

To identify the closest clusters, these algorithms use a linkage function, which maps each pair of clusters to a real number representing their proximity.

Formally, a *linkage function* is a function

$$\ell : 2^E \times 2^E \to \mathbb{R}^+.$$

The following are the linkage-functions of some of the most popular linkage-based algorithms:

- Single linkage: $\ell(A,B) = \min_{a\in A, b\in B} d(a,b)$
- Complete linkage: $\ell(A,B) = \max_{a\in A, b\in B} d(a,b)$
- Average linkage:

$$\ell(A,B) = \sum_{a\in A, b\in B} d(a,b)/(|A|\cdot|B|).$$

For all of these linkage functions,

$$\forall A,B \subseteq E, \min_{a\in A,b\in B} d(a,b) \le \ell(A,B) \le \max_{a\in A,b\in B} d(a,b). \quad (1)$$

We consider linkage-based algorithms with the well-known k-stopping criterion, which terminates a linkage-based algorithm when the data is merged into k clusters, and returns the resulting clustering.

Theorem 6.1. *Let $\mathcal{A}$ be a clustering algorithms that uses a linkage-based function ℓ to merge clusters, and stops when there are k clusters. If Eq. 1 holds for ℓ, then $\mathcal{A}$ is 1-separability-detecting.*

Proof. By way of contradiction, assume that there exists a data set X with a 1-separable k-clustering C, but $\mathcal{A}(X,k) \neq C$. Consider the first iteration of the algorithm in which the clustering stops being a refinement of C. Let C' be the clustering before this iteration. There are clusters $C'_1, C'_2, C'_3 \in C'$ such that $C'_1, C'_2 \in C_i$ for some i, $C'_3 \in C_j$ for $j \neq i$, and the algorithm merges C'_1 and C'_3.

Thus $\ell(C'_1, C'_2) \ge \ell(C'_1, C'_3)$. By Eq. 1, $\ell(C'_1, C'_2) \le \max_{a\in C'_1, b\in C'_2} d(a,b)$, and $\min_{a\in C'_1, b\in C'_3} d(a,b) \le \ell(C'_1, C'_3)$. Since C is 1-separable, $\max_{a\in C'_1, b\in C'_2} d(a,b) < \min_{a\in C'_1, b\in C'_3} d(a,b)$, so $\ell(C'_1, C'_2) < \ell(C'_1, C'_3)$, contradicting the assumption. □

We show that there are also clustering objective functions that are α-separability-detecting. Thus clustering algorithms that minimize them satisfy Theorem 4.3.

The min-diameter objective function [6] is simply the diameter of the clustering. We show that it is 1-separability-detecting.

Theorem 6.2. *Min-diameter is 1-separability-detecting.*

Proof. For a set X, assume that there exists a 1-separable k-clustering C with diameter s. For any k-clustering $C' \neq C$ there are points x, y such that $x \sim_{C'} y$ while $x \nsim_C y$. $d(x,y) > s$, thus the diameter of C' is larger than s. Thus C' is not the optimal clustering for X. □

The k-center [5] objective functions finds a clustering that minimizes the maximum radius of any cluster in the clustering. In k-center the centers are arbitrary points in the underlying space, thus the cost of a k-clustering C is $\max_{i\in[k]} \min_{t\in E} \max_{x\in C_i} d(x,t)$. In *discrete k-center* they are a subset of the input points. We show that if d satisfies the triangle inequality then k-center and discrete k-center are 2-separability detecting.

Theorem 6.3. *If d satisfies the triangle inequality then k-center and discrete k-center are 2-separability detecting.*

Proof. Assume that there exists a 2-separable k-clustering C of a set X. Then the k-center cost is at most the diameter of C. For any k-clustering $C' \neq C$ there are points x, y such that $x \nsim_C y$ while $x \sim_{C'} y$. Hence the radius of C' is at least than $\frac{1}{2} \cdot \min_{x\not\sim_C y} d(x,y) > \max_{x\sim_C y} d(x,y)$, and thus it is larger than the cost of C. The proof for discrete k-center is similar. □

6.2 Lloyd's method with furthest centroids initialization

In Section 5.2, we have shown that large clusterable sets are robust with respect to Randomized Lloyd. This does not hold for the furthest-centroid initialization method [15], which admits oligarchies.

Using the furthest-centroid initialization method [15], the initial points $p_1^0, \ldots, p_k^0$ for an input set Z are chosen as follows: p_1^0 is the point with maximum norm (or an arbitrary point if no norm exists). Then, for all i between 2 and k, p_i^0 is set to be the point in Z that maximizes the distance from the other points that were already chosen. That is, $p_i^0 = argmax_{p\in Z} \min_{j\in[i-1]} d(p, p_j^0)$.

Lemma 6.4. *Lloyd's method with furthest centroid initialization is 1-separability detecting.*

Proof. If Z has a 1-separable k-clustering C, then between-cluster distances are larger than within-cluster distances. Thus, for every $i \ge 2$, the cluster of C that includes p_i^0 is different from the clusters that include $p_1^0, \ldots, p_{i-1}^0$. Thus the clustering induced by the initial points is C. In the next iteration, $p_i^1 = \mu(C_i)$ for all $i \in [k]$, thus the clustering remains C. □

References

[1] Ben-David S. Branzei S. Ackerman, M. and D. Loker. Weighted clustering. AAAI, 2012.

[2] M. Ackerman and S. Ben-David. Discerning linkage-based algorithms among hierarchical clustering methods. IJCAI, 2011.

[3] M. Ackerman, S. Ben-David, and D. Loker. Characterization of linkage-based clustering. COLT, 2010.

[4] M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. NIPS, 2010.

[5] P.K. Agarwal and C.M. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.

[6] A. Aggarwal, H. Imai, N. Katoh, and S. Suri. Finding k points with minimum diameter and related problems. *Journal of algorithms*, 12(1):38–56, 1991.

[7] M.F. Balcan and P. Gupta. Robust hierarchical clustering. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

[8] S. Epter, M. Krishnamoorthy, and M. Zaki. Clusterability detection and initial seed selection in large datasets. In *The International Conference on Knowledge Discovery in Databases*, volume 7, 1999.

[9] E.W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[10] M.T. Gallegos and G. Ritter. A robust method for cluster analysis. *The Annals of Statistics*, 33(1):347–380, 2005.

[11] L.A. Garcia-Escudero and A. Gordaliza. Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, pages 956–969, 1999.

[12] L.A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, pages 1324–1345, 2008.

[13] L.A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A review of robust clustering methods. *Advances in Data Analysis and Classification*, 4(2):89–109, 2010.

[14] C. Hennig. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, 99(6):1154–1176, 2008.

[15] I. Katsavounidis, C.C. Jay Kuo, and Z. Zhang. A new initialization technique for generalized lloyd iteration. *Signal Processing Letters, IEEE*, 1(10):144–146, 1994.

[16] R.B. Zadeh and S. Ben-David. A uniqueness theorem for clustering. UAI, 2009.