Perturbation Resilient Clustering for k-Center and Related Problems via LP Relaxations

Chandra Chekuri
Department of Computer Science, University of Illinois, Urbana-Champaign, IL 61801, USA
chekuri@illinois.edu

Shalmoli Gupta
Department of Computer Science, University of Illinois, Urbana-Champaign, IL 61801, USA
sgupta49@illinois.edu

Abstract

We consider clustering in the perturbation resilience model that has been studied since the work of Bilu and Linial [16] and Awasthi, Blum and Sheffet [8]. A clustering instance $\mathcal{I}$ is said to be α-perturbation resilient if the optimal solution does not change when the pairwise distances are modified by a factor of α and the perturbed distances satisfy the metric property – this is the metric perturbation resilience property introduced in [4] and a weaker requirement than prior models. We make two high-level contributions.

- We show that the natural LP relaxation of k-center and asymmetric k-center is integral for 2-perturbation resilient instances. We belive that demonstrating the goodness of standard LP relaxations complements existing results [12, 4] that are based on new algorithms designed for the perturbation model.
- We define a simple new model of perturbation resilience for clustering with *outliers*. Using this model we show that the unified MST and dynamic programming based algorithm proposed in [4] exactly solves the clustering with outliers problem for several common center based objectives (like k-center, k-means, k-median) when the instances is 2-perturbation resilient. We further show that a natural LP relxation is integral for 2-perturbation resilient instances of k-center with outliers.

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

Keywords and phrases Clustering, Perturbation Resilience, LP Integrality, Outliers, Beyond Worst Case Analysis

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2018.9

Related Version A full version of the paper is available at `https://arxiv.org/abs/1806.04202`.

Funding Work on this paper supported in part by NSF grants CCF-1319376 and CCF-1526799.

Acknowledgements CC thanks Mohit Singh for initial discussions on the integrality of the LP relaxation for 2-perturbation-resilient instances of k-median. We thank Yury Makarychev for comments on Voronoi clustering for k-center.

1 Introduction

Clustering is an ubiquitous task that finds applications in numerous areas since it is a basic primitive in data analysis. Consequently, clustering methods are extensively studied in many scientific communities and there is a vast literature on this topic. In a typical clustering problem the input is a set of points with a notion of similarity (also called *distance*) between

every pair of points, and a parameter k, which specifies the desired number of clusters. The goal is to partition the points into k clusters such that points assigned to the same cluster are similar. One way to obtain this partition is to select k centers and then assign each point to the nearest center. The quality of the clustering can be measured in terms of an objective function. Some of the popular and commonly studied ones are k-median (sum of distances of points to nearest center), k-means (sum of squared distances of points to nearest center), and k-center (maximum distance between a point to its nearest center). These are center-based objective functions. Unlike some applications in Operations Research, in many clustering problems in data analysis, the objective function is a proxy to identify the clusters and the actual value of the objective function is not necessarily meaningful. Clustering is often considered in the presence of outliers. In this setting the goal is to find the best clustering of the input after removing (at most) a specified number (or fraction) of points – this is useful in practice when the input data is noisy.

Most of the natural optimization problems that arise in clustering turn out to be NP-Hard. Extensive work exists on approximation algorithm design as well as heuristics. Although clustering and its variants are intractable in the worst case, various heuristic based algorithms like Lloyd's, K-Means++ perform very well in practice and are routinely used – at the same time some of these heuristics have poor worst-case approximation performance. On the other hand algorithms designed for worst-case approximation bounds may not work well in practice or may not be sufficiently fast for large data sets. To bridge this gap between theory and practice, there has been an increasing emphasis on *beyond worst case analysis*. Several models have been proposed to understand real-world instances and why they may be computationally easier. One such model is based on the notion of *instance stability*. This is based on the assumption that typical instances have a clear underlying optimal clustering (also known as *ground-truth clustering*) which is significantly better than all other clusterings, and remains the same under small perturbations.

The notion of stability/perturbation resilience was formalized in the work of Bilu and Linial [16] initially for Max Cut, and by Awasthi, Blum and Sheffet [8] for clustering. For clustering problems, an instance $\mathcal{I}$ is said to be α-perturbation resilient for some $\alpha > 1$ if the optimum clustering remains the same even if pairwise distances between points are altered by a multiplicative factor of at most α. Intuitively, α determines the degree of resilience of the instance, with a higher value translating to more structured, and separable instances. In the past few years, there has been increasing interest in understanding stable/perturbation-resilient instances. After several papers [8, 13, 12], a recent result by Angelidakis, Makarychev and Makarychev [4] showed that 2-perturbation resilient instances of several clustering problems with center based objectives (which includes k-median, k-center, k-means) can be solved exactly in polynomial time. For k-center finding the optimum solution for $(2-\delta)$-perturbation resilient instances is NP-Hard [12]. One criticism of perturbation resilience for clustering was the assumption in some earlier works that the optimum clustering remains stable under perturbation of the original metric d even when perturbed distance d' itself may not be a metric. Interestingly the results of [4] hold even under the weaker assumption of *metric* perturbation resilience, which constrains the perturbed pairwise distances to be a metric. The results in [4] are based on a simple and unified algorithm that computes the MST T of the given set of points and then applies dynamic programming on T to find the clusters; it is only in the second step that the specific objective function is used. We believe that empirically evaluating the performance of this algorithm, and related heuristics, on real-world data is an interesting avenue and plan to study it. Our work in this paper is motivated by the existing work and several interrelated questions on theoretical concerns, that we discuss next.

One of the objectives in beyond-worst-case analysis is to explain the empirical success of existing algorithms and mathematical programming formulations. For stable instances of Max-Cut and Minimum Multiway Cut, convex relaxations are known to be integral for various bounds on the perturbation parameter [31, 4]. In the context of k-median and k-means Awasthi et al. [9] showed that if the data is generated uniformly at random from k unit balls with well-separated centers, convex relaxations (linear and semi-definite) give an optimal itegral solution under appropriate separation conditions on the centers. However, for perturbation resilient clustering instances not much is known about the the natural LP relaxations. This raises a natural question.

▶ **Question 1.** *Are the natural LP relaxations for* 2*-metric perturbation resilient instances of clustering problems integral?*

There are several advantages in proving that well-known relaxations are integral. First, they provide evidence of the goodness of the relaxation; often these relaxations also have worst-case approximation bounds. Second, when the relaxation does not give an integral solution for a given instance we can deduce that the instance is *not* perturbation resilient.

As we remarked, one major takeaway from the paper of Angelidakis et al. [4], apart from its strong theoretical results, is the simple and unified algorithm that they propose which may lead to an effective heuristic. In real-world data there is often noise, and it would be useful to develop algorithms in the more general setting of clustering with outliers. This leads us to the question,

▶ **Question 2.** *Is there any stability model under which the algorithm proposed by [4] gives optimal solution for the problem of clustering with outliers?*

We remark that even for instances without outliers, removing a small fraction of the points can lead to a residual instance which has better stability parameters than the initial one. Thus, clustering with outlier removal is relevant even when there is no explicit noise.

1.1 Our Results

In this paper we address the preceding questions and obtain the following results.

- We show that a natural LP relaxation for k-center has an optimum integral solution for 2-metric-perturbation resilience instances[1]. Thus, when running the LP on a clustering instance, either we are guaranteed to have found the optimal solution (if the LP solution is integral), or we are guaranteed that the instance is not 2-perturbation resilient (if the LP solution is not integral). The previous algorithms of Angelidakis et al. [4], and Balcan et al. [12] do not have this guarantee, and could be arbitrarily bad if the instance is not 2-PR.
- Motivated by the work of [12] we consider the *asymmetric* k-center (Asym-k-center) problem. We show that a natural LP relaxation has an optimum integral solution for 2-*metric*-perturbation resilient instances[2]. For Asym-k-center the worst-case integrality gap of the LP relaxation is known to be $\Theta(\log^* k)$ [5, 23]. Previously [12] described a specific combinatorial algorithm that outputs an optimum solution for 2-perturbation resilient instances. We obtain it via the LP relaxation in the weaker metric perturbation model.

[1] For k-center it is known [12, 14] that *any* 2-approximation algorithm yields an optimum solution for 2-perturbation resilient instances. Although the LP lower bound provides a 2-approximation it is not immediate that it would be exact for perturbation-resilience instances

[2] In the asymmetric setting the perturbed distances should satisfy triangle inequality but symmetry is not required.

- We define a simple model of perturbation resilience for clustering with *outliers*. It is a clean extension of the existing perturbation resilience model. We show that under this new model, a modification of the algorithm of Angelidakis et al. [4] gives an exact solution for the outliers problem (for k-median, k-means, k-center and outer ℓ_p based objectives). This algorithm may lead to an interesting heuristic for clustering (noisy) real-world instances. We also show that for a 2-perturbation resilient instance of k-center with outliers, a natural LP relaxation has an optimum integral solution[3].

Our results show the efficacy of LP relaxations for k-center and its variants. We also demonstrate, via a natural model, that the interesting algorithm from [4] extends to handle outliers. Perturbation resilience appears to be a simple definition but it is hard to pin down its precise implications. Prior work demonstrates that observations and algorithms that appear simple in retrospect have not been easy to find. For k-center and Asym-k-center we work with notion of perturbation resilience under Voronoi clusterings as was done in [12]; this is the more restrictive version. See Section 2 for the formal definitions.

We would like to understand the integrality gap of the natural LP relaxations for perturbation resilient instances of k-median and k-means. We believe that the following open question is quite interesting to resolve.

▶ **Question 3.** *Is there a fixed constant α such that the natural LP relaxation for k-median (similarly k-means) has an integral optimum solution for every α-perturbation resilient instance[4]?*

Please see Appendix A for other related work. Most proofs are omitted in this conference version due to space constraints. We encourage the reader to read the full version of the paper that will be available shortly on the ArXiv.

2 Preliminaries

2.1 Definitions & Notations

Clustering. An instance $\mathcal{I}$ of a clustering problem is defined by the tuple (V, d, k), where V is a set of n points, $d : V \times V \to \mathbb{R}_{\geq 0}$ is a metric distance function, and k is an integer parameter. The goal is to find a set of k distinct points $S = \{c_1, \ldots, c_k\} \subseteq V$ called *centers* such that an objective function defined over the points is optimized. The objective function, also known as clustering cost, can be defined in various ways, and depends on the problem in hand. Here, we are interested in the k-median, and k-center objectives. Given a set of centers $S = \{c_1, \ldots, c_k\}$ these objectives are defined as follows: [i] (k-median) $\texttt{cost}_d(S) = \sum_{u \in V} d(S, u)$ [ii] (k-means) $\texttt{cost}_d(S) = \sum_{u \in V} d^2(S, u)$ [iii] (k-center) $\texttt{cost}_d(S) = \max_{u \in V} d(S, u)$; where $d(S, u) = \min_{i \in \{1,\ldots,k\}} d(c_i, u)$.

The Voronoi partition induced by the centers, gives a natural way of clustering the input point set. In fact, the inherent goal of clustering is to uncover the underlying partitioning of points, and one expects with correct choice of distance modeling, "k", and objective function,

[3] It will be interesting to see whether *any* 2-approximation algorithm for k-center-outlier gives an optimum solution for 2-perturbation resilient instances. A starting point would be to check the 2-approximation algorithm in [19].

[4] It may be possible to answer this question in the positive if we additionally assume that the optimum clusters are balanced in terms of number of points. However, we feel that such an assumption does not shed light on the structure of perturbation resilient instances that are not balanced.

the Voronoi partition induced by the optimal set of centers will reveal the underlying clustering. Throughout this paper, whenever we mention *optimal clustering*, we indicate the Voronoi partition corresponding to the optimal set of centers. Thus with this dual view of the clustering problem, given a set of centers $S = \{c_1, \ldots, c_k\}$, and corresponding Voronoi partition $\mathcal{C} = \{C_1, \ldots, C_k\}$, the clustering cost can be rewritten as: [i] (k-median) $\texttt{cost}_d(\mathcal{C}, S) = \sum_{i=1}^{k} \sum_{u \in C_i} d(c_i, u)$ [ii] (k-means) $\texttt{cost}_d(\mathcal{C}, S) = \sum_{i=1}^{k} \sum_{u \in C_i} d^2(c_i, u)$ [iii] (k-center) $\texttt{cost}_d(\mathcal{C}, S) = \max_{i \in \{1,\ldots,k\}} \max_{u \in C_i} d(c_i, u)$.

So far, in the clustering problem instance, we considered the distance function d to be a metric. However, this may not always be the case. Specifically, for the k-center objective, a generalization which is also studied is the Asymmetric k-center problem (Asym-k-center), where the distance function d in the input instance $\mathcal{I} = (V, d, k)$ is an asymmetric distance function. In other words, d obeys triangle inequality, but not symmetry. That is $d(u, v) \leq d(u, w) + d(w, v)$ for all $u, v, w \in V$. However $d(u, v)$ may be not be same as $d(v, u)$. The objective is the k-center objective, but because the distance is assymetric, order matters – we define the cost in terms of distance *from the center to the points* i.e. given a center c and a point u, $d(c, u)$ is used to define cost. To reiterate, given a set of centers $S = \{c_1, \ldots, c_k\}$ and corresponding Voronoi partition (w.r.t $d(c_i, u)$) $\mathcal{C} = \{C_1, \ldots, C_k\}$, the clustering cost is: (Asym-k-center) $\texttt{cost}_d(\mathcal{C}, S) = \max_{i \in \{1,\ldots,k\}} \max_{u \in C_i} d(c_i, u)$.

Clustering with Outliers. An instance $\mathcal{I}$ of a clustering with outliers problem is defined by the tuple (V, d, k, z), where V is a set of n points, $d : V \times V \to \mathbb{R}_{\geq 0}$ is a metric distance function, and k, z are integer parameters. The goal is to identify z points $Z \subseteq V$ as *outliers* and partition the remaining $V \setminus Z$ points into k clusters such that the clustering cost is minimized. Formally, given a set of outliers Z, a set of centers $S = \{c_1, \ldots, c_k\} \subseteq V \setminus Z$, and a Voronoi partition of $V \setminus Z$, $\mathcal{C} = \{C_1, \ldots, C_k\}$ induced by S, the clustering cost is defined as: [i] (k-median-outlier) $\texttt{cost}_d(\mathcal{C}, S; Z) = \sum_{i=1}^{k} \sum_{u \in C_i} d(c_i, u)$ [ii] (k-means-outlier) $\texttt{cost}_d(\mathcal{C}, S; Z) = \sum_{i=1}^{k} \sum_{u \in C_i} d^2(c_i, u)$ [iii] (k-center-outlier) $\texttt{cost}_d(\mathcal{C}, S; Z) = \max_{i \in \{1,\ldots,k\}} \max_{u \in C_i} d(c_i, u)$.

Perturbation Resilience. A clustering instance $\mathcal{I} = (V, d, k)$ is α-metric perturbation resilient (α-PR) for a given objective function, if for any metric [5] distance function $d' : V \times V \to \mathbb{R}_{\geq 0}$, such that for all $u, v \in V$, $\frac{d(u,v)}{\alpha} \leq d'(u, v) \leq d(u, v)$, the unique optimal clustering of $\mathcal{I}' = (V, d', k)$ is identical to the unique optimal clustering of $\mathcal{I}$.

Note that after perturbation the optimal centers may change, however for the instance to be perturbation resilient, the optimal clustering i.e. Voronoi partition induced by the optimal centers must stay the same. Unless otherwise noted, for the rest of the paper α-perturbation resilient indicates metric perturbation resilience.

Outlier Perturbation Resilience. A clustering with outliers instance $\mathcal{I} = (V, d, k, z)$ is α-metric outlier perturbation resilient (α-OPR) for a given objective function, if for any metric distance function $d' : V \times V \to \mathbb{R}_{\geq 0}$, such that for all $u, v \in V$, $\frac{d(u,v)}{\alpha} \leq d'(u, v) \leq d(u, v)$, the unique optimal clustering and outliers of $\mathcal{I}' = (V, d', k, z)$ are identical to the optimal solution of $\mathcal{I}$.

It is easy to see, if a clustering with outliers instance (V, d, k, z) with unique optimal clusters $\mathcal{C}$ and outliers Z is α-OPR, then the clustering instance $(V \setminus Z, d, k)$ is α-PR.

[5] In case of Asym-k-center, we consider perturbations in which d' obeys triangle inequality, but not symmetry

Notation. For integer k, let $[k] = \{1, \ldots, k\}$. Throughout, we use V to denote the input set of points, and n is the number of points. For any clustering instance (including outlier instances), $S = \{c_1, \ldots, c_k\}$ denotes an optimal set of centers, and $\mathcal{C} = \{C_1, \ldots, C_k\}$ denotes the corresponding Voronoi partition, which we call optimal clusters. Further, for a point $p \in C_i$, we often interchangebly use the terms, p is *assigned/belongs* to center c_i or cluster C_i. For a clustering with outlier instance, Z denotes the optimal set of outliers. In case of k-center, we refer to the optimal clustering cost as *optimal radius*, and denote it as R_d^*.

2.2 Some useful lemmas

Here we state some intuitive and useful lemmas regarding k-center and and Asym-k-center instances. The proofs of these lemmas are fairly simple and can be found in the full version.

Recall, in the definition of perturbation resilience, we insisted that the optimal k clustering of the perturbed instance $\mathcal{I}'$ has to be same as the optimal k clustering of the original instance. It is not hard to show, that for Asym-k-center (and also for k-center), if a $k-1$ clustering of $\mathcal{I}'$ exists whose cost is at most the optimal cost of k clustering, then the instance is not perturbation resilient. Formally,

▶ **Lemma 1.** *Consider any* Asym-k-center *instance $\mathcal{I} = (V, d, k)$. Let $S = \{c_1, \ldots, c_k\}$ be an optimal set of centers, and $\mathcal{C} = \{C_1, \ldots C_k\}$ be the corresponding optimal clustering. The optimal radius is R_d^*. Suppose there exists a set of $k-1$ centers $S' = \{c'_1, \ldots, c'_{k-1}\}$, inducing the Voronoi partition $\mathcal{C}' = \{C'_1, \ldots, C'_{k-1}\}$, with cost $\texttt{cost}_d(\mathcal{C}', S') \leq R_d^*$. Then, the optimal clustering $\mathcal{C}$ is not unique.*

One common technique we use in multiple arguments, is perturbing the input instance in a structured way. The next two lemmas are related to that.

▶ **Lemma 2.** *Consider a set of points V, and let d be an asymmetric distance function defined over V. Let G be a complete directed graph on vertices V. The edge lengths in graph G are given by the function ℓ, where for any edge (u, v), $\frac{d(u,v)}{2} \leq \ell(u, v) \leq d(u, v)$. Then the distance function d', defined as the shortest path distance in graph G using ℓ, is a metric*[6] *2-perturbation of d.*

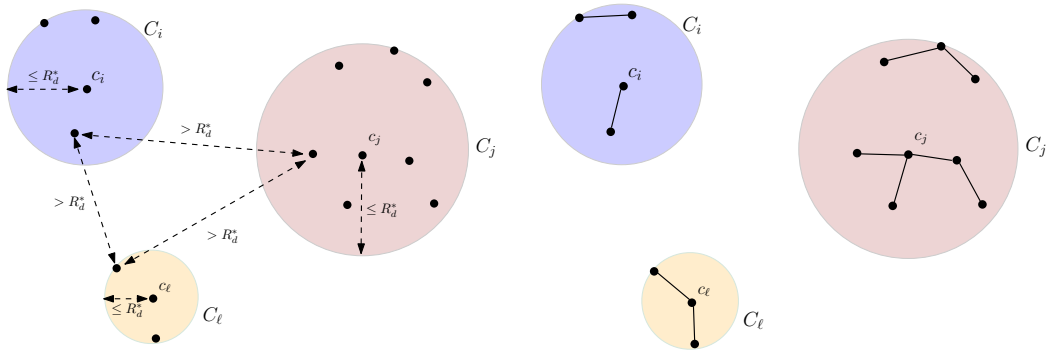
▶ **Lemma 3.** *Consider an* Asym-k-center *instance $\mathcal{I} = (V, d, k)$, and let $\mathcal{C}$ be the optimal clustering and R_d^* be the optimal radius. Let G be a complete directed graph over vertex set V. The edge lengths in graph G are given by the function ℓ, where (1) for a subset of edges E', $\ell(u, v) = \min\{d(u, v), R_d^*\}$; (2) for every other edge, $\ell(u, v) = d(u, v)$. Suppose d' is defined as the shortest path distance in graph G using ℓ. Consider the* Asym-k-center *instance $\mathcal{I}' = (V, d', k)$, let $R_{d'}^*$ be the optimal radius. If $\mathcal{C}$ is an optimal clustering in $\mathcal{I}'$, then $R_d^* = R_{d'}^*$.*

We can prove similar lemmas for the undirected version (corresponding to k-center) as well (formally stated in the full version).

3 k-center under Perturbation Resilience

In this section, we show that the natural LP relaxation for a 2-perturbation resilient k-center has an integral optimum solution.

[6] satsifies triangle inequality, and not necessarily symmetry



(a) Points in different optimal clusters are R_d^* apart. **(b)** Graph G_R for $R < R_d^*$.

Figure 1 Optimal Clusters in a 2-perturbation resilient k-center instance.

Properties of 2-perturbation resilient k-center instance: Angelidakis et al. [4] showed that in the optimal clustering of a 2-perturbation resilient k-center instance, every point is closer to its assigned center than to any point in a different cluster. In fact they show this property for general center based objectives, not just k-center. However for k-center we can show stronger structural properties : (1) any point is closer to a point in its own cluster, than to a point in a different cluster; (2) the distance between two points in two different clusters is atleast the optimal radius (see Figure 1a). We now formally state the properties (the proofs are in the full version).

▶ **Lemma 4.** *Consider a 2-perturbation resilient k-center instance $\mathcal{I} = (V, d, k)$. Let $S = \{c_1, \ldots, c_k\}$ be an optimal set of centers, and $\mathcal{C} = \{C_1, \ldots C_k\}$ be the corresponding unique optimal clustering. Consider any cluster C_i with $|C_i| \geq 2$, and let p, w be any two points in C_i. For any point q in a different cluster C_j $(i \neq j)$, we have $d(p, q) > d(p, w)$.*

▶ **Lemma 5.** *Consider a 2-perturbation resilient k-center instance $\mathcal{I} = (V, d, k)$. Let $S = \{c_1, \ldots, c_k\}$ be an optimal set of centers, and $\mathcal{C} = \{C_1, \ldots C_k\}$ be the corresponding unique optimal clustering. The optimal radius is R_d^*. Consider any point $p \in V$, and let $p \in C_i$. For any point q in a different cluster C_j $(i \neq j)$, we have $d(p, q) > R_d^*$.*

To prove the properties we use the result of Balcan et al. [12] – *any* 2-approximation algorithm for k-center finds the optimal clustering for a 2-perturbation resilient instance. They proved this result under the stronger definition of non-metric perturbation resilience, which was subsequently extended to metric perturbation resilience in an unpublished follow-up paper [14].

3.1 LP Integrality

Now, we show that as a consequence of Lemma 5, the LP relaxation for k-center is integral. Given an instance $\mathcal{I} = (V, d, k)$ of k-center and a parameter $R \geq 0$, we define the graph (also called *threshold* graph) $G_R = (V, E_R)$, where $E_R = \{(u, v) : u, v \in V, d(u, v) \leq R\}$. For a vertex v, let $\texttt{Nbr}[v] = \{u : (u, v) \in E_R\} \bigcup \{v\}$ be the neighbors (including itself). Observe, for any $R \geq R_d^*$, where R_d^* is the optimal solution cost of $\mathcal{I}$, there exists a set of k centers $S \subseteq V$, such that S *covers* V in G_R, i.e. $\bigcup_{c \in S} \texttt{Nbr}[c] = V$. Given a parameter R, we can define the following LP on graph G_R. We use y_v as an indicator variable for open centers, and x_{uv} to denote if v is assigned to u.

$$\sum_{u \in V} y_u \leq k \qquad \textbf{(kc-LP)}$$
$$x_{uv} \leq y_u \qquad \forall v \in V, u \in V$$
$$\sum_{u \in \texttt{Nbr}[v]} x_{uv} \geq 1 \qquad \forall v \in V$$
$$y_v, x_{uv} \geq 0$$

The minimum R for which **kc-LP** is feasible provides a lower bound on the optimum solution, and is the standard relaxation for k-center. It easy to see for all $R \geq R_d^*$ **kc-LP** is feasible. Further, it is well-known that the integrality gap is 2, that is, for all $R < R_d^*/2$, the LP is infeasible. However, if the k-center instance is 2-perturbation resilient, we can show that LP has no integrality gap.

▶ **Theorem 6.** *Consider a 2-perturbation resilient instance $\mathcal{I} = (V, d, k)$ of k-center. Let R_d^* be the cost of the optimal solution. Then, for any $R < R_d^*$, **kc-LP** is infeasible.*

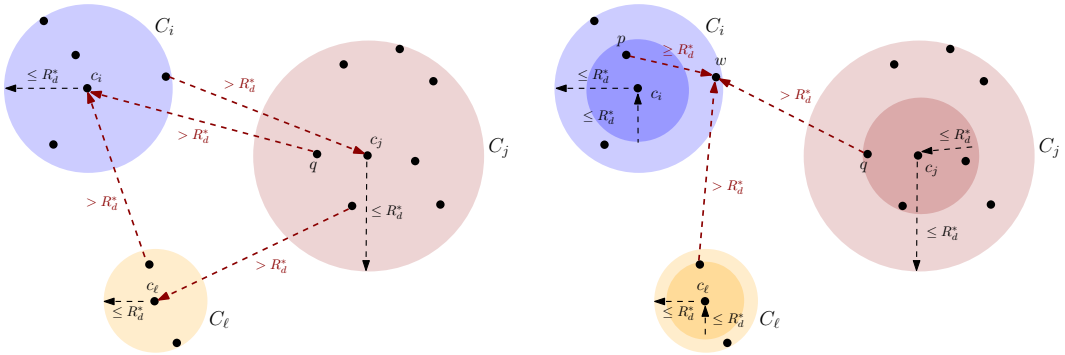
Proof. Let $C_1, \ldots C_k$ be the unique optimal clustering of instance $\mathcal{I} = (V, d, k)$, with optimal radius R_d^*. Consider an arbitrary $R < R_d^*$, and let G_R denote the corresponding threshold graph. Recall, in graph G_R the vertex set is V, and the edge set $E_R = \{(u, v) : d(u, v) \leq R\}$. According to Lemma 5, in a 2-PR instance, two points belonging to two different optimal clusters are separated by a distance of strictly more than R_d^*. Since $\mathcal{I}$ is 2-PR, graph G_R has a simple structure – for any $v \in C_i, i \in [k]$, $\texttt{Nbr}[v] \subseteq C_i$. Or in other words, the connected components of G_R are subsets of the optimal clusters (see Figure 1b).

Suppose, the k-center LP (**kc-LP**) defined over graph G_R is feasible, and (x, y) is the feasible fractional solution. Since every point is fully covered, and it can be covered only by its neighbors in G_R, we have, for all C_i, $\sum_{u \in C_i} y_u \geq 1$. Since, $\sum_{v \in V} y_v \leq k$, and the clusters $C_1, \ldots, C_k$ are disjoint, we have $\sum_{u \in C_i} y_u = 1$, for each i.

From the definition of R_d^*, there is an optimum cluster C_t such that $\min_{c \in C_t} \max_{v \in C_t} d(c, v) = R_d^*$. Let $C_t' = \{u \in C_t : y_u > 0\}$. As we argued earlier, $\sum_{u \in C_t} y_u = \sum_{u \in C_t'} y_u = 1$. Further, since for every $v \in C_t$, $\texttt{Nbr}[v] \subseteq C_t$, and v needs to be covered, we must have $C_t' \subseteq \texttt{Nbr}[v]$. Consider any $c \in C_t'$. Note that for every $v \in C_t$, c is a neighbor of v in graph G_R, i.e. $d(c, v) \leq R < R_d^*$. This implies, $\max_{v \in C_t} d(c, v) < R_d^*$ which is a contradiction. ◀

4 LP Integrality of Asym-k-center under Perturbation Resilience

We start with an LP relaxation for Asym-k-center problem by considering an unweighted directed graph on node set V. Specifically, for a parameter $R \geq 0$, we define the directed graph $G_R = (V, E_R)$, where $E_R = \{(u, v) : u, v \in V, d(u, v) \leq R\}$. For a node v, let $\texttt{Nbr}^-[v] = \{u : (u, v) \in E_R\} \bigcup \{v\}$ denote the in-neighbors, and $\texttt{Nbr}^+[v] = \{u : (v, u) \in E_R\} \bigcup \{v\}$ be the out-neighbors (including itself). Observe, for any $R \geq R_d^*$, there exists a set of k centers $S \subseteq V$, such that S *covers* V in G_R, i.e. $\bigcup_{c \in S} \texttt{Nbr}^+[c] = V$. Thus, given a parameter R, we can define the following LP relaxation on graph G_R. We use y_v as an indicator variable for open centers, and x_{uv} to denote if v is assigned to u.



(a) Cluster centers are atleast R_d^* away from points in different cluster.

(b) Cluster points R_d^* away from cluster core points, are R_d^* away from different cluster core points.

Figure 2 Properties of a 2-perturbation resilient Asym-k-center instance.

$$\sum_{u \in V} y_u \leq k \qquad \textbf{(asym-kc-LP)}$$

$$x_{uv} \leq y_u \qquad \forall v \in V, u \in V$$

$$\sum_{u \in \texttt{Nbr}^-[v]} x_{uv} \geq 1 \qquad \forall v \in V$$

$$y_v, x_{uv} \geq 0$$

For Asym-k-center, Archer [5] showed that the integrality gap is atmost $O(\log^* k)$, Infact it is tight within a constant factor [23]. The main result of the section is captured by the following theorem.

▶ **Theorem 7.** *Let $\mathcal{I} = (V, d, k)$ be a 2-perturbation resilient instance of* Asym-k-center *and let R_d^* be the cost of the optimal solution. Then, for any $R < R_d^*$, **asym-kc-LP** is infeasible.*

4.1 Properties of 2-perturbation resilient Asym-k-center instance

In Section 3 we showed that the clusters in an optimal solution to a 2-PR k-center instance have a strong separation property: $d(p, q) > R_d^*$ if p, q are in different clusters. For Asym-k-center the asymmetry in the distances does not permit a such a strong and simple separation property. However, we can show slightly weaker properties: (1) every optimal center is separated from any point in a different cluster by at least R_d^*; (2) points in a cluster which are far off from *core* points (these points have small distance "to" correponding cluster centers) in the cluster, are well-separated from core points of other clusters as well (See Figure 2a, Figure 2b). These properties suffice to prove our desired theorem. We now formally state the properties (the proofs are in the full version).

▶ **Lemma 8.** *Consider a 2-perturbation resilient* Asym-k-center *instance $\mathcal{I} = (V, d, k)$. Let $\mathcal{C} = \{C_1, \ldots C_k\}$ be the unique optimal clustering, induced by a set of centers $S = \{c_1, \ldots, c_k\}$. Let the optimal radius be R_d^*. Consider any center c_i. Then for any point q in a different cluster C_j $(i \neq j)$, we have $d(q, c_i) > R_d^*$.*

The next lemma formalizes the notion of core points and the property they enjoy.

▶ **Lemma 9.** *Consider a 2-perturbation resilient* Asym-k-center *instance $\mathcal{I} = (V, d, k)$. Let $\mathcal{C} = \{C_1, \ldots C_k\}$ be the unique optimal clustering induced by a set of centers $S = \{c_1, \ldots, c_k\}$. Let the optimal radius is R_d^*. Suppose $p \in C_i$ and $q \in C_j$ where $i \neq j$ and $d(p, c_i) \leq R_d^*$ and $d(q, c_j) \leq R_d^*$. Then for any $w \in C_i$ such that $d(p, w) \geq R_d^*$ we have $d(q, w) > R_d^*$.*

4.2 Proof of Theorem 7

Let $C_1, \ldots C_k$ be the unique optimal clustering of instance $\mathcal{I} = (V, d, k)$, with optimal radius R_d^*. Consider an arbitrary $R < R_d^*$, and let G_R denote the corresponding threshold graph. Recall, graph G_R is a directed graph defined over vertex set V, and the edge set $E_R = \{(u, v) : d(u, v) \leq R\}$. Suppose, the Asym-$k$-center LP (**asym-kc-LP**) defined over graph G_R is feasible, and (x, y) is a feasible fractional solution.

From Lemma 8, in a 2-PR instance, we have the following: if $q \notin C_i$ then $d(q, c_i) > R_d^* > R$. This implies that, in the graph G_R, for any $c_i, i \in [k]$, $\texttt{Nbr}^-[c_i] \subseteq C_i$. Let $C_i' = \{u \in \texttt{Nbr}^-[c_i] : y_u > 0\}$. Since (x, y) is a feasible solution, we must have $\sum_{u \in C_i'} y_u \geq 1$. Since, $\sum_{v \in V} y_v \leq k$, and the clusters $C_1, \ldots, C_k$ are disjoint, we have $\sum_{u \in C_i} y_u = \sum_{u \in C_i'} y_u = 1$, for all $i \in [k]$.

From the definition of R_d^* there must be a cluster C_t such that $\min_{c \in C_t} \max_{v \in C_t} d(c, v) = R_d^*$. Consider its center c_t and let $p \in C_t'$. Clearly $d(p, c_t) \leq R < R_d^*$. Furthermore, since C_t is the largest radius cluster, there exists $w \in C_t$, such that $d(p, w) \geq R_d^*$. Therefore in graph G_R, $p \notin \texttt{Nbr}^-[w]$. For any other cluster C_j, by Lemma 9, for any point $q \in C_j'$, we have $d(q, w) > R_d^*$. That is, $\texttt{Nbr}^-[w] \bigcap C_j' = \emptyset$, for any $j \neq t$. This implies w can be covered only by points that belong to C_t'. Therefore $\sum_{u \in \texttt{Nbr}^-[w]} x_{uw} \leq \sum_{u \in C_t' - p} y_u < 1$ since $y_p > 0$. This contradicts feasibility of (x, y).

5 LP Integrality of k-center-outlier under Perturbation Resilience

In this section we now consider the k-center-outlier problem. Recall that an instance $\mathcal{I} = (V, d, k, z)$ consists of a finite metric space (V, d) an integer k specifying the number of centers and an integer $z < |V|$ specifying the number of outliers that are allowed. One can write a natural LP relaxation for this problem as follows. As before, for a parameter $R \geq 0$, we define the graph $G_R = (V, E_R)$, where $E_R = \{(u, v) : u, v \in V, d(u, v) \leq R\}$. For a node v, let $\texttt{Nbr}[v] = \{u : (u, v) \in E_R\} \cup \{v\}$ be the neighbors (including itself). Observe, for any $R \geq R_d^*$, there exists a set of k centers $S \subseteq V$, and a set of outliers Z with $|Z| \leq z$, such that S *covers* $V \setminus Z$ in G_R, i.e. $\cup_{c \in S} \texttt{Nbr}[c] = V \setminus Z$. Thus, given a parameter R, we can define the following LP relaxation on graph G_R. We use y_v as an indicator variable for open centers, and x_{uv} to denote if v is assigned to u.

$$\sum_{u \in V} y_u \leq k \qquad \textbf{(kco-LP)}$$

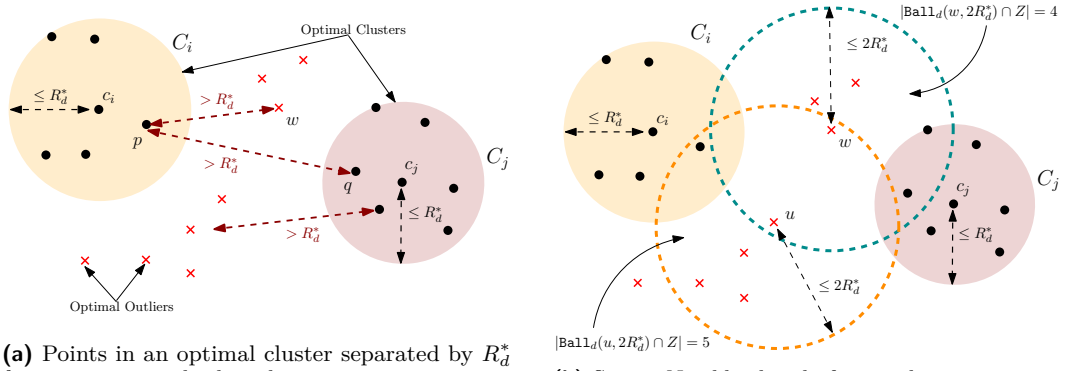
$$x_{uv} \leq y_u \qquad \forall v \in V, u \in V$$

$$\sum_{u \in V} x_{uv} \leq 1 \qquad \forall v \in V$$

$$\sum_{v \in V} \sum_{u \in V} x_{uv} \geq n - z$$

$$x_{uv} = 0 \qquad \forall v \in V, u \notin \texttt{Nbr}[v]$$

$$y_v, x_{uv} \geq 0$$



(a) Points in an optimal cluster separated by R_d^* from points outside that cluster.

(b) Sparse Neighborhood of an outlier.

Figure 3 Properties of a 2-perturbation resilient k-CENTER-OUTLIER instance.

The **kco-LP** is feasible for all $R \geq R_d^*$. The main theorem we prove in this section is as follows:

▶ **Theorem 10.** *Given a 2-perturbation resilient instance $\mathcal{I} = (V, d, k, z)$ of k-CENTER-OUTLIER with optimal cost R_d^*, **kco-LP** is infeasible for any $R < R_d^*$.*

5.1 Properties of 2-perturbation resilient k-center-outlier instance

For k-CENTER-OUTLIER we extend the properties from Section 3 that hold for 2-perturbation resilient instances. The first property shows that if p is a non-outlier point p and q is any point not in the same cluster as p (q could be an outlier) then $d(p, q) > R_d^*$. The second property is that for any outlier point q, the number of outliers in a ball of radius $2R_d^*$ is small. Specifically the number of points is strictly smaller than the size of the smallest cluster in the optimum clustering. This property makes intuitive sense, for otherwise q can define another cluster with outlier points and contradict the uniqueness of the clustering in after perturbation. We formally state them below after setting up the required notation.

Let $\mathcal{I} = (V, d, k, z)$ be a 2-outlier perturbation resilient k-CENTER-OUTLIER instance. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the optimum clustering, and Z be the set of outliers in the optimal solution of $\mathcal{I}$. Further, let $S = \{c_1, \ldots, c_k\}$ be the optimal centers inducing the clustering $\mathcal{C}$. Let the optimal cost be R_d^*. For each optimal cluster C_i, $n_i = |C_i|$ denotes its cardinality. Additionally, given a point $u \in Z$, and radius R, let $\texttt{Ball}_d(u, R) = \{v \in V : d(u, v) \leq R\}$ be the set of points in a ball of radius R centered at u.

The two main structural properties of an 2-OPR k-CENTER-OUTLIER instance we show are as follows (See Figure 3a, Figure 3b):

▶ **Lemma 11.** *Consider any non-outlier point $p \in V \setminus Z$, and let $p \in C_i$. For all $q \notin C_i$, $d(p, q) > R_d^*$.*

▶ **Lemma 12.** *For any outlier $p \in Z$, we have $|\texttt{Ball}_d(p, 2 \cdot R_d^*) \bigcap Z| < \min\{n_1, \ldots, n_k\}$.*

We observe that a much weaker version of the preceding lemma suffices for our proof of LP integrality. The weaker version states that $|\texttt{Ball}_d(p, R_d^*) \bigcap Z| < \min\{n_1, \ldots, n_k\}$. For if the statement is false, we could replace the smallest cluster with the cluster $\texttt{Ball}_d(p, R_d^*)$; this gives an alternate clustering with at most z outliers and the same optimum radius contradicting the uniqueness of the optimum solution.

5.2 Integrality Gap and Proof of Theorem 10

In this section, we show that **kco-LP** is infeasible for $R < R_d^*$. Recall in Lemma 11, we showed that the optimal clusters are well-separated from each other and also from the outliers. Therefore, in graph G_R, the connected components are either subsets of optimal clusters or outliers. As a consequence, in a fractional solution, non-outlier points can only be covered by points inside the cluster, and similarly outliers can be covered by outliers only. However, unlike k-center, here the tricky part is, the fractionally open outliers can potentially cover a lot of points. We show that this in fact is not possible because of the sparsity of an outlier's neighborhood.

Suppose the claim is not true, that is for some $R < R_d^*$, **kco-LP** has a feasible solution (x^*, y^*). Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be the set of clusters and Z be the outliers in the unique optimal solution of $\mathcal{I}$.

First, let us consider the simpler case when $y^*(Z) = 0$. Recall Lemma 11, for every $p \in C_i$, $(i \in [k])$, the distance to any $q \notin C_i$ is more than R_d^*. In other words, for any $p \in C_i$, $\texttt{Nbr}[p] \subseteq C_i$, and for any $w \in Z$, $\texttt{Nbr}[w] \bigcap V \setminus Z = \emptyset$. Therefore, $y^*(Z) = 0$ and the LP constraint $x_{uv} = 0, \forall v \in V, u \notin \texttt{Nbr}[v]$ implies (1) for any $w \in Z$, $x^*_{uw} = 0$ for all $u \in V$; (2) for any $v \in V \setminus Z$, and $w \in Z$, $x^*_{wv} = 0$. Therefore, (x^*, y^*) [restricted to $V \setminus Z$] is a feasible fractional solution for **kc-LP** defined for the k-center instance $\mathcal{I}' = (V \setminus Z, d, k)$, and parameter R. The optimal radius of $\mathcal{I}'$ is also R_d^*. Therefore by Theorem 6, we cannot have a feasible fractional solution for $R < R_d^*$, leading to a contradiction.

We focus on the case $y^*(Z) > 0$. Without loss of generality assume that the optimum clusters are numbered such that $n_1 \leq n_2 \leq \ldots \leq n_k$. For $i \in [k]$ let $a_i = y(C_i)$ and let $b = y^*(Z)$. For a point p let $\gamma_p = \sum_u x^*_{up}$ be the amount to which p is covered. For a set of points S we let $\gamma(S)$ denote $\sum_{p \in S} \gamma_p$.

▶ **Claim 5.1.** *Total coverage of outlier points, that is,* $\gamma(Z) = \sum_{p \in Z} \rho_p < bn_1$.

▶ **Claim 5.2.** *Let* C_i *be an optimum cluster such that* $a_i < 1$. *Then* $\gamma(C_i) \leq n_i a_i$.

Let $A = \{i \in [k] \mid a_i < 1\}$ be the indices of the clusters whose total y value is strictly less than 1. Since $y(V) = k$, we have $b \leq \sum_{i \in A}(1 - a_i)$.

Using the preceding two claims we can show that $\gamma(V) < n - z$ which contradicts the feasibility of (x^*, y^*).

6 Algorithm for k-median-outlier under Perturbation Resilience

In this section, we present a dynamic programming based algorithm for k-MEDIAN-OUTLIER, which gives an optimal solution when the instance is 2-perturbation resilient. First, we prove some structural properties of a 2-OPR k-MEDIAN-OUTLIER instance. They serve as the key ingredient in showing that our algorithm will return exact solution for 2-OPR instances.

This section is essentially a straight forward extension of the ideas in [4] once the model is set up. In a sense the model justifies the natural extension of the algorithm from [4] to the outlier setting.

6.1 Properties of 2-perturbation resilient k-median-outlier instance

Angelidakis et al. [4] proved that in the optimal clustering of a 2-perturbation resilient k-median instance, every point is closer to its assigned center than to any point in a different cluster. In the optimal solution of k-MEDIAN-OUTLIER, points are not only assigned to clusters, some points are identified as outliers as well. Here, we extend the result of [4]

to show that the optimal solution of a 2-OPR k-MEDIAN-OUTLIER instance satisfies the property: any non-outlier point is closer to its assigned center than to any point outside the cluster.

▶ **Lemma 13.** *Consider a 2-perturbation resilient k-MEDIAN-OUTLIER instance $\mathcal{I} = (V, d, k, z)$. Let $\mathcal{C} = \{C_1, \ldots C_k\}$, and Z be the unique optimal clustering and outliers resp. Consider any point $p \in V \setminus Z$, and let $p \in C_i$. For all $q \notin C_i$, we have $d(c_i, p) < d(p, q)$.*

6.2 Algorithm

In the previous section, we showed that in the optimal solution of a 2-perturbation resilient k-MEDIAN-OUTLIER instance, any non-outlier point is closer to its assigned center than to any point outside the cluster. This gives a nice structure to the optimal solution. In particular, the optimal clusters form subtrees in the minimum spanning tree over input point set. We leverage this property to design a dynamic programming based algorithm to identify the optimal clusters and outliers.

▶ **Lemma 14.** *Let $\mathcal{I} = (V, d, k, z)$ be a 2-perturbation resilient instance of the k-MEDIAN-OUTLIER problem. Let T be a minimum spanning tree on V. The optimal clusters of $\mathcal{I}$, $C_1, \ldots C_k$ are subtrees in T i.e. for any two points $p, q \in C_i$, all the points along the unique tree path between p, and q belongs to cluster C_i.*

Once we prove Lemma 14, a simple modification of the dynamic programming in [4] gives exact solution for k-MEDIAN-OUTLIER. The running time of the algorithm is $O\big((nkz)^2\big)$. Further, as noted in [4], the algorithm readily generalizes to give exact solution for other objectives like k-CENTER-OUTLIER, k-MEANS-OUTLIER, and more general ℓ_p objectives. The details can be found in the full version of the paper.

References

1. Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013.
2. Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k-means and euclidean k-median by primal-dual algorithms. In *FOCS*, pages 61–72, 2017.
3. Sara Ahmadian and Chaitanya Swamy. Approximation Algorithms for Clustering Problems with Lower Bounds and Outliers. In *ICALP*, pages 69:1–69:15, 2016.
4. Haris Angelidakis, Konstantin Makarychev, and Yury Makarychev. Algorithms for stable and perturbation-resilient problems. In *STOC*, pages 438–451, 2017.
5. Aaron Archer. Two $O(\log^* K)$-approximation algorithms for the asymmetric k-center problem. In *IPCO*, pages 1–14, 2001.
6. David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *SODA*, pages 1027–1035, 2007.
7. Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *FOCS*, pages 309–318, 2010.
8. Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Inf. Process. Lett.*, 112(1-2):49–54, 2012. doi:10.1016/j.ipl.2011.10.006.
9. Pranjal Awasthi, Moses Charikar, Ravishankar Krishnaswamy, and Ali Kemal Sinop. The hardness of approximation of euclidean k-means. *CoRR*, abs/1502.03316, 2015. arXiv:1502.03316.
10. Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *APPROX-RANDOM*, pages 37–49, 2012.

11 Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *SODA*, pages 1068–1077, 2009.

12 Maria-Florina Balcan, Nika Haghtalab, and Colin White. k-center clustering under perturbation resilience. In *ICALP*, pages 68:1–68:14, 2016.

13 Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. In *ICALP*, pages 63–74, 2012.

14 Maria-Florina Balcan and Colin White. Clustering under local stability: Bridging the gap between worst-case and beyond worst-case analysis. *CoRR*, abs/1705.07157, 2017. arXiv:1705.07157.

15 Yonatan Bilu, Amit Daniely, Nati Linial, and Michael E. Saks. On the practically interesting instances of MAXCUT. In *STACS*, pages 526–537, 2013.

16 Yonatan Bilu and Nathan Linial. Are stable instances easy? In *ICS*, pages 332–341, 2010.

17 Johannes Blömer, Christiane Lammersen, Melanie Schmidt, and Christian Sohler. Theoretical analysis of the k-means algorithm - A survey. *CoRR*, abs/1602.08254, 2016. arXiv:1602.08254.

18 Jaroslaw Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median and positive correlation in budgeted optimization. *ACM Trans. Algorithms*, 13(2):23:1–23:31, 2017. doi:10.1145/2981561.

19 Deeparnab Chakrabarty, Prachi Goyal, and Ravishankar Krishnaswamy. The non-uniform k-center problem. In *ICALP*, pages 67:1–67:15, 2016.

20 Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. A constant-factor approximation algorithm for the k-median problem. In *STOC*, pages 1–10, 1999.

21 Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *SODA*, pages 642–651, 2001.

22 Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In *SODA*, pages 826–835, 2008.

23 Julia Chuzhoy, Sudipto Guha, Eran Halperin, Sanjeev Khanna, Guy Kortsarz, Robert Krauthgamer, and Joseph (Seffi) Naor. Asymmetric k-center is $\log^* n$-hard to approximate. *J. ACM*, 52(4):538–551, 2005. doi:10.1145/1082036.1082038.

24 Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *FOCS*, pages 49–60, 2017.

25 T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.

26 D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10:180–184, 1985.

27 K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *J. ACM*, 50(6):795–824, 2003. doi:10.1145/950620.950621.

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

29 Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *FOCS*, pages 299–308, 2010.

30 S. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129–137, 2006. doi:10.1109/TIT.1982.1056489.

31 Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Bilu-linial stable instances of max cut and minimum multiway cut. In *SODA*, pages 890–906. SIAM, 2014.

32 Matús Mihalák, Marcel Schöngens, Rastislav Srámek, and Peter Widmayer. On the complexity of the metric TSP under stability considerations. In *SOFSEM*, pages 382–393, 2011.

33 Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *FOCS*, pages 165–176, 2006.

34 Rina Panigrahy and Sundar Vishwanathan. An $O(\log^* n)$ approximation algorithm for the asymmetric p-center problem. *J. Algorithms*, 27(2):259–268, 1998. doi:10.1006/jagm.1997.0921.

35 Aravindan Vijayaraghavan, Abhratanu Dutta, and Alex Wang. Clustering stable instances of euclidean k-means. In *NIPS*, pages 6503–6512, 2017.

A Related Work

There is extensive related work on clustering topics. Here we only mention some closely related work.

Clustering. For both k-center and asymmetric k-center tight approximation bounds are known. For k-center, already in the mid 1980's Gonzales [25] and Hochbaum & Shmoys [26] had developed remarkably simple 2-approximation algorithms, which are infact tight. Approximating asymmetric k-center is significantly harder. Panigrahy and Vishwanathan [34] designed an elegant $O(\log^* n)$ approximation algorithm, which was subsequently improved by Archer [5] to $O(\log^* k)$. Interestingly, the result is asymptotically tight [23].

For k-means and k-median– arguably the two most popular clustering problems – there is a long line of research (see [17] for a survey on k-means). The first constant factor approximation for the k-median problem was given by Charikar et al. [20], and the current best-known is a 2.675 approximation by Byrka et al. [18]; and it is NP-Hard to do better than $1 + 2/e \approx 1.736$ [27]. For k-means the best approximation known is 6.357 [2]. The k-means problem is widely used in practice as well, and the commonly used algorithm is Lloyd's algorithm, which is a special case of the EM algorithm [30]. While there is no explicit approximation guarantee of the algorithm, it performs remarkably well in practice with careful seeding [6] (this heuristic is called K-Means++).

Clustering with Outliers. The influential paper by Charikar et al. [21] initiated the work on clustering with outliers and other robust clustering problems. For k-center with outliers, they gave a greedy 3-approximation algorithm. Recently, it has been improved to 2-approximation [19]. A related problem of lower bounded k-center clustering with outliers has also been studied [3].

For k-median with outliers they gave a bicriteria approximation algorithm, which achieves an approximation ratio of $4(1 + \epsilon)$, violating the number of outliers by a factor of $(1 + \epsilon)$. The first constant factor approximation algorithm for this problem was given by Chen (the constant is not explicitly computed) [22]. Very recently, Krishnaswamy et al. [28] proposed a generic framework for clustering with outliers. It improves the results of Chen and gives the first constant factor approximation for k-means with outliers. However, the algorithm does not appear suitable for practice in its current form (See [1] for details on algorithms used in practice for clustering with outliers).

Perturbation Resilience. The notion of perturbation resilience was introduced by Bilu and Linial [16]. They originally considered it for the Max Cut problem, designing an exact

polynomial time algorithm for $O(n)$-stable instances [7] of Max Cut. It was later improved to $O(\sqrt{n})$-stable instances [15], and finally Makarychev et al. gave a polynomial time exact algorithm for $O(\sqrt{\log n} \cdot \log \log n)$-stable instances [31].

The definition of perturbation resilience naturally extends to clustering problems. Awasthi, Blum, and Sheffet [8] presented an exact algorithm for solving 3-perturbation resilient clustering problems with *separable center based objectives* (s.c.b.o) – this includes k-median, k-means, k-center. This result was later improved by Balcan and Liang [13], who gave an exact algorithm for clustering with s.c.b.o under $(1+\sqrt{2})$-perturbation resilience. Specifically for k-center and asymmetric k-center, Balcan, Haghtalab, and White [12] obtained a stronger result – any 2-approximation algorithm for k$-$*center* gives optimum solution for 2-perturbation resilient instances. Their result was later extended to metric perturbation resilience in a follow-up paper [14]. In fact, for k-center they gave a stronger result, that any 2-approximation algorithm for k-center can give an optimal solution for 2-perturbation resilient instances. They also showed the results are essentially tight unless NP = RP[8]. Recently, Angelidakis et al. [4], gave an unifying algorithm which gives exact solution for 2-perturbation resilient instances of clustering problems with center based objectives. In fact, their algorithms work under metric perturbation resilience, which is a weaker assumption. Perturbation resilience has also been studied in various other contexts, like TSP, Minimum Multiway Cut, Clustering with min-sum objectives [13, 31, 32].

Robust Perturbtion Resilience. Perturbation resilience requires optimal solution to remain unchanged under any valid perturbation. Balcan and Liang [13] relaxed this condition slightly, and defined (α, ϵ)-perturbation resilience (or robust perturbation resilience), in which at most ϵ fraction of the points can change their cluster membership under any α-perturbation. They gave a near optimal solution for k-median under $(2+\sqrt{3}, \epsilon)$-perturbation resilience, when the clusters are not too small. Further, for k-center and asymmetric k-center efficient algorithms are known for $(3, \epsilon)$-perturbation resilient instances, assuming mild size lower bound on optimal clusters [12].

Other Stability Notions. Several other stability models, and separation conditions have also been studied to better explain real-world instances. In a seminal paper Ostrovsky, Rabani, Schulman, and Swamy [33] considered k-means instances where the cost of clusterng using k is clusters is much lower than $k-1$ clusters. They showed, that popular K-Means++ algorithm achieves an $O(1)$-approximation for these instances. Subsequently there has been series of work many other models like approximation stability [11], distribution stability [7, 24], spectral separability [29, 10, 24], and more recently on additive perturbation stability [35].

[7] They used this name to denote perturbation resilient instances of Max Cut

[8] They showed, unless NP = RP, no polynomial-time algorithm can solve k-center under $(2-\epsilon)$-approximation stability, a notion that is stronger than perturbation resilience