k-Center Clustering Under Perturbation Resilience*†

Maria-Florina Balcan[1], Nika Haghtalab[2], and Colin White[3]

1 **Carnegie Mellon University, Pittsburgh, USA**
ninamf@cs.cmu.edu

2 **Carnegie Mellon University, Pittsburgh, USA**
nhaghtal@cs.cmu.edu

3 **Carnegie Mellon University, Pittsburgh, USA**
crwhite@cs.cmu.edu

Abstract

The k-center problem is a canonical and long-studied facility location and clustering problem with many applications in both its symmetric and asymmetric forms. Both versions of the problem have tight approximation factors on worst case instances: a 2-approximation for symmetric k-center and an O(log*(k))-approximation for the asymmetric version. Therefore to improve on these ratios, one must go beyond the worst case.

In this work, we take this approach and provide strong positive results both for the asymmetric and symmetric k-center problems under a very natural input stability (promise) condition called alpha-perturbation resilience [Bilu Linial, 2012], which states that the optimal solution does not change under any alpha-factor perturbation to the input distances. We show that by assuming 2-perturbation resilience, the exact solution for the asymmetric k-center problem can be found in polynomial time. To our knowledge, this is the first problem that is hard to approximate to any constant factor in the worst case, yet can be optimally solved in polynomial time under perturbation resilience for a constant value of alpha. Furthermore, we prove our result is tight by showing symmetric k-center under (2-epsilon)-perturbation resilience is hard unless NP=RP. This is the first tight result for any problem under perturbation resilience, i.e., this is the first time the exact value of alpha for which the problem switches from being NP-hard to efficiently computable has been found.

Our results illustrate a surprising relationship between symmetric and asymmetric k-center instances under perturbation resilience. Unlike approximation ratio, for which symmetric k-center is easily solved to a factor of 2 but asymmetric k-center cannot be approximated to any constant factor, both symmetric and asymmetric k-center can be solved optimally under resilience to 2-perturbations.

1998 ACM Subject Classification I.5.3 Clustering

Keywords and phrases k-center, clustering, perturbation resilience

Digital Object Identifier 10.4230/LIPIcs.ICALP.2016.68

* Full version of the paper available at http://arxiv.org/abs/1505.03924.

† This work was supported in part by grants NSF-CCF 1535967, NSF CCF-1422910, NSF CCF-145117, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, a Google Research Award, an IBM Ph.D. fellowship, and a National Defense Science & Engineering Graduate (NDSEG) fellowship.

43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016).
Editors: Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi;
Article No. 68; pp. 68:1–68:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Overview: Traditionally, the theory of algorithms has focused on the analysis of worst-case instances. While this approach has led to many elegant algorithms and strong lower bounds, it tends to be overly pessimistic of an algorithm's performance on the most typical instances of a problem. A recent line of work in the algorithms community, the so called *beyond worst case analysis* of algorithms, considers the question of designing algorithms for instances that satisfy some natural structural properties and has given rise to strong positive results [4, 5, 6, 18, 20, 21, 24]. One of the most appealing properties that has been proposed in this space is the stability of the solution to small changes in the input. Bilu and Linial [10] formalized this property in the notion of α-*perturbation resilience*, which states that the optimal solution does not change under any α-factor perturbation to the input distances.

A large body of work has sought to exploit the power of perturbation resilience in problems such as center-based clustering [5, 8, 10, 22], finding Nash equilibria in game theoretic problems [7], and the traveling salesman problem [23]. These works are focused on providing positive results for exactly solving the corresponding optimization problem under perturbation resilient instances, for example, $1+\sqrt{2}$-perturbation resilience for center based clustering, and $O(\sqrt{\log n}\log\log n)$-perturbation resilience for max-cut. In this paper we continue this line of work and provide a tight result for the canonical and long-studied k-center clustering problem, thereby completely quantifying the power of perturbation resilience for this problem. We show that $\alpha = 2$ is the moment where the problem switches from NP-hard to efficiently computable – specifically, we show that by assuming 2-perturbation resilience, the exact solution for the k-center problem can be found in polynomial time; we also show that k-center under $(2-\epsilon)$-perturbation resilience cannot be solved in polynomial time unless $NP = RP$. Our results apply to both symmetric and asymmetric k-center, illustrating a surprising relationship between symmetric and asymmetric k-center instances under perturbation resilience. Unlike approximation ratio, for which symmetric k-center is easily solved to a factor of 2 but asymmetric k-center cannot be approximated to any constant factor, both symmetric and asymmetric k-center can be solved optimally under resilience to 2-perturbations. Overall, this is the first tight result quantifying the power of perturbation resilience for a canonical combinatorial optimization problem.

The k-center problem is a canonical and long-studied clustering problem with many applications to facility location, data clustering, image classification, and information retrieval [11, 12, 13, 14, 16, 25]. For example, it can be used to solve the problem of placing k fire stations spaced throughout a city to minimize the maximum time for a fire truck to reach any location, given the pairwise travel times between important locations in the city. In the symmetric k-center problem the distances are assumed to be symmetric, while in the asymmetric k-center problem they are not; however in both cases they satisfy the triangle inequality. Formally, given a set of n points S, a distance function $d : S \times S \to R^+$ satisfying the triangle inequality (and symmetry in the symmetric case), and an integer k, our goal is to find k centers $\{c_1, \ldots, c_k\}$ to minimize $\max_{p\in S} \min_i d(c_i, p)$.

Both forms of k-center admit tight approximation bounds. For symmetric k-center, several 2- approximation algorithms have been found starting in the mid 1980s (e.g., [16, 19]). This is the best possible approximation factor by a simple reduction from set cover. On the other hand, the asymmetric k-center problem is a prototypical problem where the best known approximation is superconstant and is matched by a lower bound. For the asymmetric k-center problem, an $O(\log^*(n))$-approximation algorithm was found by Vishwanathan [25], and later improved to $O(\log^*(k))$ by Archer [1]. This approximation ratio was shown to be

asymptotically tight by the work of Chuzhoy et al. [14], which built upon a sequence of papers establishing the hardness of approximating d-uniform hypergraph covering (culminating in [15]).

Perturbation resilience has a natural interpretation for both symmetirc and asymmetric k-center: it can be viewed as a stability condition in the presence of uncertainties involved in measurements. For example, small fluctuations in the travel time between a fire station and locations in the city, which are caused by different levels of traffic at different times of day, should not drastically affect the optimal placement of fire stations. Furthermore, perturbation resilience can be viewed as a condition on an instance under which the optimal solution satisfies a form of privacy. For instance, if the actions of no individuals (such as how they drive to work, or the amount of network traffic they are using in a network application) can affect the overall state of the problem drastically then the individual's actions cannot be detected by looking at the optimal solution.

There is a large body of work on instances satisfying perturbation resilience and other natural notions of stability on problems ranging from clustering to data privacy to social networks to topic modeling [2, 3, 17, 18, 20, 21, 24]. For discussion of related work, see the full version of the paper.

Our Results: In this work we consider both symmetric and asymmetric k-center under perturbation resilience and give tight results for both forms. In addition, we consider more robust and weaker variants of perturbation resilience, and give strong results for these problems as well. A summary of our results and techniques used to achieve them are as follows:

1. *Efficient algorithm for symmetric and asymmetric k-center under 2-perturbation resilience.* This directly improves over the result of Balcan and Liang [8] for symmetric k-center under $1+\sqrt{2}$-perturbation resilience. We show that *any* α-approximation algorithm returns the optimal solution for an α-perturbation resilient instance, thus showing there exists an optimal algorithm for symmetric k-center under 2-perturbation resilience. For the asymmetric result, we first construct a "symmetrized set" by only considering points that demonstrate a rough symmetry. Then we prove strong structural results about the symmetrized set which motivates a novel algorithm for detecting clusters locally.
2. *Hardness of symmetric k-center under $(2-\epsilon)$-perturbation resilience.* This shows that our perturbation-resilience results are tight for both symmetric and asymmetric k-center. For this hardness result, we use a reduction from a variant of perfect dominating set. To show that this variant is itself hard, we construct a chain of parsimonious reductions (reductions which conserve the number of solutions) from 3-dimensional matching to perfect dominating set.
3. *Efficient algorithms for symmetric and asymmetric k-center under $(3,\epsilon)$-perturbation resilience.* A clustering instance satisfies (α,ϵ)-perturbation resilience if $\leq \epsilon n$ points switch clusters under any α-perturbation. We assume the optimal clusters are of size $> 2\epsilon n$ (the problem is NP-hard without this assumption). We show that if any single point p is close to an optimal cluster other than its own, then $k-1$ centers achieve the optimal score under a carefully constructed 3-perturbation. Any other point we add to the set of centers must create a clustering that is ϵ-close to $\mathcal{OPT}$, and we show all of these sets cannot simultaneously be consistent with one another, thus causing a contradiction. A key concept in our analysis is defining the notion of a *cluster-capturing center*, which allows us to reason about which points can capture a cluster when its center is removed.
4. *Efficient algorithm for any center-based clustering objective under weak center proximity.* Weak center proximity asks that each point be closer to its own center than to any point

from any other cluster, but note that it allows a cluster center to be closer to points from different clusters than to its own. Thus it is not at all obvious whether efficient optimal clustering is possible in such a setting. We present a novel linkage-based algorithm that is able to do so. It works by iteratively running single linkage as a subroutine until all clusters are balanced, and then removing all but the very last link.

The novelty of our results are manifold. First, our work is the first to provide a tight perturbation resilience result, thereby painting the complete picture for k-center under perturbation resilience. Second, this is the first result where a problem is not approximable to any constant in the worst-case, but can be optimally solved under resilience to small constant perturbations. Third, we are the first to consider an asymmetric problem under stability. Our results here illustrate a stark contrast between worst-case analysis and analysis of algorithms under stability. Unlike approximation ratio, for which symmetric k-center is easily solved to a factor of 2 but asymmetric k-center cannot be approximated to any constant factor, both symmetric and asymmetric k-center can be solved optimally under the same constant level of resilience.

2 Preliminaries

We define a clustering instance as (S, d), where S is a set of n points and $d : S \times S \to \mathbb{R}_{\geq 0}$ is a distance function. In the k-center problem, the goal is to find a set of points $\vec{p} = \{p_1, \ldots, p_k\} \subseteq S$ called *centers* such that the maximum distance from any point to its closest center is minimized. More formally, in the k-center problem, given a Voronoi partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ induced by a set of centers $\vec{p} = \{p_1, \ldots, p_k\}$ (where for all $1 \leq i \leq k$, $p_i \in P_i$), we refer to $\mathcal{P}$ as a clusering, and define its cost by $\Phi(\mathcal{P}) = \max_{i \in [k]} \max_{v \in P_i} d(p_i, v)$. We indicate by $\mathcal{OPT}$ the clustering $\{C_1, \ldots, \mathcal{C}_k\}$ with minimum cost, we denote the optimal centers $\{c_1, \ldots, c_k\}$, and we denote the optimal cost $\Phi(\mathcal{OPT})$ by r^*, the maximum cluster radius.

We study the k-center clustering of instance (S, d) under two types of distance functions, *symmetric* and *asymmetric*. A symmetric distance function is a metric. An asymmetric distance function satisfies all the properties of a metric space, except for symmetry. That is, it may be the case that for some $p, q \in S$, $d(p, q) \neq d(q, p)$. Note that the k-center objective function for asymmetric instances is the same as the symmetric case, the maximum distance *from the center to the points*, where the order now matters.

We consider *perturbation resilience*, a notion of stability introduced by Bilu & Linial [10]. Perturbation resilience implies that the optimal clustering does not change under small perturbations of the distance measure. Formally, d' is called an α-perturbation of distance function d, if for all $p, q \in S$, $d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$.[1] Perturbation resilience is defined formally as follows.

▶ **Definition 1.** A clustering instance (S, d) satisfies *α-perturbation resilience* for k-center, if for any α-perturbation d' of d, the optimal k-center clustering under d' is unique and equal to $\mathcal{OPT}$.

Note that the optimal centers may change, but the Voronoi partition $C_1, \ldots, C_k$ induced by them must stay the same. We do *not* assume that d' satisfies the triangle inequality.[2] We

[1] WLOG, we only consider perturbations in which the distances increases because we can scale the distances to simulate decreasing distances.

[2] This is well-justified, as the data may be gathered from heuristics or an average of measurements.

also consider a more robust variant of α-perturbation resilience, called (α, ϵ)-perturbation resilience, that allows a small change in the optimal clustering when distances are perturbed. To this end, we say that two clusterings $\mathcal{C}$ and $\mathcal{C}'$ are ϵ-*close*, if only an ϵ-fraction of the input points are clustered differently in the two clusterings, i.e., $\min_\sigma \sum_{i=1}^{k} |C_i \setminus C'_{\sigma(i)}| \leq \epsilon n$, where σ is a permutation on $[k]$. Formally,

▶ **Definition 2.** A clustering instance (S, d) satisfies (α, ϵ)-*perturbation resilience* for k-center, if for any α-perturbation d' of d, any optimal k-center clustering $\mathcal{C}'$ under d' is ϵ-close to $\mathcal{OPT}$.

We use ϵ-*far* to denote two clusters which are not ϵ-close. We also discuss the strictly stronger notion of approximation stability [6], which requires *any* α-approximation (not just a Voronoi partition) to be ϵ-close to $\mathcal{OPT}$. This is formally defined in Section 3.3. In Section 5, we define *center-based* objectives [8], a more general class of clustering functions which includes objective functions such as k-center, k-median, and k-means. Throughout this work, we use $B_r(c)$ to denote a ball of radius r centered at point r. Also for a point p and a set D, $d(p, D)$ denotes the distance from p to the farthest point in D.

3 2-perturbation resilience

In this section, we provide efficient algorithms for finding $\mathcal{OPT}$ for symmetric and asymmetric instances of k-center under 2-perturbation resilience. Our result directly improves on the result of Balcan and Liang for symmetric k-center under $(1 + \sqrt{2})$-perturbation resilience [8]. We also show that it is NP-hard to recover $\mathcal{OPT}$ even in the symmetric k-center instance under $(2 - \epsilon)$-approximation stability. As an immediate consequence, our results are tight for both perturbation resilience and approximation stability, for symmetric and asymmetric k-center instances. This is the first problem for which the exact value of perturbation resilience is found ($\alpha = 2$), where the problem switches from efficiently computable to NP-hard.

In the remainder of this section, first we show that any α-approximation algorithm returns the optimal solution for α-perturbation resilient instances. An immediate consequence is an algorithm for symmetric k-center under 2-perturbation resilience. Then we provide a novel algorithm for asymmetric k-center under 2-perturbation resilience.

3.1 Approximation algorithms under perturbation resilience

The following lemma allows us to reason about a specific type of α-perturbation we construct. This lemma will be important throughout the analysis in this section and in Section 4.

▶ **Lemma 3.** *For all $\alpha \geq 1$, given an α-perturbation d' of d with the following property: for all p, q, if $d(p, q) \geq r^*$ then $d'(p, q) \geq \alpha r^*$. Then the optimal cost under d' is αr^*.*

Proof. Clearly the optimal cost under d' cannot be greater than αr^*, since d' is an α-perturbation. Suppose there exists a set of centers $c'_1, \ldots, c'_k$ under d' that achieves a cost $< \alpha r^*$. Then for all i and all $p \in C'_i$, $d'(c'_i, p) < \alpha r^*$. But then by assumption, $d(c'_i, p) < r^*$. This implies that $c'_1, \ldots, c'_k$ achieve an optimal cost $< r^*$ under d, which is a contradiction. ◀

The following theorem will imply that any α-approximation algorithm for k-center will return the optimal solution on clustering instances that are α-perturbation resilient.

▶ **Theorem 4.** *Given a clustering instance (S, d) satisfying α-perturbation resilience for asymmetric k-center. Given a set C of k centers which is an α-approximation, i.e., $\forall p \in S$, $\exists c \in C$ s.t. $d(c, p) \leq \alpha r^*$. Then the Voronoi partition induced by C is the optimal clustering.*

Proof. For a point $p \in S$, let $c(p) := \text{argmin}_{c \in C} d(c, p)$, the closest center in C to p. The idea is to construct an α-perturbation in which C is the optimal solution by increasing all distances except between p and $c(p)$, for all p. Then the theorem will follow by using the definition of perturbation resilience.

By assumption, $\forall p \in S$, $d(c(p), p) \leq \alpha r^*$. Create a perturbation d' as follows. Increase all distances by a factor of α, except for all $p \in S$, set $d'(c(p), p) = \min(\alpha d(c(p), p), \alpha r^*)$ (recall in Definition 1, the perturbation need not satisfy the triangle inequality). Then no distances were increased by more than a factor of α. And since we had that $d(c(p), p) \leq \alpha r^*$, no distances decrease either. Therefore, d' is an α-perturbation of d. By Lemma 3, the optimal cost for d' is αr^*. Also, C achieves cost $\leq \alpha r^*$ by construction, so C is an optimal set of centers under d'. Then by α-perturbation resilience, the Voronoi partition induced by C under d' is the optimal clustering.

Finally, we show the Voronoi partition of C under d is the same as the Voronoi partition of C under d'. Given $p \in S$ whose closest point in C is $c(p)$ under d, then under d', all distances from p to $C \setminus \{c(p)\}$ increased by exactly α, and $d(p, c(p))$ increased by $\leq \alpha$. Therefore, the closest point in C to p under d' is still $c(p)$. ◀

An immediate consequence is that we have exact algorithms for symmetric k-center under 2-perturbation resilience, and asymmetric k-center under $O(\log^*(k))$-perturbation resilience. Now we show it is possible to substantially improve the latter result.

3.2 Asymmetric k-center algorithm

One of the challenges involved in dealing with asymmetric k-center instances is the fact that even though for all $p \in C_i$, $d(c_i, p) \leq r^*$, $d(p, c_i)$ might be arbitrarily large. Such points for which $d(p, c_i) \gg r^*$ pose a challenge to the structure of the clusters, as they can be very close to points or even centers of other clusters. To deal with this challenge, we first define a set of "good" points, A, such that $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$. Intuitively speaking, these points behave similarly to a set of points with symmetric distances up to a distance r^*. To explore this, we define a desirable property of A with respect to the optimal clustering.

▶ **Definition 5.** A is said to *respect the structure of* $\mathcal{OPT}$ if (1) $c_i \in A$ for all $i \in [k]$, and (2) for all $p \in S \setminus A$, if $A(p) := \arg\min_{q \in A} d(q, p) \in C_i$, then $p \in C_i$.

For all i, define $C'_i = C_i \cap A$ (which is in fact the optimal clustering of A, although we do not need to prove this). Satisfying Definition 5 implies that if we can optimally cluster A, then we can optimally cluster the entire instance (formalized in Theorem 8). Thus our goal is to show that A does indeed respect the structure of $\mathcal{OPT}$, and to show how to return $C'_1, \ldots, C'_k$.

Intuitively, A is similar to a symmetric 2-perturbation resilient clustering instance. However, some structure is no longer there, for instance, a point p may be at distance $\leq 2r^*$ from every point in a different cluster, which is not true for 2-perturbation resilient instances. This implies we cannot simply run a 2-approximation algorithm on the set A, as we did in the previous section. However, we show that the remaining structural properties are sufficient to optimally cluster A. To this end, we define two properties and show how they lead to an algorithm that returns $C'_1, \ldots, C'_k$, and help us prove that A respects the structure of $\mathcal{OPT}$.

The first of these properties requires each point to be closer to its center than any point in another cluster. That is, *Property (1): For all $p \in C'_i$ and $q \in C'_j$, $i \neq j$, $d(c_i, p) < d(q, p)$.*

The second property requires that any point within distance r^* of a cluster center belongs to that cluster. That is, *Property (2): For all $i \neq j$ and $q \in C_j$, $d(q, c_i) > r^*$.*[3]

Let us illustrate how these properties allow us to optimally cluster A.[4] Consider a ball of radius r^* around a center c_i. By Property 2, such a ball exactly captures C'_i. Furthermore, by Property 1, any point in this ball is closer to the center than to points outside of the ball. Is this true for a ball of radius r^* around a general point p? Not necessarily. If this ball contains a point $q \in C'_j$ from a different cluster, then q will be closer to a point outside the ball than to p (namely, c_j, which is guaranteed to be outside of the ball by Property 2). This allows us to determine that the center of such a ball must not be an optimal center.

This structure motivates our Algorithm 1 for asymmetric k-center under 2-perturbation resilience. At a high level, we start by constructing the set A (which can be done easily in polynomial time). Then we create the set of all balls of radius r^* around all points in A (if r^* is not known, we can use a guess-and-check wrapper). Next, we prune this set by throwing out any ball that contains a point farther from its center than to a point outside the ball. We also throw out any ball that is a subset of another one. Our claim is that the remaining balls are exactly $C'_1, \ldots, C'_k$. Finally, we add the points in $S \setminus A$ to their closest point in A.

Algorithm 1 Asymmetric k-center algorithm under 2-PR

Input: Asymmetric k-center instance (S, d), distance r^* (or try all possible candidates).

1. Build set $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$
2. $\forall c \in A$, construct $G_c = B_{r^*}(c)$ (the ball of radius r^* around c).
3. $\forall G_c$, if $\exists p \in G_c$, $q \notin G_c$ s.t. $d(q, p) < d(c, p)$, then throw out G_c.
4. $\forall p, q$ s.t. $G_p \subseteq G_q$, throw out G_p.
5. $\forall p \notin A$, add p to G_q, where $q = \arg\min_{s \in A} d(s, p)$.

Output: Output the sets $G_1, \ldots, G_k$.

▶ **Lemma 6.** *Properties 1 & 2 hold for asymmetric k-center instances under 2-perturbation resilience.*

Proof sketch. For Property 2, assume that there exists c_i and $q \in C_j$, $i \neq j$, such that $d(q, c_i) \leq r^*$. We construct a 2-perturbation in which q becomes the center for C_i. Increase all distances by a factor of 2, except for the distances from q to C_i, which we increase until they reach $2r^*$. By Lemma 3, this 2-perturbation achieves a cost of $2r^*$. However, q is distance $2r^*$ to C_i, so it must replace c_i as an optimal center. Then q and c_j are no longer in the same cluster, causing a contradiction.

The first property was shown to hold for symmetric instances by Awasthi et al. and the same proof can be used for asymmetric instances. This proof appears in the full version. ◀

▶ **Lemma 7.** *A respects the structure of $\mathcal{OPT}$.*

We defer this proof to the full version of the paper.

[3] Property (1) first appeared in the work of Awasthi et al. [5], for symmetric clustering instances. A weaker variation of Property (2) was introduced by Balcan and Liang [8], which showed that in $1 + \sqrt{2}$-perturbation resilient instances for any cluster C_i with radius r_i, $B_{r_i}(c_i) = C_i$. Our Property (2) shows that this is true *for a universal radius, r^**, even for 2-perturbation resilient instances, *and* even for asymmetric instances.

[4] Other algorithms work, such as single linkage with dynamic programming at the end to find the minimum cost pruning of k clusters. However, our algorithm is able to recognize optimal clusters *locally* (without a complete view of the point set).

▶ **Theorem 8.** *Algorithm 1 returns the exact solution for asymmetric k-center under 2-perturbation resilience.*

Proof. First we must show that after step 4, the remaining sets are exactly $C'_1, \ldots, C'_k = C_1 \cap A, \ldots, C_k \cap A$. We prove this in three steps: the sets G_{c_i} correspond to C'_i, these sets are not thrown out in steps 3 and 4, and all other sets are thrown out in steps 3 and 4. Because of Lemma 6, we can use Properties 1 and 2.

For all i, $G_{c_i} = C'_i$: From Lemma 7, all centers are in A, so G_{c_i} will be created in step 2. For all $p \in C_i$, $d(c_i, p) \leq r^*$. For all $q \notin C'_i$, then by Property 2, $d(q, c_i) > r^*$ (and since $c_i, q \in A$, $d(c_i, q) > r^*$ as well). For all i, G_{c_i} is not thrown out in step 3: Given $s \in G_{c_i}$ and $t \notin G_{c_i}$. Then $s \in C'_i$ and $t \in C'_j$ for $j \neq i$. If $d(t, s) < d(c_i, s)$, then we get a contradiction from Property 1. For all non-centers p, G_p is thrown out in step 3 or 4: From the previous paragraph, $G_{c_i} = C'_i$. If $G_p \subseteq G_{c_i}$, then G_p will be thrown out in step 4 (if $G_p = G_{c_i}$, it does not matter which set we keep, so WLOG say that we keep G_{c_i}). Then if G_p is not thrown out in step 4, $\exists s \in G_p \cap C'_j$, $j \neq i$. If $s = c_j$, then $d(p, c_j) \leq r^*$ and we get a contradiction from Property 2. So, we can assume s is a non-center (and that $c_j \notin G_p$). But $d(c_j, s) < d(p, s)$ from Property 1, and therefore G_p will be thrown out in step 3. Thus, the remaining sets after step 4 are exactly $C'_1, \ldots, C'_k$.

Finally, by Lemma 7, for each $p \in C_i \setminus A$, $A(p) \in C_i$, so p will be added to G_{c_i}. Therefore, the final output is $C_1, \ldots, C_k$. ◀

3.3 Hardness for k-center under $(2 - \epsilon)$-approximation stability

In this section, we consider approximation stability, introduced by Balcan et al. [6], which is strictly stronger than perturbation resilience. We show that if symmetric k-center under $(2-\epsilon)$-approximation stability can be solved in polynomial time, then $NP = RP$, even under the condition that the optimal clusters are all $\geq \frac{n}{2k}$. Because approximation stability is stronger than perturbation resilience, this result implies k-center under $(2 - \epsilon)$-perturbation resilience is hard as well. Similarly, symmetric k-center is a special case of asymmetric k-center, so we get the same hardness results for asymmetric k-center. This proves that Theorem 8 is tight.

Approximation stability requires constant approximations to the optimal cost to differ from $\mathcal{OPT}$ by at most an ϵ-fraction of the points.

▶ **Definition 9.** A clustering instance (S, d) satisfies (α, ϵ)-*approximation stability* for k-center, if for any partition $\mathcal{C}'$ with objective value r' (not necessarily a Voronoi partition), if $r' \leq \alpha r^*$, then $\mathcal{C}'$ is ϵ-close to $\mathcal{OPT}$.

It is not hard to see that (α, ϵ)-approximation stability implies (α, ϵ)-perturbation resilience, as the optimal clustering under any α-perturbation costs at most αr^* under the original distance function, d. So, a violating instance of (α, ϵ)-perturbation resilience induces a partition which costs $\leq \alpha r^*$ and is ϵ-far from $\mathcal{OPT}$, and therefore is not (α, ϵ)-approximation stable.

▶ **Theorem 10.** *There is no polytime algorithm for finding the optimal k-center clustering under $(2 - \epsilon)$-approximation stability, even when assuming all optimal clusters are size $\geq \frac{n}{2k}$, unless $NP = RP$.*

We show a reduction from a special case of Dominating Set which we call Unambiguous-Balanced-Perfect Dominating Set. A reduction from Perfect Dominating Set (Dominating Set with the additional constraint that for all dominating sets of size $\leq k$, each vertex is hit by

exactly one dominator) to the problem of clustering under $(2-\epsilon)$-center proximity was shown in [9] (α-center proximity is the property that for all $p \in C_i$ and $j \neq i$, $\alpha d(c_i, p) < d(c_j, p)$, and it follows from α-perturbation resilience). Our contribution is to show that Perfect Dominating Set remains hard under two additional conditions. First, in the case of a YES instance, each dominator must hit at least $\frac{n}{2k}$ vertices (which translates to clusters having size at least $\frac{n}{2k}$ as well). Second, we are promised that there is at most one dominating set of size $\leq k$ (which is required for establishing approximation stability for the resulting clustering instance).

4 Robust perturbation resilience

In this section, we consider (α, ϵ)-perturbation resilience. We show that under $(3, \epsilon)$-perturbation resilience, there is an algorithm that recovers $\mathcal{OPT}$ for symmetric k-center, and an algorithm that returns a solution that is ϵ-close to $\mathcal{OPT}$ for asymmetric k-center. For both of these results, we assume a lower bound on the size of the optimal clusters, $|C_i| > 2\epsilon n$ for all $i \in [k]$. We show the lower bound on cluster sizes is necessary; in its absence, the problem becomes NP-hard for all values of $\alpha \geq 1$ and $\epsilon > 0$. The theorems in this section require a careful reasoning about sets of centers under different perturbations that cannot all simultaneously be valid.

4.1 Symmetric k-center

We show that for any $(3, \epsilon)$-perturbation resilient k-center instance such that $|C_i| > 2\epsilon n$ for all $i \in [k]$, $\mathcal{OPT}$ can be found by simply thresholding the input graph using distance r^* and outputting the connected components. A nice feature of our result is that the Single Linkage algorithm, a fast algorithm widely used in practice, is sufficient to optimally cluster these instances.

▶ **Theorem 11.** *Given a $(3, \epsilon)$-perturbation resilient k-center instance (S, d) where all optimal clusters are $> \max(2\epsilon n, 3)$. Then the optimal clusters in $\mathcal{OPT}$ are exactly the connected components of the threshold graph G_{r^*} of the input distances.*

Proof idea. Since each optimal cluster center is distance r^* from all points in its cluster, it suffices to show that any two points in different clusters are at least r^* apart from each other. Assume on the contrary that there exist $p \in C_i$ and $q \in C_j$, $i \neq j$, such that $d(p, q) \leq r^*$. First we find a set of $k+2$ points and a 3-perturbation d', such that every size k subset of the points are optimal centers under d'. Then we show how this leads to a contradiction under $(3, \epsilon)$-perturbation resilience.

From our assumption, p is distance $\leq 3r^*$ from every point in $C_i \cup C_j$ (by the triangle inequality). Under a 3-perturbation in which all distances are blown up by a factor of 3 except $d(p, C_i \cup C_j)$, then replacing c_i and c_j with p would still give us a set of $k-1$ centers that achieve the optimal score. But, *would this contradict $(3, \epsilon)$-perturbation resilience?* Indeed, not! Perturbation resilience requires exactly k *distinct* centers.[5] The key challenge is to pick a final "dummy" center to guarantee that the Voronoi partition is ϵ-far from $\mathcal{OPT}$. The dummy center might "accidentally" be the closest center for almost all points in C_i or C_j. Even worse, it might be the case that the new center sets off a chain reaction in which it

[5] This distinction is well-motivated; if for some application, the best k-center solution is to put two centers at the same location, then we could achieve the exact same solution with $k-1$ centers. That implies we should have been running k'-center for $k' = k-1$ instead of k.

becomes center to a cluster C_x, and c_x becomes center to C_j, which would also result in a partition that is not ϵ-far from $\mathcal{OPT}$.

To deal with the chain reactions, we crucially introduce the notion of a *cluster capturing center* (CCC). c_x is a CCC for C_y, if for all but ϵn points $p \in C_y$, $d(c_x, p) \leq r^*$ and for all $i \neq x, y$, $d(c_x, p) < d(c_i, p)$. Intuitively, a CCC exists if and only if c_x is a valid center for C_y when c_y is taken out of the set of optimal centers (i.e., a chain reaction will occur). We argue that if a CCC does not exist then every dummy center we pick must be close to either C_i or C_j, since there are no chain reactions. If there does exist a CCC c_x for C_y, then we cannot reason about what happens to the dummy centers under our d'. However, we can define a new d'' by increasing all distances except $d(c_x, C_y)$, which allows us to take c_y out of the set of optimal centers, and then any dummy center must be close to C_x or C_y. There are no chain reactions because *we already know c_x is the best center for C_y* among the original optimal centers. Thus, whether or not there exists a CCC, we can find $k+2$ points close to the entire dataset by picking points from both C_i and C_j (resp. C_x and C_y).

Because of the assumption that all clusters are size $> 2\epsilon n$, for every 3-perturbation there must be a bijection between clusters and centers, where the center is closest to the majority of points in the corresponding cluster. We show that all size k subsets of the $k+2$ points cannot simultaneously admit bijections that are consistent with one another.

Note that Theorem 10 implies $(2-\delta, \epsilon)$- perturbation resilient k-center is hard for $\delta > 0$, even when the optimal clusters are large. Therefore, the value of α we achieve is within one of optimal. ◀

4.2 Lower bound on cluster sizes

Before moving to the asymmetric case, we show that the lower bound on the cluster sizes in Theorem 11 is necessary. Without this lower bound, clustering becomes hard, even assuming (α, ϵ)-perturbation resilience for any α and ϵ. This reduction follows from k-center (the details appear in the full version).

▶ **Theorem 12.** *For all $\alpha \geq 1$ and $\epsilon > 0$, finding the optimal solution for k-center under (α, ϵ)-perturbation resilience is NP-hard.*

4.3 Asymmetric k-center

In the asymmetric case, we consider the definition of the symmetric set A from Section 3, $A = \{p \mid \forall q, d(q,p) \leq r^* \implies d(p,q) \leq r^*\}$. We might first ask whether A respects the structure of $\mathcal{OPT}$, as it did under 2-perturbation resilience. Namely, whether *Condition 1:* all centers are in A, and *Condition 2:* $\arg\min_{q \in A} d(q,p) \in C_i \implies p \in C_i$ hold. This is not the case for either condition. We explore to what degree these conditions are violated.

We call a center c_i "bad" if it is not in the set A, i.e., $\exists q \notin C_i$ and $d(q, c_i) \leq r^*$. When a bad center c_i exists, we can take it out of the set of optimal centers, and we can pick an arbitrary dummy center which must be close to C_i or a CCC for C_i. In our symmetric argument, we arrived at a contradiction by showing that two dummy centers which capture the same cluster, must be close by the triangle inequality. This logic breaks down for asymmetric distances. In the full version of the paper, we show an example of an instance with a bad center that satisfies (α, ϵ)-perturbation resilience. However, it turns out that *no* instance can have more than 6 bad centers under $(3, \epsilon)$-perturbation resilience, assuming all optimal clusters have size $> 2\epsilon n$. So Condition 1 is satisfied for all but a constant number of centers. However, Condition 2 may not be satisfied for up to ϵn points. Therefore, even if we fully cluster A, we will only get ϵ-close to $\mathcal{OPT}$.

Although up to 6 clusters may have no intersection with A, each point that does belong to A is distance r^* from its center and distance $2r^*$ from its entire cluster. This motivates the following algorithm. First, we run a symmetric k-center 2-approximation algorithm on A, for $k-6 \leq k' \leq k$. For instance, iteratively pick an unmarked point, and mark all points distance $2r^*$ away from it [19]. This gives us a 2-approximation for the centers in A, and thus a 3-approximation for S minus the clusters with no centers in A. Then we brute force search for the remaining ≤ 6 centers to find a 3-approximation for S. Under $(3,\epsilon)$-perturbation resilience, this 3-approximation must be ϵ-close to $\mathcal{OPT}$. We formally state the algorithm and theorem below, and we defer the proof to the full version of this paper. The main technical challenge is in proving that no instance can have more than 6 bad centers.

Algorithm 2 $(3,\epsilon)$-Perturbation Resilient Asymmetric k-center

Input: Asymmetric k-center instance (S,d), r^* (or try all possible candidates).

1. Build set $A = \{p \mid \forall q, d(q,p) \leq r^* \implies d(p,q) \leq r^*\}$.
2. Create the threshold graph G with vertices A, and threshold distance r^*. Define a new symmetric k-center instance with A, using the lengths of the paths in the threshold graph.
3. Run a symmetric k-center 2-approximation algorithm on the symmetrized instance. Start with $k' = k-6$, and increase k' by 1 until the algorithm returns a solution with radius $\leq 2r^*$.
4. Brute force over all size $k-x$ subsets of C and all size x subsets of S for $x \leq 6$, to find a set of size k which is $3r^*$ from all points in S. Denote this set by C'.

Output: Output the Voronoi tiling $G_1, \ldots, G_k$ using C' as the centers.

▶ **Theorem 13.** *Algorithm 2 runs in polytime and outputs a clustering that is ϵ-close to $\mathcal{OPT}$, for $(3,\epsilon)$-perturbation resilient asymmetric k-center instances s.t. all optimal clusters are size $> 2\epsilon n$.*

5 Weak center proximity

In this section, we consider any center-based objective, not just k-center. A clustering objective function is *center-based* if the solution can be defined by choosing a set of centers $\{c_1, c_2, \ldots, c_k\} \subseteq S$, and partitioning S into k clusters $\mathcal{OPT} = \{C_1, C_2, \ldots, C_k\}$ by assigning each point to its closest center. Furthermore:

1. The objective value of a given clustering is a weighted sum or maximum of the individual cluster scores.
2. Given a proposed single cluster, its score can be computed in polynomial time.

k-median, k-means, and k-center are all center-based objectives.

Here, we show a novel algorithm that finds the optimal clustering in instances that satisfy two simple properties: each point is closer to its center than to any point in a different cluster, and we can recognize optimal clusters as soon as they are formed. Formally, we define these properties as:

1. **Weak Center Proximity:** For all $p \in C_i$ and $q \in C_j$, $d(c_i,p) < d(p,q)$.
2. **Cluster Verifiability:** There exists a polytime computable function $f : 2^S \to \mathbb{R}$ that for $B \subseteq S$, if there is $i \in [k]$ such that $B \subset C_i$, then $f(B) < 0$, and if $B \supseteq C_i$, then $f(B) \geq 0$.

Examples of cluster verifiable instances include any instance where all the optimal clusters are the same size ($f(B) = |B| - \frac{n}{k}$), or where all the optimal clusters have the same k-median/k-means cost ($f(B) = \Phi(B) - \Phi(\mathcal{OPT})$).

For any center-based objective, weak center proximity is a consequence of 2-perturbation resilience (i.e., Lemma 6), so, our algorithm relies on a much weaker assumption than α-perturbation resilience for $\alpha \geq 2$, when instances are cluster verifiable.

All existing algorithms and analysis for α-perturbation resilience require that for all $p \in C_i$ and $q \in C_j$, $d(c_i, p) < d(c_i, q)$. It is not at all obvious how one can even proceed without such a property, as in its absence, clusters can 'overlap'. That is, for a cluster with center c_i and radius r, we can not assume that $B_r(c_i)$ only includes points from C_i. Our challenge is then in showing that even in absence of this property, there is still enough structure imposed by the weak center proximity and cluster verifiability to find the optimal clustering efficiently.

Our Algorithm 3 is a novel linkage based procedure. Given a clustering instance (S, d), we will start with a graph $G = (S, E)$ where $E = \emptyset$. In each round, we do single linkage on the components in G, except we do not merge two components if both are supersets of optimal clusters (indicated by $f(B) \geq 0$). Put the single linkage edges from this round in a set A. This will continue until every component is a superset of an optimal cluster. Then we throw away the set A except for the *very last edge* that was added. We will prove this last edge is never between two points from different clusters, so we add that single edge to E and then recur. Here, we present a proof sketch of our main theorem. The details can be found in the full version.

Algorithm 3 Clustering under weak center proximity and cluster verifiability

Input: Clustering instance (S, d), function f, and $k \leq |S|$.

Set $G = (S, E)$ and $E = \emptyset$. While there are more than k components in G, repeat (1) and (2):

1. Set $A = \emptyset$. While there exists a component B in $G' = (S, E \cup A)$ such that $f(B) < 0$, add (p, q) to A, where $d(p, q)$ is minimized such that p and q are in different components in G' and at least one of these components B has $f(B) < 0$.
2. Take the last edge e that was added to A, and put $e \in E$.

Output: Output the components of G.

▶ **Theorem 14.** *Given a center-based clustering instance satisfying weak center proximity and cluster verifiability, Algorithm 3 outputs $\mathcal{OPT}$ in polynomial time.*

Proof Sketch. It suffices to show that step (b) never adds an edge between two points from different clusters. We proceed by induction. Assume it is true up to iteration t of the first while loop. Now assume towards contradiction that in round t, the last edge added to A is between two points $p \in C_i$ and $q \in C_j$, $i \neq j$. WLOG, for the component in G' that includes p, called P', we have $f(P') < 0$, otherwise the merge would not have happened. Furthermore, $c_i \in P'$ by weak center proximity. Then $f(P') < 0$ implies that $C_i \setminus P'$ is nonempty, so call it P. The component(s) in G corresponding to P are strict subsets of C_i, therefore, $f(P) < 0$. So they must merge to another component, and by weak center proximity, the closest component is P', but this contradicts our assumption that (p, q) was the last edge added to A. ◀

6 Conclusions

Our work pushes the understanding of (promise) stability conditions farther in three ways. We are the first to design computationally efficient algorithms to find the optimal clustering under α-perturbation resilience with a constant value of α for a problem that is hard to

approximate to any constant factor in the worst case, thereby demonstrating the power of perturbation resilience. Furthermore, we demonstrate the limits of this power by showing the first tight results in this space for both perturbation resilience and approximation stability. Finally, we show a surprising relation between symmetric and asymmetric instances, in that they are equivalent under resilience to 2-perturbations, which is in stark contrast to their widely differing tight approximation factors.

References

1 Aaron Archer. Two o (log* k)-approximation algorithms for the asymmetric k-center problem. In *Integer Programming and Combinatorial Optimization*, pages 1–14. Springer, 2001.

2 Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models – going beyond SVD. In *53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 1–10, 2012.

3 Pranjal Awasthi, Afonso S. Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200, 2015. doi:10.1145/2688073.2688116.

4 Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 309–318, 2010.

5 Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

6 Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077, 2009.

7 Maria-Florina Balcan and Mark Braverman. Approximate nash equilibria under stability conditions. Technical report, 2010.

8 Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. In *Automata, Languages, and Programming*, pages 63–74. Springer, 2012.

9 Shalev Ben-David and Lev Reyzin. Data stability in clustering: A closer look. In *Algorithmic Learning Theory*, pages 184–198. Springer, 2012.

10 Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(05):643–660, 2012.

11 Fazli Can. Incremental clustering for dynamic information processing. *ACM Transactions on Information Systems (TOIS)*, 11(2):143–164, 1993.

12 Fazli Can and ND Drochak. Incremental clustering for dynamic document databases. In *Proceedings of the 1990 Symposium on Applied Computing*, pages 61–67, 1990.

13 Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 626–635, 1997.

14 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. *Journal of the ACM (JACM)*, 52(4):538–551, 2005.

15 Irit Dinur, Venkatesan Guruswami, Subhash Khot, and Oded Regev. A new multilayered pcp and the hardness of hypergraph vertex cover. *SIAM Journal on Computing*, 34(5):1129–1146, 2005.

16 Martin E Dyer and Alan M Frieze. A simple heuristic for the p-centre problem. *Operations Research Letters*, 3(6):285–288, 1985.

17 Rishi Gupta, Tim Roughgarden, and C Seshadhri. Decompositions of triangle-dense graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 471–482. ACM, 2014.

18 Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340, 2013.

19 Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

20 Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 299–308, 2010.

21 Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+ ε)-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings-Annual Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

22 Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Bilu-linial stable instances of max cut and minimum multiway cut. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 890–906. SIAM, 2014.

23 Matúš Mihalák, Marcel Schöngens, Rastislav Šrámek, and Peter Widmayer. On the complexity of the metric tsp under stability considerations. In *SOFSEM 2011: Theory and Practice of Computer Science*, pages 382–393. Springer, 2011.

24 Tim Roughgarden. Beyond worst-case analysis, 2014. URL: `http://theory.stanford.edu/~tim/f14/f14.html`.

25 Sundar Vishwanathan. An o(log*n) approximation algorithm for the asymmetric p-center problem. In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1–5, 1996. URL: `http://dl.acm.org/citation.cfm?id=313852.313861`.