k-center Clustering under Perturbation Resilience

MARIA-FLORINA BALCAN, Carnegie Mellon University, USA
NIKA HAGHTALAB, Cornell University, USA
COLIN WHITE, RealityEngines.AI, USA

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 natural input stability (promise) condition called *α-perturbation resilience* [15], which states that the optimal solution does not change under any α-factor perturbation to the input distances. We provide algorithms that give strong guarantees simultaneously for stable and non-stable instances: Our algorithms always inherit the worst-case guarantees of clustering approximation algorithms and output the optimal solution if the input is 2-perturbation resilient. In particular, we show that if the input is only perturbation resilient on part of the data, our algorithm will return the optimal clusters from the region of the data that is perturbation resilient while achieving the best worst-case approximation guarantee on the remainder of the data. Furthermore, we prove that our result is tight by showing symmetric k-center under $(2-\epsilon)$-perturbation resilience is hard unless $NP = RP$.

The impact of our results is multifaceted. First, to our knowledge, asymmetric k-center 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 α. This is also the first tight result for any problem under perturbation resilience, i.e., this is the first time the exact value of α for which the problem switches from being NP-hard to efficiently computable has been found. Furthermore, 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. Finally, our guarantees in the setting where only part of the data satisfies perturbation resilience make these algorithms more applicable to real-life instances.

This article combines and extends results appearing in ICALP 2016 "k-center Clustering under Perturbation Resilience" and arxiv 2017 "Clustering under Local Stability: Bridging the Gap between Worst-Case and Beyond Worst-Case Analysis." This work was supported in part by NSF grants CCF-1910321, CCF 1535967, CCF-1422910, CCF-145117, IIS-1618714, an AWS Machine Learning Research Award, a Bloomberg Data Science research grant, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, a Google Research Award, an IBM Ph.D. fellowship, a National Defense Science & Engineering Graduate (NDSEG) fellowship, and by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures program. Most of this work was completed while Nika Haghtalab and Colin White were students at Carnegie Mellon University.

Authors' addresses: M.-F. Balcan, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213; email: ninamf@cs.cmu.edu; N. Haghtalab, Cornell University, 616 Thurston Ave, Ithaca, NY, 14853; email: nika@cs.cornell.edu; C. White, RealityEngines.AI, 1099 Folsom St. San Francisco, CA, 94103; email: crwhite@cs.cmu.edu.

1549-6325/2020/03-ART22
https://doi.org/10.1145/3381424

CCS Concepts: • **Theory of computation → Facility location and clustering**;

Additional Key Words and Phrases: Beyond worst-case analysis, clustering, perturbation resilience

ACM Reference format:
Maria-Florina Balcan, Nika Haghtalab, and Colin White. 2020. *k*-center Clustering under Perturbation Resilience. *ACM Trans. Algorithms* 16, 2, Article 22 (March 2020), 39 pages.
https://doi.org/10.1145/3381424

1 INTRODUCTION

Clustering is a fundamental problem in combinatorial optimization with a wide range of applications including bioinformatics, computer vision, text analysis, and countless others. The underlying goal is to partition a given set of points to maximize similarity within a partition and minimize similarity across different partitions. A common approach to clustering is to consider an objective function over all possible partitionings and seek solutions that are optimal according to the objective. Given a set of points (and a distance metric), common clustering objectives include finding k centers to minimize the sum of the distance from each point to its closest center (k-median) or to minimize the maximum distance from a point to its closest center (k-center).

Traditionally, the theory of clustering (and more generally, the theory of algorithms) has focused on the analysis of worst-case instances [4, 16–18, 20, 27, 39]. For example, it is well known that popular objective functions are provably NP-hard to optimize exactly or even approximately (APX-hard) [27, 32, 38], so research has focused on finding approximation algorithms. While this perspective has led to many elegant approximation algorithms and lower bounds for worst-case instances, it is often overly pessimistic of an algorithm's performance on "typical" instances or real-world instances. A rapidly developing line of work in the algorithms community, the so-called *beyond worst-case analysis* of algorithms (BWCA), considers the design and analysis of problem instances under natural structural properties that may be satisfied in real-world applications. For example, the popular notion of α-perturbation resilience, introduced by Bilu and Linial [15], considers instances such that the optimal solution does not change when the input distances are allowed to increase by up to a factor of $\geq \alpha$. The goals of BWCA are twofold: *(1)* to design new algorithms with strong performance guarantees under the added assumptions [8, 29, 36, 44] and *(2)* to prove strong guarantees under BWCA assumptions for existing algorithms used in practice [40, 43, 45]. An example of goal *(1)* is a series of works focused on finding exact algorithms for k-median, k-means, and k-center clustering under α-perturbation resilience [1, 6, 11]. The goal in this line of work is to find the minimum value of $\alpha \geq 1$ for which optimal clusterings of α-perturbation resilient instances can be found efficiently. Two examples of goal *(2)* are as follows: Ostrovsky et al. showed that k-means++ outputs a near-optimal clustering as long as the data satisfy a natural clusterability criterion [43]; and Spielman and Teng [45] established that the expected runtime of the simplex method is $O(n)$ under smoothed analysis.

In approaches for answering goals *(1)* and *(2)*, researchers have developed an array of sophisticated tools exploiting the structural properties of such instances leading to algorithms that output the optimal solution. However, overly exploiting a BWCA assumption can lead to algorithms that perform poorly when the input data do not exactly satisfy the given assumption. Indeed, recent analyses and technical tools are susceptible to small deviations from the BWCA assumptions that can propagate when just a small fraction of the data does not satisfy the assumption. For example, some recent algorithms make use of a dynamic programming subroutine that crucially needs the entire instance to satisfy the specific structure guaranteed by the BWCA assumption. To continue the efforts of BWCA in bridging the theory-practice gap, it is essential to study

Table 1. Our Results over All Variants of *k*-center under Perturbation Resilience

Problem	Guarantee	α	Metric	Local	Theorem
Symmetric *k*-center under α-PR	$\mathcal{OPT}$	2	Yes	Yes	Theorem 5.1
Asymmetric *k*-center under α-PR	$\mathcal{OPT}$	2	Yes	Yes	Theorem 5.2
Symmetric *k*-center under (α, ϵ)-PR	$\mathcal{OPT}$	3	No	Yes	Theorem 6.8
Asymmetric *k*-center under (α, ϵ)-PR	ϵ-close	3	No	No	Theorem 6.17

algorithms whose guarantees degrade gracefully to address scenarios that present mild deviations from the standard BWCA assumptions. Another downside of existing approaches is that BWCA assumptions are often not efficiently verifiable. This creates a catch-22 scenario: It is only useful to run the algorithms if the data satisfy certain assumptions, but a user cannot check these assumptions efficiently. For example, by nature of α-perturbation resilience (that the optimal clustering does not change under *all* α-perturbations of the input), it is not known how to test this condition without computing the optimal clustering over $\Omega(2^n)$ different perturbations. To alleviate these issues, in this work, we also focus on what we propose should be a third goal for BWCA: *(3)* to show (new or existing) algorithms whose performance degrades gracefully on instances that only partially meet the BWCA assumptions.

1.1 Our Results and Techniques

In this work, we address goals *(1)*, *(2)*, and *(3)* of BWCA by providing algorithms that give the optimal solution under perturbation resilience and also perform well when the data are partially perturbation resilient or not at all perturbation resilient. These algorithms act as an interpolation between worst-case and beyond worst-case analysis. We focus on the symmetric/asymmetric *k*-center objective under perturbation resilience. Our algorithms simultaneously output the optimal clusters from the stable regions of the data while achieving state-of-the-art approximation ratios over the rest of the data. In most cases, our algorithms are natural modifications to existing approximation algorithms, thus achieving goal *(2)* of BWCA. To achieve these two-part guarantees, we define the notion of perturbation resilience on a subset of the datapoints. All prior work has only studied perturbation resilience as it applies to the entire dataset. Informally, a subset $S' \subseteq S$ satisfies α-perturbation resilience if all points $v \in S'$ remain in the same optimal cluster under any α-perturbation to the input. We show that our algorithms return all optimal clusters from these locally stable regions. Most of our results also apply under the recently defined, weaker condition of *α-metric perturbation resilience* [1], which states that the optimal solution cannot change under the metric closure of any α-perturbation. We list all our results in Table 1 and give a summary of the results and techniques below.

k-center under 2-perturbation resilience. In Section 3, we show that *any* 2-approximation algorithm for *k*-center will always return the clusters satisfying 2-perturbation resilience. Therefore, since there are well-known 2-approximation algorithms for symmetric *k*-center, our analysis shows that these will output the optimal clustering under 2-perturbation resilience. For asymmetric *k*-center, we give a new algorithm that outputs the optimal clustering under 2-perturbation resilience. It works by first computing the "symmetrized set," or the points that demonstrate a rough symmetry. We show how to optimally cluster the symmetrized set, and then we show how to add back the highly asymmetric points into their correct clusters.

Hardness of symmetric k-center under $(2-\delta)$-perturbation resilience. In Section 3.3, we prove that there is no polynomial time algorithm for symmetric *k*-center under $(2-\delta)$-perturbation resilience unless $NP = RP$, which shows that our perturbation resilience results are tight for both

symmetric and asymmetric k-center. In particular, it implies that we have identified the exact threshold ($\alpha = 2$) where the problem switches from efficiently computable to NP-hard 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 that preserve the number of solutions) from 3-dimensional matching to perfect dominating set.

Our upper bound for asymmetric k-center under 2-perturbation resilience and lower bound for symmetric k-center under $(2 - \delta)$-perturbation resilience 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. Overall, this is the first tight result quantifying the power of perturbation resilience for a canonical combinatorial optimization problem.

Local perturbation resilience. In Section 5, we apply our results from Section 3 to the local perturbation resilience setting. For symmetric k-center, we show that any 2-approximation algorithm outputs all optimal clusters from 2-perturbation resilient regions. For asymmetric k-center, we design a new algorithm based on the worst-case $O(\log^* n)$ approximation algorithm due to Vishwanathan [48], which is tight [21]. We give new insights into this algorithm, which allow us to show that a modification of the algorithm outputs all optimal clusters from 2-perturbation resilient regions, while keeping the worst-case $O(\log^* n)$ guarantee overall. If the entire dataset satisfies 2-perturbation resilience, then our algorithm outputs the optimal clustering. We combine the tools of Vishwanathan with the perturbation resilience assumption to prove this two-part guarantee. Specifically, we use the notion of a *center-capturing vertex (CCV)*, which is used in the first phase of the approximation algorithm to pull out supersets of clusters. We show that each optimal center from a 2-perturbation resilient subset is a CCV and satisfies a separation property; we prove this by carefully constructing a 2-perturbation in which points from other clusters cannot be too close to the center without causing a contradiction. The structure allows us to modify the approximation algorithm of Vishwanathan [48] to ensure that optimal clusters from perturbation resilient subsets are pulled out separately in the first phase. All of our guarantees hold under the weaker notion of metric perturbation resilience.

Efficient algorithms for symmetric and asymmetric k-center under $(3, \epsilon)$-perturbation resilience. In Section 6, we consider (α, ϵ)-perturbation resilience, which states that at most ϵn total points can swap in or out of each cluster under any α-perturbation. For symmetric k-center, we show that any 2-approximation algorithm will return the optimal clusters from $(3, \epsilon)$-perturbation resilient regions, assuming a mild lower bound on optimal cluster sizes; and for asymmetric k-center, we give an algorithm that outputs a clustering that is ϵ-close to the optimal clustering (see Section 2 for the formal definition). Our main structural tool is showing that if any single point v is close to an optimal cluster other than its own, then $k - 1$ centers achieve the optimal radius 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 the optimal clustering, and we show that 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.

1.2 Related Work

k-center. There are three classic 2-approximation algorithms for k-center from the 1980s [24, 27, 30] that are known to be tight [30]. Asymmetric k-center proved to be a much harder problem. The

first nontrivial result was an $O(\log^* n)$ approximation algorithm [48], and this was later improved to $O(\log^* k)$ [2]. This result was later proven to be asymptotically tight [21].

Perturbation resilience. Perturbation resilience was introduced by Bilu and Linial [15], who showed algorithms that output the optimal solution for Max Cut under $O(n)$-perturbation resilient instances. This was later improved by Bilu et al. [14] to $O(\sqrt{n})$-perturbation resilience and by Makarychev et al. [40] to $O(\sqrt{\log n} \log \log n)$-perturbation resilience). The study of clustering under perturbation resilience was initiated by Awasthi et al. [6], who provided an optimal algorithm for center-based clustering objectives (which include k-median, k-means, and k-center clustering, as well as other objectives) under 3-perturbation resilience. This result was improved by Balcan and Liang [11], who showed an algorithm for center-based clustering under $(1 + \sqrt{2})$-perturbation resilience. They also gave a near-optimal algorithm for k-median under $(2 + \sqrt{3}, \epsilon)$-perturbation resilience when the optimal clusters are not too small.

Recently, Angelidakis et al. [1] gave algorithms for center-based clustering (including k-median, k-means, and k-center) under 2-perturbation resilience and defined the more general notion of metric perturbation resilience, although their algorithm does not extend to the (α, ϵ)-perturbation resilience or local perturbation resilience settings. Cohen-Addad and Schwiegelshohn [22] showed that local search outputs the optimal k-median, k-means, and k-center solution when the data satisfy a stronger variant of 3-perturbation resilience, in which both the optimal clustering and optimal centers are not allowed to change under any 3-perturbation. Perturbation resilience has also been applied to other problems, such as Min Multiway Cut, the Traveling Salesman Problem, finding Nash Equilibria, Metric Labeling, and Facility Location [10, 37, 40–42].

Subsequent work. Vijayaraghavan et al. [47] study k-means under additive perturbation resilience, in which the optimal solution cannot change under additive perturbations to the input distances. The notion of *additive* perturbation resilience is similar but orthogonal to the more common notion of (multiplicative) perturbation resilience. Deshpande et al. [23] gave an algorithm for Euclidean k-means under α-perturbation resilience, which runs in time linear in n and the dimension d, and exponentially in k and $\frac{1}{\alpha-1}$. Chekuri and Gupta [19] showed the natural LP relaxation of k-center and asymmetric k-center is integral for 2-perturbation resilient instances. They also define a new model of perturbation resilience for clustering with outliers, and they show the algorithm of Angelidakis et al. [1] exactly solves clustering with outliers under 2-perturbation resilience; and they further show the natural LP relaxation for k-center with outliers is integral for 2-perturbation resilient instances. Their algorithms have the desirable property that either they output the optimal solution or they guarantee the input did not satisfy 2-perturbation resilience (but note this is not the same thing as determining whether or not a given instance satisfies perturbation resilience). Friggstad et al. [26] show that for any fixed $\epsilon > 0$, $(1 + \epsilon)$-perturbation resilient instances of k-means in doubling metrics can be solved in polynomial time. They also show that in Euclidean space for a non-constant dimension, there exists a fixed $\epsilon_0 > 0$ such that there is no PTAS for $(1 + \epsilon_0)$-perturbation resilient k-means, unless $NP = RP$, assuming a conjecture that they call stable SAT.

Other stability notions. A related notion, approximation stability [8], states that any α-approxima- tion to the objective must be ϵ-close to the target clustering. There are several positive results for k-means, k-median [8, 12, 28], and min-sum [8, 9, 49] under approximation stability. Approximation stability is a stronger notion than perturbation resilience. Formally, approximation stability implies (α, ϵ)-perturbation resilience when the parameters α and ϵ are the same (consequently, when $\epsilon = 0$, approximation stability implies α-perturbation resilience) [8]. Ostrovsky et al. [43] show how to efficiently cluster instances in which the k-means clustering cost is much

lower than the $(k-1)$-means cost. Kumar and Kannan [35] give an efficient clustering algorithm for instances in which the projection of any point onto the line between its cluster center to any other cluster center is a large additive factor closer to its own center than the other center. This result was later improved along multiple axes by Awasthi and Sheffet [7]. These other notions of stability are similar but orthogonal to perturbation resilience in the sense that neither definition implies perturbation resilience or vice versa. There are many other works that show positive results for different natural notions of stability in various settings [3, 5, 28, 29, 35, 36, 44].

2 PRELIMINARIES AND BASIC PROPERTIES

A clustering instance (S, d) consists of a set S of n points, a distance function $d : S \times S \to \mathbb{R}_{\geq 0}$, and an integer k. For a point $u \in S$ and a set $A \subseteq S$, we define $d(A, u) = \min_{v \in A} d(v, u)$. The k-center objective is to find a set of points $X = \{x_1, \ldots, x_k\} \subseteq S$ called *centers* to minimize $\max_{v \in S} d(X, v)$. We denote $\mathrm{Vor}_{X,d}(x) = \{v \in S \mid x = \mathrm{argmin}_{y \in X} d(y, v)\}$,[1] the Voronoi tile of $x \in X$ induced by X on the set of points S in metric d, and we denote $\mathrm{Vor}_{X,d}(X') = \bigcup_{x \in X'} \mathrm{Vor}_X(x)$ for a subset $X' \subseteq X$. We often write $\mathrm{Vor}_X(x)$ and $\mathrm{Vor}_X(X')$ when d is clear from context. We refer to the Voronoi partition induced by X as a clustering (we only consider Voronoi partitions as valid k-center solutions). Throughout the article, we denote the clustering with the minimum cost with respect to d by $\mathcal{OPT} = \{C_1, \ldots, C_k\}$, we denote the radius of $\mathcal{OPT}$ by r^*, and we denote the optimal centers by $c_1, \ldots, c_k$, where c_i is the center of C_i for all $1 \leq i \leq k$. We use $B_r(c)$ to denote a ball of radius r centered at point c.

Some of our results assume distance functions that are metrics, and some of our results assume *asymmetric* distance functions. A distance function d is a *metric* if

(1) for all u, v, $d(u, v) \geq 0$,
(2) for all u, v, $d(u, v) = 0$ if and only if $u = v$,
(3) for all u, v, w, $d(u, w) \leq d(u, v) + d(v, w)$, and
(4) for all u, v, $d(u, v) = d(v, u)$.

An *asymmetric* distance function satisifies *(1)*, *(2)*, and *(3)*, but not *(4)*.

Now, we formally define *perturbation resilience*, a notion introduced by Bilu and Linial [15] for Max Cut and by Awasthi et al. [6] for clustering. We say that d' is an α-perturbation of the distance function d, if for all $u, v \in S$, $d(u, v) \leq d'(u, v) \leq \alpha d(u, v)$.[2]

Definition 2.1 (Perturbation Resilience). A clustering instance (S, d) satisfies *α-perturbation resilience* (α-PR) if for any α-perturbation d' of d, the optimal clustering C' under d' is unique and equal to $\mathcal{OPT}$.

Note that the optimal *centers* might change under an α-perturbation, but the optimal *clustering* must stay the same. This is a well-studied assumption for clustering problems; however, one downside is that it assumes *every point* must stay in its own optimal cluster following a perturbation. This motivates a relaxed variant of α-perturbation resilience, called (α, ϵ)-perturbation resilience, that allows a small change in the optimal clustering when distances are perturbed. We say that two clusterings C and C' are ϵ-close if $\min_\sigma \sum_{i=1}^{k} \left| C_i \setminus C'_{\sigma(i)} \right| \leq \epsilon n$, where σ is a permutation on $[k] = \{1, \ldots, k\}$.

[1] In general, $\mathrm{argmin}_{y \in X} d(y, v)$ might be a set. All of the Voronoi tilings defined in this work (unless otherwise noted) are provably unique due to the perturbation resilience assumptions defined later.
[2] We only consider perturbations in which the distances increase, because without loss of generality, we can scale the distances to simulate decreasing distances.

Definition 2.2 ((α, ϵ)-perturbation Resilience). A clustering instance (S, d) satisfies (α, ϵ)- perturbation resilience if for any α-perturbation d' of d, each optimal clustering C' under d' is ϵ-close to $\mathcal{OPT}$.

In Definitions 2.1 and 2.2, we do not assume that the α-perturbations satisfy the triangle inequality. Angelidakis et al. [1] recently studied the weaker definition in which the α-perturbations must satisfy the triangle inequality, called *metric perturbation resilience*. We can update these definitions accordingly. For symmetric clustering objectives, α-metric perturbations are restricted to metrics. For asymmetric clustering objectives, the α-metric perturbations must satisfy the directed triangle inequality.

Definition 2.3 (Metric Perturbation Resilience). A clustering instance (S, d) satisfies *α-metric perturbation resilience* (α-MPR) if for any α-metric perturbation d' of d the optimal clustering C' under d' is unique and equal to $\mathcal{OPT}$.

In our arguments, we will sometimes convert a non-metric perturbation d' into a metric perturbation by taking the *metric completion* d'' of d' (also referred to as the *shortest-path metric* on d') by setting the distances in d'' as the length of the shortest path on the graph whose edges are the lengths in d'. Note that for all u, v, we have $d(u, v) \leq d''(u, v)$, since d was originally a metric.

2.1 Local Perturbation Resilience

In the previous section, we defined α-perturbation resilience and the more relaxed (α, ϵ)-perturbation resilience. However, even assuming (α, ϵ)-perturbation resilience is strong in the sense that it applies to every cluster, e.g., the entire clustering instance cannot change by more than an ϵ fraction after a perturbation. Now, we define perturbation resilience for an optimal cluster rather than the entire dataset. All prior works have considered perturbation resilience with respect to the entire dataset.

Definition 2.4 (Local Perturbation Resilience). Given a clustering instance (S, d), a cluster C satisfies α-perturbation resilience (α-PR) if for each α-perturbation d' of d each optimal clustering C' under d' contains C.

As a sanity check, we show that a clustering is perturbation resilient if and only if every optimal cluster satisfies perturbation resilience.

Fact 2.5. *A clustering instance (S, d) satisfies α-PR if and only if each optimal cluster satisfies α-PR.*

Proof. Given a clustering instance (S, d), the forward direction follows by definition: Assume (S, d) satisfies α-PR, and given an optimal cluster C_i, then for each α-perturbation d', the optimal clustering stays the same under d'; therefore, C_i is contained in the optimal clustering under d'. Now, we prove the reverse direction. Given a clustering instance with optimal clustering C, and given an α-perturbation d', let the optimal clustering under d' be C'. For each $C_i \in C$, by assumption, C_i satisfies α-PR, so $C_i \in C'$. Therefore, $C = C'$. □

Next, we define the local version of (α, ϵ)-PR. Two clusters C_i and C_j are ϵ-close if $|C_i \setminus C_j| + |C_j \setminus C_i| \leq \epsilon$.

Definition 2.6 (Local (α, ϵ)-perturbation Resilience). Given a clustering instance (S, d), an optimal cluster C satisfies (α, ϵ)-PR if for any α-perturbation d' of d each optimal clustering C' under d' contains a cluster C' that is ϵ-close to C.

In Sections 5 and 6, we will consider a slightly stronger notion of local perturbation resilience. Informally, an optimal cluster satisfies α-strong local perturbation resilience if it is α-PR and all

nearby optimal clusters are also α-PR. We will sometimes be able to prove guarantees for clusters satisfying strong local perturbation resilience that are not true under standard local perturbation resilience.

Definition 2.7 (Strong Local Perturbation Resilience). Given a clustering instance (S, d), a cluster C satisfies α-strong local perturbation resilience (α-SLPR) if for each C' such that there exists $u \in C$, $v \in C'$, and $d(u, v) \leq r^*$, then C' is α-PR (any cluster that is close to C must be α-PR).

Note that a typical use for local perturbation resilience is when we can divide the input point set into two (or more) sections and some sections consist of clusters that satisfy local perturbation resilience while other sections do not. In this case, all clusters in the perturbation resilient section also satisfy strong local perturbation resilience, except potentially the clusters on the border of the section (see Figure 5).

To conclude this section, we state a lemma for asymmetric (and symmetric) k-center, which allows us to reason about a specific class of α-perturbations that will be important throughout the article. We give two versions of the lemma, each of which will be useful in different sections of the article.

Lemma 2.8. *Given a clustering instance (S, d) and $\alpha \geq 1$,*

(1) assume we have 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^*.
(2) assume we have an α-perturbation d' of d with the following property: for all u, v, either $d'(u, v) = \min(\alpha r^*, \alpha d(u, v))$ or $d'(u, v) = \alpha d(u, v)$. Then the optimal cost under d' is αr^*.

Proof.

(1) Assume there exists a set of centers $C' = \{c'_1, \ldots, c'_k\}$ whose k-center cost under d' is strictly less than αr^*. Then for all i and $s \in \text{Vor}_{C', d'}(c'_i)$, $d'(c'_i, s) < \alpha r^*$, implying $d(c'_i, s) < r^*$ by construction. It follows that the k-center cost of C' under d is r^*, which is a contradiction. Therefore, the optimal cost under d' must be αr^*.
(2) Given u, v such that $d(u, v) \geq r^*$, then $d'(u, v) \geq \alpha r^*$ by construction. Now the proof follows from part one. □

3 k-CENTER UNDER PERTURBATION RESILIENCE

In this section, we provide efficient algorithms for finding the optimal clustering for symmetric and asymmetric instances of k-center under 2-perturbation resilience. Our results directly improve on the result by Balcan and Liang [11] for symmetric k-center under $(1 + \sqrt{2})$-perturbation resilience. We also show that it is NP-hard to recover $\mathcal{OPT}$ even for symmetric k-center instance under $(2 - \delta)$-perturbation resilience. As an immediate consequence, our results are tight for both 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.

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. Next, we provide a novel algorithm for asymmetric k-center under 2-perturbation resilience. Finally, we show hardness of k-center under $(2 - \delta)$-PR.

3.1 α-approximations are Optimal under α-PR

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

Theorem 3.1. *Given a clustering instance (S, d) satisfying α-perturbation resilience for asymmetric k-center and a set C of k centers that is an α-approximation, i.e., for all $p \in S$, $d(C, p) \leq \alpha r^*$, 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 2.1, 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 2.8, the optimal cost for d' is αr^*. Also, C achieves cost at most αr^* 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 at most α. Therefore, the closest point in C to p under d' is still $c(p)$. □

3.2 Asymmetric k-center under 2-PR

An immediate consequence of Theorem 3.1 is that we have an exact algorithm for symmetric k-center under 2-perturbation resilience by running a simple 2-approximation algorithm (e.g., References [24, 27, 30]). However, Theorem 3.1 only gives an algorithm for asymmetric k-center under $O(\log^*(k))$-perturbation resilience. Next, we show it is possible to substantially improve the latter result.

One of the challenges involved in dealing with asymmetric k-center instances is the fact that even though for all $p \in C_i$, $d(c_i, p) \leq r^*$, the reverse distance, $d(p, c_i)$, 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 the notion of a *center-capturing vertex* [48].

Definition 3.2. Given an asymmetric k-center clustering instance (S, d), a point $v \in S$ is a *center-capturing vertex* (CCV) if for all $u \in S$, $d(u, v) \leq r^*$ implies $d(v, u) \leq r^*$.

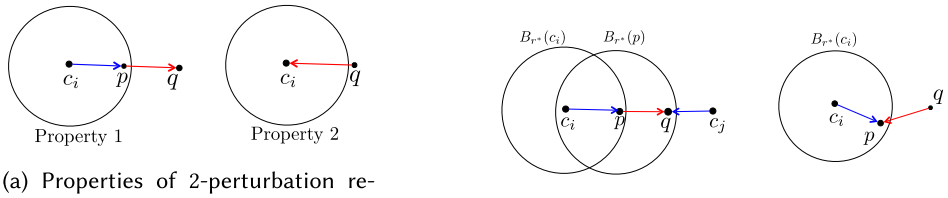
As the name suggests, each CCV $p \in C_i$ "captures" its center in the sense that $d(p, c_i) \leq r^*$. We define the set of center-capturing vertices $A = \{p \mid p \text{ is a CCV}\}$. Intuitively speaking, these points behave similarly to a set of points with symmetric distances up to a distance r^*. To explore this, we define a desirable property of A with respect to the optimal clustering.

Definition 3.3. 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). Satisfying Definition 3.3 implies that if we can optimally cluster A, then we can optimally cluster the entire instance (formalized in Theorem 3.6). 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



(a) Properties of 2-perturbation resilience

(b) Demonstrating the correctness of Algorithm 1

Fig. 1. Properties of a 2-perturbation resilient instance of aymmetric k-center that are used for clustering.

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 [6].

Property (1): For all $p \in C'_i$ and $q \in C'_{j \neq i}$, $d(c_i, p) < d(q, p)$.

The second property requires that any point within distance r^* of a cluster center belongs to that cluster.

Property (2): For all $i \neq j$ and $q \in C_j$, $d(q, c_i) > r^*$ (see Figure 1). A weaker version of this property was introduced by Balcan and Liang [11].

Let us illustrate how these properties allow us to optimally cluster A.[3] 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 in polynomial time (if r^* is not known, then we can use a guess-and-check wrapper). Then, we create the set of all balls of radius r^* around all points in A. 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.

Formal details of our analysis.

Lemma 3.4. *Properties 1 and 2 hold for asymmetric k-center instances satisfying 2-perturbation resilience.*

Proof. *Property 1:* Assume false, $d(q, p) \leq d(c_i, p)$. The idea will be that, since q is in A, it is close to its own center, so we can construct a perturbation in which q replaces its center c_j. Then p will join q's cluster, causing a contradiction. Construct the following d':

$$d'(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = q,\ t \in C_j \cup \{p\}, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

[3]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).

ALGORITHM 1: Asymmetric k-center algorithm under 2-PR

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

Create symmetric set

• Build set $A = \{p \mid \forall q, d(q,p) \leq r^* \implies d(p,q) \leq r^*\}$

Create candidate balls

• $\forall c \in A$, construct $G_c = \{p \in A \mid d(c,p) \leq r^*\}$.

• Define $\mathcal{G} = \{G_c\}_{c \in A}$.

Prune balls

• $\forall G_c$, if $\exists p \in G_c$, $q \in A \setminus G_c$ such that $d(q,p) < d(c,p)$, then set $\mathcal{G} = \mathcal{G} \setminus G_c$.

• $\forall p, q$ such that $G_p \subseteq G_q$, set $\mathcal{G} = \mathcal{G} \setminus G_p$.

Insert remaining points

• $\forall p \notin A$, add p to G_q, where $q = \arg\min_{s \in A} d(s,p)$.

Output: $\mathcal{G}$

This is a 2-perturbation, because for all $q' \in C_j \cup \{p\}$, $d(q,q') \leq 2r^*$. Then, by Lemma 2.8, the optimal cost is $2r^*$. The set of centers $\{c_\ell\}_{i=1}^k \setminus \{c_j\} \cup \{q\}$ achieves the optimal cost, since q is distance $2r^*$ from C_j, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). Then for all c_ℓ, $d'(q,p) \leq d'(c_i,p) \leq d'(c_\ell,p)$. Then, we can construct a 2-perturbation in which q becomes the center of C_j, and then q is the best center for p, so we have a contradiction.

Property 2: Assume on the contrary that there exists $q \in C_j$, $i \neq j$ such that $d(q,c_i) \leq r^*$. Now, we will define a d' in which q can become a center for C_i.

$$d'(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = q,\ t \in C_i, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

This is a 2-perturbation, because for all $p \in C_i$, $d(q,p) \leq 2r^*$. Then, by Lemma 2.8, the optimal cost is $2r^*$. The set of centers $\{c_\ell\}_{i=1}^k \setminus \{c_i\} \cup \{q\}$ achieves the optimal cost, since q is distance $2r^*$ from C_i, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). But the clustering with centers $\{c_\ell\}_{i=1}^k \setminus \{c_i\} \cup \{q\}$ is different from $\mathcal{OPT}$, since (at the very least) q and c_i are in different clusters. This contradicts 2-perturbation resilience. □

Lemma 3.5. *The set A respects the structure of $\mathcal{OPT}$.*

Proof. From Lemma 3.4, we can use Property 2 in our analysis. First, we show that $c_i \in A$ for all $i \in [k]$. Given c_i, $\forall p \in C_i$, then $d(c_i,p) \leq r^*$ by definition of $\mathcal{OPT}$. $\forall q \notin C_i$, then by Property 2, $d(q,c_i) > r^*$. It follows that for any point $p \in S$, it cannot be the case that $d(p,c_i) \leq r^*$ and $d(c_i,p) > r^*$. Therefore, $c_i \in A$.

Now, we show that for all $p \in S \setminus A$, if $A(p) \in C_i$, then $p \in C_i$. Given $p \in S \setminus A$, let $p \in C_i$ and assume towards contradiction that $q = A(p) \in C_j$ for some $i \neq j$. We will construct a 2-perturbation d' in which q replaces c_j as the center for C_j and p switches from C_i to C_j, causing a contradiction. We construct d' as follows: All distances are increased by a factor of 2 except for $d(q,p)$ and $d(q,q')$ for all $q' \in C_j$. These distances are increased by a factor of 2 up to $2r^*$. Formally,

$$d'(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = q,\ t \in C_j \cup \{p\}, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

This is a 2-perturbation, because $d(q,C_j) \leq 2r^*$. Then, by Lemma 2.8, the optimal cost is $2r^*$. The set of centers $\{c_\ell\}_{i=1}^k \setminus \{c_j\} \cup \{q\}$ achieves the optimal cost, since q is distance $2r^*$ from C_j, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). But consider the point p. Since all centers are in A and q is the closest point to p in A, then q is the center for p under d'. Therefore, the optimal clustering under d' is different from $\mathcal{OPT}$, so we have a contradiction. □

Now, we are ready to show Algorithm 1 returns the optimal clustering.

Theorem 3.6. *Algorithm 1 returns the exact solution for asymmetric k-center under 2-perturbation resilience.*

Proof. In this proof, we refer to the first line of **Prune balls** in Algorithm 1 as **Pruning step 1** and the second line as **Pruning step 2**. First, we must show that after **Pruning step 2**, the remaining sets are exactly $C_i' = C_i \cap A$ for all $i \in [k]$. We prove this in three steps: The sets G_{c_i} correspond to C_i', these sets are not thrown out in **Pruning step 1** and **Pruning step 2**, and all other sets are thrown out in steps **Pruning step 1** and **Pruning step 2**. Because of Lemma 3.4, we can use Properties 1 and 2.

From Lemma 3.5, 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 $q \notin C_i'$, by Property 2, we have $d(q, c_i) > r^*$, and, since c_i and q are in A, we have $d(c_i, q) > r^*$ as well. It follows that $G_{c_i} = C_i'$.

Given $s \in G_{c_i}$ and $t \in A \setminus G_{c_i}$, we have $s \in C_i'$ and $t \in C_j'$ for some $j \neq i$. If $d(t, s) < d(c_i, s)$, then we get a contradiction from Property 1. Therefore, for all i, G_{c_i} is not thrown out in step **Pruning step 1**.

Recall we showed that $G_{c_i} = C_i'$ for all i. If $G_p \subseteq G_{c_i}$, then G_p will be thrown out in **Pruning step 2** (if $G_p = G_{c_i}$, it does not matter which set we keep, so without loss of generality, say that we keep G_{c_i}). If G_p is not thrown out in **Pruning step 2**, then there must exist $s \in G_p \cap C_j'$ for some $j \neq i$. 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 **Pruning step 1**. We conclude that for all non-centers p, G_p is thrown out in **Pruning step 1** or **Pruning step 2**. Thus, the remaining sets after **Pruning step 2** are exactly $C_1', \ldots, C_k'$.

Finally, by Lemma 3.5, 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 of *k*-center under Perturbation Resilience

In this section, we show NP-hardness for k-center under $(2 - \delta)$-perturbation resilience. We show that if there exists a polynomial time algorithm that returns the optimal solution for symmetric k-center under $(2 - \delta)$-perturbation resilience,[4] then $NP = RP$ even under the condition that the optimal clusters all have size at least $\frac{n}{2k}$. Because symmetric k-center is a special case of asymmetric k-center, we have the same hardness results for asymmetric k-center. This proves Theorem 3.6 is tight with respect to the level of perturbation resilience assumed.

Theorem 3.7. *There is no polynomial time algorithm for finding the optimal k-center clustering under $(2 - \delta)$-perturbation resilience, even when assuming all optimal clusters have size at least $\frac{n}{2k}$, unless $NP = RP$.*

We show a reduction from a special case of Dominating Set that we call Unambiguous-Balanced-Perfect Dominating Set. Below, we formally define this problem and all intermediate problems. Part of our reduction is based on the proof of Ben-David and Reyzin [13], who showed a reduction from a variant of dominating set to the weaker problem of clustering under $(2 - \delta)$-center proximity. 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. We use four NP-hard problems in a chain of reductions. Here, we define all of these problems up front. We introduce the "balanced" variants of two existing problems.

[4] In fact, our result holds even under the strictly stronger notion of *approximation stability* [8].

Definition 3.8 (3-Dimensional Matching (3DM) [33]). We are given three disjoint sets X_1, X_2, and X_3 each of size m, and a set T such that $t \in T$ is a triple $t = (x_1, x_2, x_3)$ where $x_1 \in X_1$, $x_2 \in X_2$, and $x_3 \in X_3$. The problem is to find a set $M \subseteq T$ of size m that exactly hits all the elements in $X_1 \cup X_2 \cup X_3$. In other words, for all pairs $(x_1, x_2, x_3), (y_1, y_2, y_3) \in M$, it is the case that $x_1 \neq y_1$, $x_2 \neq y_2$, and $x_3 \neq y_3$.

Definition 3.9 (Balanced-3-Dimensional Matching (B3DM)). This is the 3DM problem (X_1, X_2, X_3, T) with the additional constraint that $2m \leq |T| \leq 3m$, where $|X_1| = |X_2| = |X_3| = m$.

Definition 3.10 (Perfect Dominating Set (PDS) [13]). Given a graph $G = (V, E)$ and an integer k, the problem is to find a set of vertices $D \subseteq V$ of size k such that for all $v \in V \setminus D$, there exists *exactly one* $d \in D$ such that $(d, v) \in E$ (then, we say d "hits" v).

Definition 3.11 (Balanced-Perfect-Dominating Set (BPDS)). This is the PDS problem (G, k) with the additional assumption that if the graph has n vertices and a dominating set of size k exists, then each vertex in the dominating set hits at least $\frac{n}{2k}$ vertices.

Additionally, each problem has an "Unambiguous" variant, which is the added constraint that the problem has at most one solution. Valiant and Vazirani [46] showed that Unambiguous-3SAT is hard unless $NP = RP$. To show the Unambiguous version of another problem is hard, one must establish a parsimonious polynomial time reduction from Unambiguous-3SAT to that problem. A parsimonious reduction is one that preserves the number of solutions. For two problems A and B, we denote $A \leq_{par} B$ to mean there is a reduction from A to B that is parsimonious *and* polynomial time. Some common reductions involve 1-to-1 mappings and are therefore trivially parsimonious, but many other common reductions are not parsimonious. For instance, the standard reduction from 3SAT to 3DM is not parsimonious [34], yet there is a more roundabout series of reductions that are all parsimonious. To prove Theorem 3.7, we start with the claim that Unambiguous-BPDS is hard unless $NP = RP$. We use a parsimonious series of reductions from 3SAT to B3DM to BPDS. All of these reductions are from prior work, yet we verify parsimony and balancedness.

Lemma 3.12. *There is no polynomial time algorithm for Unambiguous-BPDS unless $NP = RP$.*

Proof. We use a series of parsimonious reductions from 3SAT to B3DM to BPDS. Then it follows from the result by Valiant and Vazirani [46] that there is no polynomial time algorithm for Unambiguous-BPDS unless $NP = RP$.

To show that B3DM is NP-hard, we use the reduction of Dyer and Frieze [25], who showed that Planar-3DM is NP-hard. While planarity is not important for the purpose of our problems, their reduction from 3SAT has two other nice properties that we crucially use. First, the reduction is parsimonious, as pointed out by Hunt III et al. [31]. Second, given their 3DM instance X_1, X_2, X_3, T, each element in $X_1 \cup X_2 \cup X_3$ appears in either two or three tuples in T. (Dyer and Frieze [25] mention this observation just before their Theorem 2.3.) From this, it follows that $2m \leq |T| \leq 3m$, and so their reduction proves that B3DM is NP-hard via a parsimonious reduction from 3SAT.

Next, we reduce B3DM to BPDS using a reduction similar to the reduction by Ben-David and Reyzin [13]. Their reduction maps every element in $X_1 \cup X_2 \cup X_3 \cup T$ to a vertex in V and adds one extra vertex v to V. There is an edge from each element $(x_1, x_2, x_3) \in T$ to the corresponding elements $x_1 \in X_1$, $x_2 \in X_2$, and $x_3 \in X_3$. Furthermore, there is an edge from v to every element in T. Ben-David and Reyzin [13] show that if the 3DM instance is a YES instance with matching $M \subseteq T$, then the minimum dominating set is $v \cup M$. Now, we will verify this same reduction can be used to reduce B3DM to BPDS. If we start with B3DM, then our graph has $|X_1| + |X_2| + |X_3| + |T| + 1 \leq 6m + 1$ vertices, since $|T| \leq 3m$, so $n \leq 6m + 1$. Also note that in the YES instance, the dominating set is size $m + 1$ by construction. Therefore, to verify the reduction to BPDS, we must show that

in the YES instance, each node in the dominating set hits $\geq \frac{6m+1}{2(m+1)}$ nodes. Given $t \in M$, t hits 3 nodes in the graph and $\frac{n}{2(m+1)} \leq \frac{6m+1}{2m+2} \leq 3$. The final node in the dominating set is v, and v hits $|T| - m \geq 2m - m = m$ nodes, and $\frac{6m+1}{2(m+1)} \leq m$ when $m \geq 3$. Therefore, the resulting instance is a BPDS instance.

Now, we have verified that there exists a parsimonious reduction 3SAT $\leq_{par}$ BPDS, so it follows that there is no polynomial time algorithm for Unambiguous-BPDS unless $NP = RP$. □

Now, we can prove Theorem 3.7 by giving a reduction from Unambiguous-BPDS to k-center clustering under $(2-\delta)$-perturbation resilience, where all clusters are size $\geq \frac{n}{2k}$. We use the same reduction as Ben-David and Reyzin [13], but we must verify that the resulting instance is $(2-\delta)$-perturbation resilient. Note that our reduction requires the Unambiguous variant while the reduction of Ben-David and Reyzin [13] does not, since we are reducing to a stronger problem.

Proof of Theorem 3.7. From Lemma 3.12, there is no polynomial time algorithm for Unambiguous BPDS unless $NP = RP$. Now for all $\delta > 0$, we reduce from Unambiguous-BPDS to k-center clustering and show the resulting instance has all cluster sizes $\geq \frac{n}{2k}$ and satisfies $(2-\delta)$-perturbation resilience.

Given an instance of Unambiguous-BPDS, for every $v \in V$, create a point $v \in S$ in the clustering instance. For every edge $(u, v) \in E$, let $d(u, v) = 1$, otherwise let $d(u, v) = 2$. Since all distances are either 1 or 2, the triangle inequality is trivially satisfied. Then a k-center solution of cost 1 exists if and only if there exists a dominating set of size k.

Since each vertex in the dominating set hits at least $\frac{n}{2k}$ vertices, the resulting clusters will be size at least $\frac{n}{2k} + 1$. Additionally, if there exists a dominating set of size k, then the corresponding optimal k-center clustering has cost 1. Because this dominating set is perfect and unique, any other clustering has cost 2. It follows that the k-center instance is $(2-\delta)$-perturbation resilient. □

4 k-CENTER UNDER METRIC PERTURBATION RESILIENCE

In this section, we extend the results from Section 3 to the metric perturbation resilience setting [1]. We first give a generalization of Lemma 2.8 to show that it can be extended to metric perturbation resilience. Then, we show how this immediately leads to corollaries of Theorem 3.1 and Theorem 3.6 extended to the metric perturbation resilience setting.

Recall that in the proofs from the previous section, we created α-perturbations d' by increasing all distances by α, except a few distances $d(u, v) \leq \alpha r^*$ that we increased to $\min(\alpha d(u, v), \alpha r^*)$. In this specific type of α-perturbation, we used the crucial property that the optimal clustering has cost αr^* (Lemma 2.8). However, d' may be highly non-metric, so our challenge is arguing that the proof still goes through after taking the metric completion of d' (recall the metric completion of d' is defined as the shortest path metric on d'). In the following lemma, we show that Lemma 2.8 remains true after taking the metric completion of the perturbation.

Lemma 4.1. *Given $\alpha \geq 1$ and an asymmetric k-center clustering instance (S, d) with optimal radius r^*, let d'' denote an α-perturbation such that for all u, v, either $d''(u, v) = \min(\alpha r^*, \alpha d(u, v))$ or $d''(u, v) = \alpha d(u, v)$. Let d' denote the metric completion of d''. Then d' is an α-metric perturbation of d, and the optimal cost under d' is αr^*.*

Proof. By construction, $d'(u, v) \leq d''(u, v) \leq \alpha d(u, v)$. Since d satisfies the triangle inequality, we have that $d(u, v) \leq d'(u, v)$, so d' is a valid α-metric perturbation of d.

Now given u, v such that $d(u, v) \geq r^*$, we will prove that $d'(u, v) \geq \alpha r^*$. By construction, $d''(u, v) \geq \alpha r^*$. Then, since d' is the metric completion of d'', there exists a path $u = u_0 - u_1 - \cdots - u_{s-1} - u_s = v$ such that $d'(u, v) = \sum_{i=0}^{s-1} d'(u_i, u_{i+1})$ and for all $0 \leq i \leq s-1$, $d'(u_i, u_{i+1}) = d''(u_i, u_{i+1})$.

Case 1: there exists an i such that $d''(u_i, u_{i+1}) \geq \alpha r^*$. Then $d'(u, v) \geq \alpha r^*$ and we are done.

Case 2: for all $0 \leq i \leq s-1$, $d''(u_i, u_{i+1}) < \alpha r^*$. Then, by construction, $d'(u_i, u_{i+1}) = d''(u_i, u_{i+1}) = \alpha d(u_i, u_{i+1})$, and so $d'(u, v) = \sum_{i=0}^{s-1} d'(u_i, u_{i+1}) = \alpha \sum_{i=0}^{s-1} d(u_i, u_{i+1}) \geq \alpha d(u, v) \geq \alpha r^*$.

We have proven that for all u, v, if $d(u, v) \geq r^*$, then $d'(u, v) \geq \alpha r^*$. Then, by Lemma 2.8, the optimal cost under d' must be αr^*. □

Recall that metric perturbation resilience states that the optimal solution does not change under any metric perturbation to the input distances. In the proofs of Theorems 3.1 and 3.6, the only perturbations constructed were the type as in Lemma 2.8. Since Lemma 4.1 shows that even the metric closures of these perturbations still have cost at most αr^*, Theorems 3.1 and 3.6 are true even under metric perturbation resilience.

Corollary 4.2. *Given a clustering instance (S, d) satisfying α-metric perturbation resilience for asymmetric k-center and a set C of k centers that is an α-approximation, i.e., $\forall p \in S$, $\exists c \in C$ such that $d(c, p) \leq \alpha r^*$, then the Voronoi partition induced by C is the optimal clustering.*

Corollary 4.3. *Algorithm 1 returns the exact solution for asymmetric k-center under 2-metric perturbation resilience.*

5 k-CENTER UNDER LOCAL PERTURBATION RESILIENCE

In this section, we further extend the results from Sections 3 and 4 to the local perturbation resilience setting. First, we show that any α-approximation to k-center will return each optimal α-MPR cluster, i.e., Corollary 4.2 holds even in the local perturbation resilience setting. Then for asymmetric k-center, we show that a natural modification to the $O(\log^* n)$ approximation algorithm of Vishwanathan [48] leads to an algorithm that maintains its performance in the worst case while exactly returning each optimal cluster located within a 2-MPR region of the dataset. This generalizes Corollary 4.3.

5.1 Symmetric k-center

In Section 3, we showed that any α-approximation algorithm for k-center returns the optimal solution for instances satisfying α-perturbation resilience (and this was generalized to metric perturbation resilience in the previous section). In this section, we extend this result to the local perturbation resilience setting. We show that any α-approximation will return each (local) α-MPR cluster. For example, if a clustering instance is half 2-perturbation resilient, then running a 2-approximation algorithm will return the optimal clusters for half the dataset and a 2-approximation for the other half.

Theorem 5.1. *Given an asymmetric k-center clustering instance (S, d), a set C of k centers that is an α-approximation, and a clustering C defined as the Voronoi partition induced by C, then each α-MPR cluster is contained in C.*

The proof is very similar to the proof of Theorem 3.1. The key difference is that we reason about each perturbation resilient cluster individually rather than reasoning about the global structure of perturbation resilience.

Proof of Theorem 5.1. Given an α-approximate solution C to a clustering instance (S, d), and given an α-MPR cluster C_i, we will create an α-perturbation as follows: Define $C(v) := \text{argmin}_{c \in C} d(c, v)$. For all $v \in S$, set $d''(v, C(v)) = \min\{\alpha r^*, \alpha d(v, C(v))\}$. For all other points $u \in S$, set $d''(v, u) = \alpha d(v, u)$. Then, by Lemma 4.1, the metric completion d' of d'' is an α-perturbation of d with optimal cost αr^*. By construction, the cost of C is $\leq \alpha r^*$ under d'; therefore, C is an

optimal clustering. Denote the set of centers of C by C. By definition of α-MPR, there exists $v_i \in C$ such that $\text{Vor}_{C,d'}(v_i) = C_i$. Now, given $v \in C_i$, $\text{argmin}_{u \in C} d'(u, v) = v_i$, so, by construction, $\text{argmin}_{u \in C} d(u, v) = v_i$. Therefore, $\text{Vor}_{C,d}(v_i) = C_i$, so $C_i \in \mathcal{C}$. □

5.2 Asymmetric k-center

In Section 3, we gave an algorithm that outputs the optimal clustering for asymmetric k-center under 2-perturbation resilience (Algorithm 1 and Theorem 3.6), and we extended it to metric perturbation resilience in Section 4. In this section, we extend the result further to the local perturbation resilience setting, and we show how to add a worst-case guarantee of $O(\log^* n)$. Specifically, we give a new algorithm, which is a natural modification to the $O(\log^* n)$ approximation algorithm of Vishwanathan [48], and show that it maintains the $O(\log^* n)$ guarantee in the worst case and furthermore, for each perturbation resilient cluster C_i, there is a cluster outputted by the algorithm that is a superset of C_i and does not contain any other perturbation resilient cluster. As a consequence, if the entire clustering instance satisfies 2-metric perturbation resilience, then the output of our algorithm is the optimal clustering.

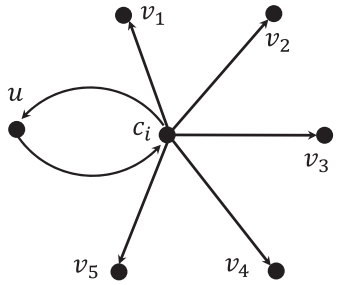
Theorem 5.2. *Given an asymmetric k-center clustering instance (S, d) of size n with optimal clustering $\{C_1, \ldots, C_k\}$, for each 2-MPR cluster C_i, there exists a cluster outputted by Algorithm 3 that is a superset of C_i and does not contain any other 2-MPR cluster. Furthermore, the overall clustering returned by Algorithm 3 is an $O(\log^* n)$-approximation.*

At the end of this section, we will also show an algorithm that outputs an optimal cluster C_i exactly, if C_i and any optimal cluster near C_i are 2-MPR.

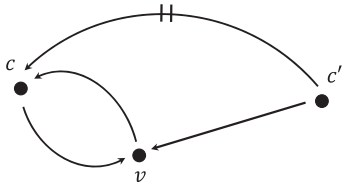
Approximation algorithm for asymmetric k-center. We start with a recap of the $O(\log^* n)$ approximation algorithm by Vishwanathan [48]. This was the first nontrivial algorithm for asymmetric k-center, and the approximation ratio was later proven to be tight by Reference [21]. To explain the algorithm, it is convenient to think of asymmetric k-center as a set covering problem. Given an asymmetric k-center instance (S, d), define the directed graph $D_{(S,d)} = (S, A)$, where $A = \{(u, v) \mid d(u, v) \leq r^*\}$. For a point $v \in S$, we define $\Gamma_{\text{in}}(v)$ and $\Gamma_{\text{out}}(v)$ as the set of vertices with an arc to and from v, respectively, in $D_{(S,d)}$. The asymmetric k-center problem is equivalent to finding a subset $C \subseteq S$ of size k such that $\cup_{c \in C} \Gamma_{\text{out}}(c) = S$. We also define $\Gamma^x_{\text{in}}(v)$ and $\Gamma^x_{\text{out}}(v)$ as the set of vertices that have a path of length $\leq x$ to and from v in $D_{(S,d)}$, respectively, and we define $\Gamma^x_{\text{out}}(A) = \bigcup_{v \in A} \Gamma^x_{\text{out}}(v)$ for a set $A \subseteq S$ and similarly for $\Gamma^x_{\text{in}}(A)$. It is standard to assume the value of r^* is known; since it is one of $O(n^2)$ distances, the algorithm can search for the correct value in polynomial time.

Recall the notion of a center-capturing vertex from Section 3.2: A point v is a CCV if for all $u \in S$, $d(u, v) \leq r^*$ implies $d(v, u) \leq r^*$. In other words, v is a CCV if $\Gamma_{\text{in}}(v) \subseteq \Gamma_{\text{out}}(v)$. Therefore, v's entire cluster is contained inside $\Gamma^2_{\text{out}}(v)$, which is a nice property that the approximation algorithm exploits (see Figure 2(a)). At a high level, the approximation algorithm has two phases. In the first phase, the algorithm iteratively picks a CCV v arbitrarily and removes all points in $\Gamma^2_{\text{out}}(v)$. This continues until there are no more CCVs. For every CCV picked, the algorithm is guaranteed to remove an entire optimal cluster. In the second phase, the algorithm runs $\log^* n$ rounds of a greedy set-cover subroutine on the remaining points (see Algorithm 2). To prove the second phase terminates in $O(\log^* n)$ rounds, the analysis crucially assumes there are no CCVs among the remaining points. We refer the reader to Reference [48] for these details.

Description of our algorithm and analysis. We show a modification to the approximation algorithm of Vishwanathan [48] leads to simultaneous guarantees in the worst case and under local perturbation resilience. Note that the set of all CCVs is identical to the symmetric set A defined



(a) u is a CCV, so it is distance $2r^*$ to its entire cluster.



(b) c satisfies CCV-proximity, so it is closer to v than c' is to v.

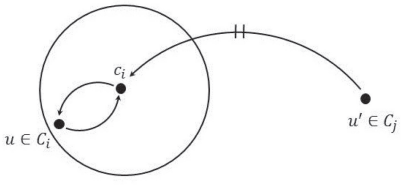
Fig. 2. Examples of a center-capturing vertex (left), and CCV-proximity (right).

ALGORITHM 2: $O(\log^* n)$ approximation algorithm for asymmetric k-center [48]

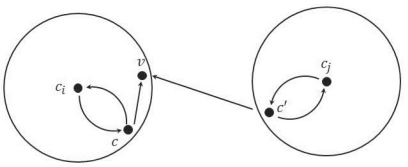
Input: Asymmetric k-center instance (S, d), optimal radius r^* (or try all possible candidates)
Set $C = \emptyset$
Phase I: Remove arbitrary CCVs
While there exists an unmarked CCV
• Pick an unmarked CCV c, add c to C, and mark all vertices in $\Gamma^2_{\text{out}}(c)$
Phase II: Recursive set cover
Set $A_0 = S \setminus \Gamma^5_{\text{out}}(C)$, $i = 0$.
While $|A_i| > k$:
• Set $A'_{i+1} = \emptyset$.
• While there exists an unmarked point in A_i:
—Pick $v \in S$, which maximizes $\Gamma^5_{\text{out}}(v) \cap A_i$, mark points in $\Gamma^5_{\text{out}}(v) \cap A_i$, and add v to A'_{i+1}.
• Set $A_{i+1} = A'_{i+1} \cap A_0$ and $i = i + 1$
Output: Centers $C \cup A_{i+1}$

in Section 3.2. In Section 3.2, we showed that all centers are in A; therefore, all centers are CCVs, assuming 2-PR. In this section, we have that the center of a 2-MPR cluster is a CCV (Lemma 5.4), which is true by definition of r^*, $(C_i \subseteq \Gamma_{\text{out}}(c_i))$ and by using the definition of 2-MPR $(\Gamma_{\text{in}}(c_i) \subseteq C_i)$. Since each 2-MPR center is a CCV, we might hope that we can output the 2-MPR clusters by iteratively choosing a CCV v and removing all points in $\Gamma^2_{\text{out}}(v)$. However, using this approach, we might remove two or more centers from 2-MPR clusters in the same iteration, which means we would not output one separate cluster for each 2-MPR cluster. If we try to get around this problem by iteratively choosing a CCV v and removing all points in $\Gamma^1_{\text{out}}(v)$, then we may not remove one full cluster in each iteration; so, for example, some of the 2-MPR clusters may be cut in half.

The key challenge is thus carefully specifying which nearby points get marked by each CCV c chosen by the algorithm. We fix this problem with two modifications that carefully balance the two guarantees. First, any CCV c chosen will mark points in the following way: For all $c' \in \Gamma_{\text{in}}(c)$, mark all points in $\Gamma_{\text{out}}(c')$. Intuitively, we still mark points that are two hops from c, but the first hop must go backwards, i.e., mark v such that there exists c' and $d(c', c) \leq r^*$ and $d(c', v) \leq r^*$. This gives us a useful property: If the algorithm picks a CCV $c \in C_i$ and it marks a different 2-MPR center c_j, then the middle hop must be a point q in C_j. However, we know from perturbation resilience that $d(c_j, q) < d(c, q)$. This fact motivates the final modification to the algorithm. Instead of picking



(a) c_i satisfies center-separation, so $\Gamma_{\text{in}}(c_i) \subseteq C_i$.



(b) $c \in C_i$ satisfies weak CCV-proximity, so it is closer to $v \in C_i$ than $c' \in C_j$ is to v.

Fig. 3. Examples of center-separation (left) and weak CCV-proximity (right).

arbitrary CCVs, we require the algorithm to choose CCVs with an extra structural property that we call *CCV-proximity* (Definition 5.3) (see Figure 2(b)). Intuitively, a point c satisfying CCV-proximity must be closer than other CCVs to each point in $\Gamma_{\text{in}}(c)$. Going back to our previous example, c will NOT satisfy CCV-proximity, because c_j is closer to q, but we will be able to show that all 2-MPR centers do satisfy CCV-proximity. Thus, Algorithm 3 works as follows: It first chooses points satisfying CCV-proximity and marks points according to the rule mentioned earlier. When there are no more points satisfying CCV-proximity, the algorithm chooses regular CCVs. Finally, it runs Phase II as in Algorithm 2. This ensures that Algorithm 3 will output each 2-MPR center in its own cluster.

Details for Theorem 5.2. Now, we formally define CCV-proximity. The other properties in the following definition, *center-separation* and *weak CCV-proximity*, are defined in terms of the optimal clustering, so they cannot be explicitly used by an algorithm, but they will simplify all of our proofs.

Definition 5.3.

(1) An optimal center c_i satisfies *center-separation* if any point within distance r^* of c_i belongs to its cluster C_i. That is, $\Gamma_{\text{in}}(c_i) \subseteq C_i$ (see Figure 3(a)).[5]
(2) A CCV $c \in C_i$ satisfies *weak CCV-proximity* if, given a CCV $c' \notin C_i$ and a point $v \in C_i$, we have $d(c, v) < d(c', v)$ (see Figure 3(b)).[6]
(3) A point c satisfies *CCV-proximity* if it is a CCV, and each point in $\Gamma_{\text{in}}(c)$ is closer to c than any CCV outside of $\Gamma_{\text{out}}(c)$. That is, for all points $v \in \Gamma_{\text{in}}(c)$ and CCVs $c' \notin \Gamma_{\text{out}}(c)$, $d(c, v) < d(c', v)$ (see Figure 2(b)).[7]

Next, we prove that all 2-MPR centers satisfy center-separation, and all CCVs from a 2-MPR cluster satisfy CCV-proximity and weak CCV-proximity.

Lemma 5.4. *Given an asymmetric k-center clustering instance (S, d) and a 2-MPR cluster C_i,*
(1) *c_i satisfies center-separation,*
(2) *any CCV $c \in C_i$ satisfies CCV-proximity,*
(3) *any CCV $c \in C_i$ satisfies weak CCV-proximity.*

Proof. Given an instance (S, d) and a 2-MPR cluster C_i, we show that C_i has the desired properties.

[5]Center-separation is the local-PR equivalent of property 2 from Section 3.2.
[6]This is a variant of α-center proximity [6], a property defined over an entire clustering instance, which states for all i, for all $v \in C_i$, $j \neq i$, we have $\alpha d(c_i, v) < d(c_j, v)$. Our variant generalizes to local-PR, asymmetric instances, and general CCVs.
[7]This is similar to a property in Reference [11].

Center separation: Assume there exists a point $v \in C_j$ for $j \neq i$ such that $d(v, c_i) \leq r^*$. The idea is to construct a 2-perturbation in which v becomes the center for C_i.

$$d''(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = v, t \in C_i, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

d'' is a valid 2-perturbation of d, because for each point $u \in C_i$, $d(v,u) \leq d(v,c_i) + d(c_i,u) \leq 2r^*$. Define d' as the metric completion of d''. Then, by Lemma 4.1, d' is a 2-metric perturbation with optimal cost $2r^*$. The set of centers $\{c_{i'}\}_{i'=1}^k \setminus \{c_i\} \cup \{v\}$ achieves the optimal cost, since v is distance $2r^*$ from C_i, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). If v is a noncenter, then $\{c_{i'}\}_{i'=1}^k \setminus \{c_i\} \cup \{v\}$ is a valid set of k centers. If $v = c_j$, then add an arbitrary point $v' \in C_j$ to this set of centers (it still achieves the optimal cost, since adding another center can only decrease the cost). Then in this new optimal clustering, c_i's center is a point in $\{c_{i'}\}_{i'=1}^k \setminus \{c_i\} \cup \{v, v'\}$, none of which are from C_i. We conclude that C_i is no longer an optimal cluster, contradicting 2-MPR.

Weak CCV-proximity: Given a CCV $c \in C_i$, a CCV $c' \in C_j$ such that $j \neq i$, and a point $v \in C_i$, assume to the contrary that $d(c', v) \leq d(c, v)$. We will construct a perturbation in which c and c' become centers of their respective clusters, and then v switches clusters. Define the following perturbation d'':

$$d''(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = c, t \in C_i \text{ or } s = c', t \in C_j \cup \{v\}, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

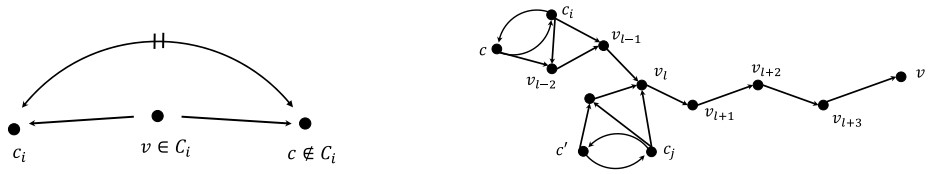
d'' is a valid 2-perturbation of d, because for each point $u \in C_i$, $d(c,u) \leq d(c,c_i) + d(c_i,u) \leq 2r^*$, for each point $u \in C_j$, $d(c',u) \leq d(c',c_j) + d(c_j,u) \leq 2r^*$, and $d(c',v) \leq d(c,v) \leq d(c,c_i) + d(c_i,v) \leq 2r^*$. Define d' as the metric completion of d''. Then, by Lemma 4.1, d' is a 2-metric perturbation with optimal cost $2r^*$. The set of centers $\{c_{i'}\}_{i'=1}^k \setminus \{c_i, c_j\} \cup \{c, c'\}$ achieves the optimal cost, since c and c' are distance $2r^*$ from C_i and C_j, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). Then, since $d'(c', v) \leq d(c, v)$, v can switch clusters, contradicting perturbation resilience.

CCV-proximity: First, we show that c_i is a CCV. By center-separation, we have that $\Gamma_{\text{in}}(c_i) \subseteq C_i$, and by definition of r^*, we have that $C_i \subseteq \Gamma_{\text{out}}(c_i)$. Therefore, $\Gamma_{\text{in}}(c_i) \subseteq C_i \subseteq \Gamma_{\text{out}}(c_i)$, so c_i is a CCV. Now given a point $v \in \Gamma_{\text{in}}(c_i)$ and a CCV $c \notin \Gamma_{\text{out}}(c_i)$, from center-separation and definition of r^*, we have $v \in C_i$ and $c \in C_j$ for $j \neq i$. Then, from weak CCV-proximity, $d(c_i, v) < d(c, v)$. □

Now using Lemma 5.4, we can prove Theorem 5.2.

Proof of Theorem 5.2. First, we explain why Algorithm 3 retains the approximation guarantee of Algorithm 2. Given any CCV $c \in C_i$ chosen in Phase I, since c is a CCV, then $c_i \in \Gamma_{\text{out}}(c)$, and by definition of r^*, $C_i \subseteq \Gamma_{\text{out}}(c_i)$. Therefore, each chosen CCV always marks its cluster, and we start Phase II with no remaining CCVs. This condition is sufficient for Phase II to return an $O(\log^* n)$ approximation (Theorem 3.1 from Vishwanathan [48]).

Next, we claim that for each 2-MPR cluster C_i, there exists a cluster outputted by Algorithm 3 that is a superset of C_i and does not contain any other 2-MPR cluster. To prove this claim, we first show there exists a point from C_i satisfying CCV-proximity that cannot be marked by any point from a different cluster in Phase I. From Lemma 5.4, c_i satisfies CCV-proximity and center-separation. If a point $c \notin C_i$ marks c_i, then there exists $v \in \Gamma_{\text{in}}(c) \cap \Gamma_{\text{in}}(c_i)$. By center-separation, $c_i \notin \Gamma_{\text{out}}(c)$, and therefore, since c is a CCV, $c \notin \Gamma_{\text{out}}(c_i)$. But then, from the definition of CCV-proximity for c_i and c, we have $d(c, v) < d(c_i, v)$ and $d(c_i, v) < d(c, v)$, so we have reached a contradiction (see Figure 4(a)).



(a) A point $c \notin C_i$ cannot mark c_i without causing a contradiction.

(b) Case 2 in the proof of Theorem 5.5: c' is closer to v than c is to v.

Fig. 4. Example of Algorithm 3 (left) and the proof of Theorem 5.5 (right).

ALGORITHM 3: Algorithm for asymmetric k-center under perturbation resilience

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

Set $C = \emptyset$.

Phase I: Remove special CCVs

- While there exists an unmarked CCV:

  —Pick an unmarked point c that satisfies CCV-proximity. If no such c exists, then pick an arbitrary unmarked CCV instead. Add c to C, and $\forall c' \in \Gamma_{\text{in}}(c)$, mark $\Gamma_{\text{out}}(c')$.
- For each $c \in C$, let V_c denote c's Voronoi tile of the marked points induced by C.

Phase II: Recursive set cover

- Run Phase II as in Algorithm 2, outputting A_{i+1}.
- Compute the Voronoi diagram $\{V_c'\}_{c \in C \cup A_{i+1}}$ of $S \setminus \Gamma^5_{\text{out}}(C)$ induced by $C \cup A_{i+1}$
- For each c in C, set $V_c' = V_c \cup V_c'$

Output: Sets $\{V_c'\}_{c \in C \cup A_{i+1}}$

At this point, we know a point $c \in C_i$ will always be chosen by the algorithm in Phase I. To finish the proof, we show that each point v from C_i is closer to c than to any other point $c' \notin C_i$ chosen as a center in Phase I. Since c and c' are both CCVs, this follows directly from weak CCV-proximity.[8] □

5.2.1 Strong local perturbation resilience. Theorem 5.2 shows that Algorithm 3 will output each 2-PR center in its own cluster. Given some 2-PR center c_i, it is unavoidable that c_i might mark a *non* 2-PR center c_j and capture all points in its cluster. In this section, we show that Algorithm 3 with a slight modification outputs each 2-strong local perturbation resilient cluster exactly. Recall that intuitively an optimal cluster C_i satisfies α-strong local perturbation resilience if all nearby clusters satisfy α-perturbation resilience (Definition 2.7).

Intuitively, the nearby 2-PR clusters "shield" C_i from all other points (see Figure 5). The only modification is that at the end of Phase II, instead of calculating the Voronoi diagram using the metric d, we assign each point $v \in S \setminus \Gamma^5_{\text{out}}(C)$ to the point in $C \cup A_{i+1}$, which minimizes the path length in $D_{(S,d)}$, breaking ties by distance to first common vertex in the shortest path.

Theorem 5.5. *Given an asymmetric k-center clustering instance (S, d) with optimal clustering $C = \{C_1, \ldots, C_k\}$, consider a 2-PR cluster C_i. Assume that for all C_j for which there is $v \in C_j$, $u \in C_i$, and $d(u, v) \leq r^*$, we have that C_j is also 2-PR (C_i satisfies 2-strong local perturbation resilience). Then Algorithm 4 returns C_i exactly.*

[8] It is possible that a center c' chosen in Phase 2 may be closer to v than c is to v, causing c' to "steal" v; this is unavoidable. This is why Algorithm 3 separately computes the Voronoi tiling from Phase I and Phase II, and so the final output is technically not a valid Voronoi tiling over the entire instance S.

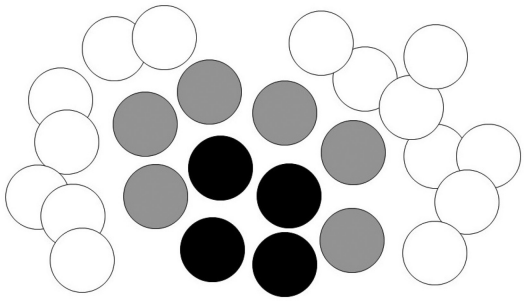


Fig. 5. The white clusters are optimal clusters with no structure, the gray clusters are 2-PR clusters, and the black clusters are 2-PR clusters that only have neighbors that are also 2-PR (Theorem 5.5). Algorithm 4 outputs the black clusters exactly.

ALGORITHM 4: Outputting optimal clusters for asymmetric k-center under stability

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

Set $C = \emptyset$.

Phase I: Remove special CCVs

- While there exists an unmarked CCV:
  - —Pick an unmarked point c that satisfies CCV-proximity. If no such c exists, then pick an arbitrary unmarked CCV instead. Add c to C, and $\forall c' \in \Gamma_{\text{in}}(c)$, mark $\Gamma_{\text{out}}(c')$.
- For each $c \in C$, let V_c denote c's Voronoi tile of the marked points induced by C.

Phase II: Recursive set cover

- Run Phase II as in Algorithm 2, outputting A_{i+1}.

Phase III: Assign points to centers

- For each $v \in S \setminus \Gamma^5_{\text{out}}(C)$, assign v to the center $c \in C \cup A_{i+1}$ with the minimum path length in $D_{(S,d)}$ from c to v, breaking ties by distance to first common vertex in the shortest path.
- Let V'_c denote the set of vertices in $v \in S \setminus \Gamma^5_{\text{out}}(C)$ assigned to c.
- For each c in C, set $V'_c = V_c \cup V'_c$

Output: Sets $\{V'_c\}_{c \in C \cup A_{i+1}}$

Proof. Given a 2-PR cluster C_i with the property in the theorem statement, by Theorem 5.2, there exists a CCV $c \in C_i$ from Phase I satisfying CCV-proximity such that $C_i \subseteq V_c$. Our goal is to show that $V_c = C_i$. First, we show that $\Gamma_{\text{in}}(c) \subseteq C_i$, which will help us prove the theorem. Assume towards contradiction that there exists a point $v \in \Gamma_{\text{in}}(c) \setminus C_i$. Let $v \in C_j$. Since c is a CCV, we have $v \in \Gamma_{\text{out}}(c)$, so C_j must be 2-PR by definition. By Lemma 5.4, c_j is a CCV and $d(c_j, v) < d(c, v)$. But this violates CCV-proximity of c, so we have reached a contradiction. Therefore, $\Gamma_{\text{in}}(c) \subseteq C_i$.

To finish the proof, we must show that $V_c \subseteq C_i$. Assume towards contradiction there exists $v \in V_c \setminus C_i$ at the end of the algorithm.

Case 1: v was marked by c in Phase I. Let $v \in C_j$. Then there exists a point $u \in \Gamma_{\text{in}}(c)$ such that $v \in \Gamma_{\text{out}}(u)$. From the previous paragraph, we have that $\Gamma_{\text{in}}(c) \subseteq C_i$, so $u \in C_i$. Therefore, $v \in \Gamma_{\text{out}}(u)$ implies C_j must be 2-PR. Since v is from a different 2-PR cluster, it cannot be contained in V_c, so we have reached a contradiction.

Case 2: v was not marked by c in Phase I. Denote the shortest path in $D_{(S,d)}$ from c to v by $(c = v_0, v_1, \ldots, v_{L-1}, v_L = v)$. Let $v_\ell \in C_j$ denote the first vertex on the shortest path that is not in C_i (such a vertex must exist, because $v \notin C_i$). Then $v_{\ell-1} \in C_i$ and $d(v_{\ell-1}, v_\ell) \leq r^*$, so C_j is 2-PR by

the assumption in Theorem 5.5. Let c' denote the CCV chosen in Phase I such that $C_j \subseteq V_{c'}$. Then, by weak CCV-proximity from Lemma 5.4, we have $d(c', v_\ell) < d(c, v_\ell)$.

Case 2a: $d(c, v_\ell) \leq r^*$. Then v_ℓ is the first vertex on the shortest path from c to v and c' to v, so v_ℓ is the first common vertex on the shortest paths. Since $d(c', v_\ell) < d(c, v_\ell)$, the algorithm will choose c' as the center for v (see Figure 4(b)).

Case 2b: $d(c, v_\ell) > 2r^*$. But, since $c' \in C_j$ is a CCV, we have $d(c', v_\ell) \leq d(c', c_j) + d(c_j, v_\ell) \leq 2r^*$, so the shortest path from c' to v_ℓ on $D_{(S,d)}$ is at most 2, and the shortest path from c to v_ℓ on $D_{(S,d)}$ is at least 3. Since v_ℓ is on the shortest path from c to v, it follows that the shortest path from c' to v is strictly shorter than the shortest path from c to v.

Case 2c: $r^* < d(c, v_\ell) \leq 2r^*$. In this case, we will show that $d(c', v_\ell) \leq r^*$, and therefore, we conclude the shortest path from c' to v is strictly shorter than the shortest path from c to v, as in Case 2b. Assume towards contradiction that $d(c', v_\ell) > r^*$. Then, we will create a 2-perturbation in which c and c' become centers for their own clusters and v_ℓ switches clusters. Define the following perturbation d':

$$d'(s,t) = \begin{cases} \min(2r^*, 2d(s,t)) & \text{if } s = c,\ t \in C_i \text{ or } s = c',\ t \in C_j \setminus \{v_\ell\}, \\ d(s,t) & \text{if } s = c,\ t = v_\ell, \\ 2d(s,t) & \text{otherwise.} \end{cases}$$

d' is a valid 2-perturbation of d, because for each point $u \in C_i$, $d(c,u) \leq d(c,c_i) + d(c_i,u) \leq 2r^*$, for each point $u \in C_j$, $d(c',u) \leq d(c',c_j) + d(c_j,u) \leq 2r^*$ and $d(c, v_\ell) \leq 2r^*$. Therefore, d' does not decrease any distances (and, by construction, d' does not increase any distance by more than a factor of 2). If the optimal cost is $2r^*$, then the set of centers $\{c_{i'}\}_{i'=1}^{k} \setminus \{c_i, c_j\} \cup \{c, c'\}$ achieves the optimal cost, since c and c' are distance $2r^*$ from all points in $C_i \cup \{v_\ell\}$ and C_j, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $2r^*$). Then, by perturbation resilience, it must be the case that $d'(c', v_\ell) < d'(c, v_\ell)$, which implies $2d(c', v_\ell) < d(c, v_\ell)$. But $d(c', v_\ell) > r^*$ and $d(c, v_\ell) \leq 2r^*$, so, we have a contradiction. Now, we assume the optimal cost of d' is less than $2r^*$. Note that all distances $d(s,t)$ were increased to $2d(s,t)$ or $\min(2d(s,t), 2r^*)$ except for $d(c, v_\ell)$. Therefore, c must be a center for v_ℓ under d', or else the optimal cost would be exactly $2r^*$ by Lemma 4.1. But it contradicts perturbation resilience to have c and c_ℓ in the same optimal cluster under a 2-perturbation. This completes the proof. □

6 k-CENTER UNDER (α, ϵ)-PERTURBATION RESILIENCE

In this section, we consider (α, ϵ)-perturbation resilience. First, we show that any 2-approximation algorithm for symmetric k-center must be optimal under $(3, \epsilon)$-perturbation resilience (Theorem 6.1). Next, we show how to extend this result to local perturbation resilience (Theorem 6.8). Then, we give an algorithm for asymmetric k-center, which returns a clustering that is ϵ-close to $\mathcal{OPT}$ under $(3, \epsilon)$-perturbation resilience (Theorem 6.17). For all 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 $\epsilon > 0$ (Theorem 6.18). The theorems in this section require a careful reasoning about sets of centers under different perturbations that cannot all simultaneously be optimal.

6.1 Symmetric k-center

We show that for any $(3, \epsilon)$-perturbation resilient k-center instance such that $|C_i| > 2\epsilon n$ for all $i \in [k]$, there cannot be any pair of points from different clusters that are distance $\leq r^*$. This structural result implies that simple algorithms will return the optimal clustering, such as running any 2-approximation algorithm or running the Single Linkage algorithm, which is a fast algorithm widely used in practice for its simplicity.

Theorem 6.1. *Given a $(3, \epsilon)$-perturbation resilient symmetric k-center instance (S, d) where all optimal clusters are size $> \max(2\epsilon n, 3)$, then the optimal clusters in $\mathcal{OPT}$ are exactly the connected components of the threshold graph $G_{r^*} = (S, E)$, where $E = \{(u, v) \mid d(u, v) \leq r^*\}$.*

First, we explain the high-level idea behind the proof.

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 greater than r^* apart from each other. Assume on the contrary that there exist $p \in C_i$ and $q \in C_{j \neq i}$ 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.

Here is how we find a set of $k + 2$ points and a perturbation d' such that all size k subsets are optimal centers under d'. 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 the distances from p to $C_i \cup C_j$, then replacing c_i and c_j with p would still give us a set of $k - 1$ centers that achieve the optimal cost. But, *would this contradict $(3, \epsilon)$-perturbation resilience?* Not necessarily! Perturbation resilience requires exactly k *distinct* centers.[9] 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 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 $\mathcal{OPT}$.

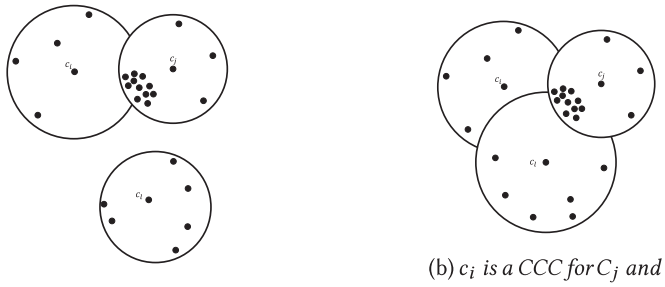
To deal with the chain reactions, we crucially introduce the notion of a *cluster capturing center* (CCC). A cluster capturing center (CCC) is not to be confused with a center-capturing vertex (CCV), defined by Vishwanathan [48] and used in the previous section. c_x is a CCC for C_y if for all but ϵn 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 it is much harder to reason about what happens to the dummy centers under d', since there may be chain reactions. 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 (respectively, 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.

Formal analysis. We start out with a simple implication from the assumption that $|C_i| > 2\epsilon n$ for all i.

Fact 6.2. *Given a clustering instance that is (α, ϵ)-perturbation resilient for $\alpha \geq 1$, and all optimal clusters have size $> 2\epsilon n$, then for any α-perturbation d', for any set of optimal centers $c'_1, \ldots, c'_k$ of*

[9] 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.



(a) c_i *is a CCC for* C_j. (b) c_i *is a CCC for* C_j *and* c_x *is a CCC2 for* C_j.

Fig. 6. (a) Definition of a CCC and (b) definition of a CCC2.

d', for each optimal cluster C_i, there must be a unique center c_i' that is the center for more than half of the points in C_i under d'.

This fact follows simply from the definition of (α, ϵ)-perturbation resilience (under d', at most ϵn points in the optimal solution can change clusters) and the assumption that all optimal clusters are size $> 2\epsilon n$. Now, we formally define a CCC.

Definition 6.3. A center c_i is a *first-order cluster-capturing center* (CCC) for C_j if for all $x \neq j$, for more than half of the points $p \in C_j$, $d(c_i, p) < d(c_x, p)$ and $d(c_i, p) \leq r^*$ (see Figure 6 (a)). c_i is a *second-order cluster-capturing center* (CCC2) for C_j if there exists a c_l such that for all $x \neq j, l$, for more than half of points $p \in C_j$, $d(c_i, p) < d(c_x, p)$ and $d(c_i, p) \leq r^*$ (see Figure 6(b)).

Each cluster C_j can have at most one CCC c_i, because c_i is closer than any other center to more than half of C_j. Every CCC is a CCC2, since the former is a stronger condition. However, it is possible for a cluster to have multiple CCC2's.[10] We needed to define CCC2 for the following reason: Assuming there exist $p \in C_i$ and $q \in C_j$ that are close, and we replace c_i and c_j with p in the set of centers. It is possible that c_j is a CCC for C_i, but this does not help us, since we want to analyze the set of centers after removing c_j. However, if we know that c_x is a CCC2 for C_i (it is the best center for C_i, disregarding c_j), then we know that c_x will be the best center for C_i after replacing c_i and c_j with p. Now, we use this definition to show that if two points from different clusters are close, then we can 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'. To formalize this notion, we give one more definition.

Definition 6.4. A set $C \subseteq S$ (β, γ)-hits S if for all $s \in S$ there exist β points in C at distance $\leq \gamma r^*$ to s.

Note that if a set C of $k + 2$ points $(3, 3)$-hits S, then any size k subset of C is still $3r^*$ from every point in S, and later, we will show that means there exists a perturbation d' such that every size k subset must be an optimal set of centers.

Lemma 6.5. *Given a clustering instance satisfying $(3, \epsilon)$-perturbation resilience such that all optimal clusters are size $> 2\epsilon n$ and there are two points from different clusters that are $\leq r^*$ apart from each other, then there exists a set $C \subseteq S$ of size $k + 2$ that $(3, 3)$-hits S.*

Proof. First, we prove the lemma assuming that a CCC2 exists, and then we prove the other case. When a CCC2 exists, we do not need the assumption that two points from different clusters are close.

[10] In fact, a cluster can have at most three CCC2's, but we do not use this in our analysis.

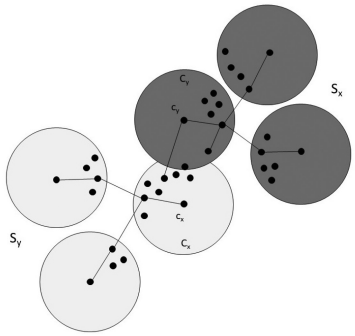


Fig. 7. Case 1 of Lemma 6.5.

Case 1: There exists a CCC2. If there exists a CCC, then denote c_x as a CCC for C_y. If there does not exist a CCC, then denote c_x as a CCC2 for C_y. We will show that all points are close to either C_x or C_y. c_x is distance $\leq r^*$ to all but ϵn points in C_y. Therefore, $d(c_x, c_y) \leq 2r^*$ and so c_x is distance $\leq 3r^*$ to all points in C_y. Consider the following d'.

$$d'(s,t) = \begin{cases} \min(3r^*, 3d(s,t)) & \text{if } s = c_x,\ t \in C_y, \\ 3d(s,t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation, because $d(c_x, t) \leq 3r^*$ for all $t \in C_y$. Then, by Lemma 2.8, the optimal cost is $3r^*$. Given any $u \in S$, the set of centers $\{c_l\}_{i=1}^k \setminus \{c_y\} \cup \{u\}$ achieves the optimal cost, since c_x is distance $3r^*$ from C_y, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $3r^*$). Therefore, this set of centers must create a partition that is ϵ-close to $\mathcal{OPT}$, or else there would be a contradiction. Then, from Fact 6.2, one of the centers in $\{c_l\}_{i=1}^k \setminus \{c_y\} \cup \{u\}$ must be the center for the majority of points in C_y under d'. If this center is c_ℓ, $\ell \neq x, y$, then for the majority of points $v \in C_y$, $d(c_\ell, v) \leq r^*$ and $d(c_\ell, v) < d(c_z, v)$ for all $z \neq \ell, y$. Then, by definition, c_ℓ is a CCC for C_y. But then ℓ must equal x, so we have a contradiction. Note that if some c_ℓ has for the majority of $v \in C_y$, $d(c_\ell, v) \leq d(c_z, v)$ (non-strict inequality) for all $z \neq \ell, y$, then there is another equally good partition in which c_ℓ is not the center for the majority of points in C_y, so we still obtain a contradiction. Therefore, either u or c_x must be the center for the majority of points in C_y under d'.

If c_x is the center for the majority of points in C_y, then u must be the center for the majority of points in C_x (it cannot be a different center c_ℓ, since c_x is a better center for C_x than c_ℓ by definition). Therefore, each $u \in S$ is distance $\leq r^*$ to all but ϵn points in either C_x or C_y.

Now partition all the non-centers into two sets S_x and S_y, such that

$$S_x = \{u' \mid \text{for the majority of points } v' \in C_x,\ d(u', v') \leq r^*\}, \text{ and}$$

$$S_y = \{u' \mid u' \notin S_x \text{ and for the majority of points } v' \in C_y,\ d(u', v') \leq r^*\}.$$

Then given $u', v' \in S_x$, there exists an $s \in C_x$ such that $d(u', v') \leq d(u', s) + d(s, v') \leq 2r^*$ (since both points are close to more than half of points in C_x). Similarly, any two points $u', v' \in S_y$ are $\leq 2r^*$ apart (see Figure 7).

Now, we will find a set of $k + 2$ points that $(3, 3)$-hits S. For now, assume that S_x and S_y are both nonempty. Given an arbitrary pair $p \in S_x$, $q \in S_y$, we claim that $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hits S. Given a non-center $s \in C_i$ such that $i \neq x$ and $i \neq y$, without loss of generality, let $s \in S_x$. Then c_i, p, and c_x are all distance $3r^*$ to s. Furthermore, c_i, p, and c_x are all distance $3r^*$ to c_i. Given a point $s \in C_x$, then c_x, c_y, and p are distance $3r^*$ to s, because $d(c_x, c_y) \leq 2r^*$, and a similar argument holds for $s \in C_y$. Therefore, $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hits S.

If $S_x = \emptyset$ or $S_y = \emptyset$, then we can prove a slightly stronger statement: For each pair of non-centers $\{p, q\}$, $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hits S. Without loss of generality, let $S_y = \emptyset$. Given a point $s \in C_i$ such that $i \neq x$ and $i \neq y$, then c_i, c_x, and p are all distance $3r^*$ to s. Given a point $s \in C_x$, then p, c_x, and c_y are all distance $\leq 3r^*$ to s. Given a point $s \in C_y$, then p, c_x, and c_y are all distance $\leq 3r^*$ to s, because $s, p \in S_x$ implies $d(s, p) \leq 2r^*$. Thus, we have proven case 1.

Case 2: There does not exist a CCC2. Now, we use the assumption that there exist $p \in C_x$, $q \in C_y$, $x \neq y$, such that $d(p, q) \leq r^*$. Then, by the triangle inequality, p is distance $\leq 3r^*$ to all points in C_x and C_y. Consider the following d':

$$d'(s,t) = \begin{cases} \min(3r^*, 3d(s,t)) & \text{if } s = p,\ t \in C_x \cup C_y, \\ 3d(s,t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation, because $d(p, t) \leq 3r^*$ for all $t \in C_x \cup C_y$. Then, by Lemma 2.8, the optimal cost is $3r^*$. Given any $s \in S$, the set of centers $\{c_l\}_{i=1}^k \setminus \{c_x, c_y\} \cup \{p, s\}$ achieves the optimal cost, since p is distance $3r^*$ from $C_x \cup C_y$, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $3r^*$). Therefore, this set of centers must create a partition that is ϵ-close to $\mathcal{OPT}$, or else there would be a contradiction. Then, from Fact 6.2, one of the centers in $\{c_l\}_{i=1}^k \setminus \{c_x, c_y\} \cup \{p, s\}$ must be the center for the majority of points in C_x under d'.

If this center is c_ℓ for $\ell \neq x$ and $\ell \neq y$, then for the majority of points $t \in C_x$, $d(c_\ell, t) \leq r^*$ and $d(c_\ell, t) < d(c_z, t)$ for all $z \neq \ell, x, y$. Then, by definition, c_ℓ is a CCC2 for C_x, and we have a contradiction. Note that if some c_ℓ has for the majority of $t \in C_x$, $d(c_\ell, t) \leq d(c_z, t)$ (non-strict inequality) for all $z \neq \ell, y$, then there is another equally good partition in which c_ℓ is not the center for the majority of points in C_y, so we still obtain a contradiction.

Similar logic applies to the center for the majority of points in C_y. Therefore, p and s must be the centers for C_x and C_y. Since s was an arbitrary noncenter, all noncenters are distance $\leq r^*$ to all but ϵn points in either C_x or C_y.

Similar to Case 1, we now partition all the non-centers into two sets S_x and S_y, such that

$$S_x = \{u \mid \text{for the majority of points } v \in C_x,\ d(u, v) \leq r^*\}, \text{ and}$$

$$S_y = \{u \mid u \notin S_x \text{ and for the majority of points } v \in C_y,\ d(u, v) \leq r^*\}.$$

As before, each pair of points in S_x are distance $\leq 2r^*$ apart and similarly for S_y. It is no longer true that $d(c_x, c_y) \leq 2r^*$, however, we can prove that for both S_x and S_y, there exist points from two distinct clusters each. From the previous paragraph, given a non-center $s \in C_i$ for $i \neq x, y$, we know that p and s are centers for C_x and C_y. With an identical argument, given $t \in C_j$ for $j \neq x, y, i$, we can show that q and t are centers for C_x and C_y. It follows that S_x and S_y both contain points from at least two distinct clusters.

Now, we finish the proof by showing that for each pair $u \in S_x$, $v \in S_y$, $\{c_\ell\}_{\ell=1}^k \cup \{u, v\}$ $(3, 3)$-hits S. Given a non-center $s \in C_i$, without loss of generality, $s \in S_x$, then there exists $j \neq i$ and $t \in C_j \cap S_x$. Then c_i, c_j, and u are $3r^*$ to s and c_i, c_x, and u are $3r^*$ to c_i. In the case where $i = x$, then c_i, c_j, and u are $3r^*$ to c_i. This concludes the proof. □

So far, we have shown that by just assuming two points from different clusters are close, we can find a set of $k + 2$ points that $(3, 3)$-hits S. Now, we will show that such a set leads to a contradiction under $(3, \epsilon)$-perturbation resilience. Specifically, we will show there exists a perturbation d' such that any size k subset can be an optimal set of centers. But it is not possible that all $\binom{k+2}{k}$ of these sets of centers simultaneously create partitions that are ϵ-close to $\mathcal{OPT}$. First, we state a lemma that proves there does exist a perturbation d' such that any size k subset is an optimal set of centers.

Lemma 6.6. *Given a k-center clustering instance (S, d), given $z \geq 0$, and given a set $C \subseteq S$ of size $k + z$ that $(z + 1, \alpha)$-hits S, there exists an α-metric perturbation d' such that all size k subsets of C are optimal sets of centers under d'.*

Proof. Consider the following perturbation d'':

$$d''(s,t) = \begin{cases} \min(\alpha r^*, \alpha d(s,t)) & \text{if } s \in C \text{ and } d(s,t) \leq \alpha r^*, \\ \alpha d(s,t) & \text{otherwise.} \end{cases}$$

By Lemma 4.1, the metric closure d' of d'' is an α-metric perturbation with optimal cost αr^*. Given any size k subset $C' \subseteq C$, then for all $v \in S$, there is still at least one $c \in C'$ such that $d(c, v) \leq \alpha r^*$; therefore, by construction, $d'(c, v) \leq \alpha r^*$. It follows that C' is a set of optimal centers under d'. □

Next, we state a fact that helps clusters rank their best centers from the set of $k + 2$ points. For each cluster C_i, we would like to have a ranking of all points such that for a given d' and set of k centers, the center for C_i is the highest point in the ranking. The following fact shows this ranking is well defined:

Fact 6.7. *Given a k-center clustering instance (S, d) with optimal clustering $C = \{C_1, \ldots, C_k\}$ such that for all $i \in [k]$, $|C_i| > 2\epsilon n$, let d' denote an α-perturbation of d. There are rankings $R_{x,d'}$ for all $x \in [k]$ such that for any optimal set of centers $c'_1, \ldots, c'_k$ under d', the center that is closest in d' to all but ϵn points in C_x is the highest-ranked point in $R_{x,d'}$.*[11]

Proof. Assume the fact is false. Then there exists a d', a cluster C_i, two points p and q, and two sets of k centers $p, q \in C$ and $p, q \in C'$ that achieve the optimal cost under d', but p is the center for C_i in C while q is the center for C_i in C'. Then p is closer than all other points in C to all but ϵn points in C_i. Similarly, q is closer than all other points in C' to all but ϵn points in C_i. Since $|C_i| > 2\epsilon n$, this causes a contradiction. □

We also define $R_{x,d',C} : C \to [n']$ as the ranking specific to a set of centers C, where $|C| = n'$. Now, we can prove Theorem 6.1.

Proof of Theorem 6.1. It suffices to prove that any two points from different clusters are at distance $> r^*$ from each other. Assume towards contradiction that this is not true. Then, by Lemma 6.5, there exists a set C of size $k + 2$ that $(3, 3)$-hits S. From Lemma 6.6, there exists a 3-metric perturbation d' such that all size k subsets of C are optimal sets of centers under d'. Consider the ranking of each cluster for C over d' guaranteed from Fact 6.7. We will show this ranking leads to a contradiction.

Consider the set of all points ranked 1 or 2 by any cluster, formally, $\{p \in C \mid \exists i \text{ s.t. } R_{i,d',C} \leq 2\}$. This set is a subset of C, since we are only considering the rankings of points in C, so it is size $\leq k + 2$. Note that a point cannot be ranked both 1 and 2 by a cluster. Then as long as $k > 2$, it follows by the Pigeonhole Principle that there exists a point $c \in C$ that is ranked in the top two by two different clusters. Formally, there exists x and y such that $x \neq y$, $R_{x,d',C}(c) \leq 2$ and $R_{y,d',C}(c) \leq 2$. Denote u and v such that $R_{x,d',C}(u) = 1$ and $R_{y,d',C}(v) = 1$. If u or v is equal to c, then redefine it to an arbitrary center in $C \setminus \{c, u, v\}$. Consider the set of centers $C' = C \setminus \{u, v\}$, which is optimal under d' by construction. But then, from Fact 6.7, c is the center for all but ϵn points in both C_x and C_y, contradicting Fact 6.2. This completes the proof. □

[11] Formally, for each C_x, there exists a bijection $R_{x,d'} : S \to [n]$ such that for all sets of k centers C that achieve the optimal cost under d', we have $c = \text{argmin}_{c' \in C} R_{x,d'}(c')$ if and only if $\text{Vor}_C(c)$ is ϵ-close to C_x.

6.2 Local Perturbation Resilience

Now, we extend the argument from the previous section to local perturbation resilience. First, we state our main structural result, which is that any pair of points from different $(3, \epsilon)$-PR clusters must be distance $> r^*$ from each other. Then, we will show how the structural result easily leads to an algorithm for $(3, \epsilon)$-SLPR clusters.

THEOREM 6.8. *Given a k-center clustering instance (S, d) with optimal radius r^* such that all optimal clusters are size $> 2\epsilon n$ and there are at least three $(3, \epsilon)$-PR clusters, then for each pair of $(3, \epsilon)$-PR clusters C_i and C_j, for all $u \in C_i$ and $v \in C_j$, we have $d(u, v) > r^*$.*

Before we prove this theorem, we show how it implies an algorithm to output the optimal $(3, \epsilon)$-SLPR clusters exactly. Since the distance from each point to its closest center is $\leq r^*$, a corollary of Theorem 6.8 is that any 2-approximate solution must contain the optimal $(3, \epsilon)$-SLPR clusters, as long as the 2-approximation satisfies two sensible conditions: *(1)* for every point v and its assigned center u (so we know $d(u, v) \leq 2r^*$), $\exists w$ such that $d(u, w)$ and $d(w, v)$ are $\leq r^*$ and *(2)* there cannot be multiple clusters outputted in the 2-approximation that can be combined into one cluster with radius smaller than r^*. Both of these properties are easily satisfied using quick pre- or post-processing steps.[12]

THEOREM 6.9. *Given a k-center clustering instance (S, d) such that all optimal clusters are size $> 2\epsilon n$ and there are at least three $(3, \epsilon)$-PR clusters, then any 2-approximate solution satisfying conditions* (1) *and* (2) *must contain all optimal $(3, \epsilon)$-SLPR clusters.*

PROOF. Given such a clustering instance, then Theorem 6.8 ensures that there is no edge of length r^* between points from two different $(3, \epsilon)$-PR clusters. Given a $(3, \epsilon)$-SLPR cluster C_i, it follows that there is no point $v \notin C_i$ such that $d(v, C_i) \leq r^*$. Therefore, given a 2-approximate solution $\mathcal{C}$ satisfying condition *(1)*, any $u \in C_i$ and $v \notin C_i$ cannot be in the same cluster. This is because in the graph of datapoints where edges signify a distance $\leq r^*$, C_i is an isolated component. Finally, by condition *(2)*, C_i must not be split into two clusters. Therefore, $C_i \in \mathcal{C}$. □

Proof idea for Theorem 6.8. The high-level idea of this proof is similar to the proof of Theorem 6.1. In fact, the first half is very similar to Lemma 6.5: We show that if two points from different PR clusters are close together, then there must exist a set of $k + 2$ points C that $(3, 3)$-hits the entire point set. In the previous section, we arrived at a contradiction by showing that it is not possible that all $\binom{k+2}{k}0$ subsets of C can be centers that are ϵ-close to $\mathcal{OPT}$. However, the weaker local PR assumption poses a new challenge.

As in the previous section, we will still argue that all size k subsets of C cannot stay consistent with the $(3, \epsilon)$-PR clusters using a ranking argument that maps optimal clusters to optimal centers, but our argument will be to establish conditional claims that narrow down the possible sets of ranking lists. For instance, assume there is a $(3, \epsilon)$-PR cluster C_i that ranks c_i first and ranks c_j second. Then under subsets C', which do not contain c_i, c_j is the center for a cluster C_i', which is ϵ-close to C_i. Therefore, a different point in C' must be the center for all but ϵn points in C_j (and it cannot be a different center c_ℓ without causing a contradiction). This is the basis for Lemma 6.13, which is the main workhorse lemma in the proof of Theorem 6.8. By building up conditional statements, we are able to analyze every possibility of the ranking lists for the three $(3, \epsilon)$-PR clusters and show that all of them lead to contradictions, proving Theorem 6.8.

[12] For condition *(1)*, before running the algorithm, remove all edges of distance $> r^*$, and then take the metric completion of the resulting graph. For condition *(2)*, given the radius $\hat{r}$ of the outputted solution, for each $v \in S$, check if the ball of radius $\hat{r}$ around v captures multiple clusters. If so, then combine them.

Formal analysis of Theorem 6.8. We start with a local perturbation resilience variant of Fact 6.2.

Fact 6.10. *Given a k-center clustering instance (S, d) such that all optimal clusters have size $> 2\epsilon n$, let d' denote an α-perturbation with optimal centers $C' = \{c'_1, \ldots, c'_k\}$. Let $\mathcal{C}'$ denote the set of (α, ϵ)-PR clusters. Then there exists a one-to-one function $f : C' \to C'$ such that for all $C_i \in \mathcal{C}'$, $|Vor_{C,d'}(f(C_i)) \cap C_i| \geq |C_i| - \epsilon n$. That is, the optimal cluster in d' whose center is $f(C_i)$ contains all but ϵn of the points in C_i.*

In words, for any set of optimal centers under an α-perturbation, each PR cluster can be paired to a unique center. This follows simply because all optimal clusters are size $> 2\epsilon n$, yet under a perturbation, $< \epsilon n$ points can switch out of each PR cluster. Because of this fact, for a perturbation d' with set of optimal centers C and an (α, ϵ)-PR cluster C_x, we will say that c is the center for C_x under d' if c is the center for all but ϵn points in C_x. Now, we are ready to prove the first half of Theorem 6.8, stated in the following lemma. The proof is similar to Lemma 6.5.

Lemma 6.11. *Given a k-center clustering instance (S, d) such that all optimal clusters are size $> 2\epsilon n$ and there exist two points at distance r^* from different $(3, \epsilon)$-PR clusters, then there exists a partition $S_x \cup S_y$ of the non-centers $S \setminus \{c_\ell\}_{\ell=1}^k$ such that for all pairs $p \in S_x$, $q \in S_y$, $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hits S.*

Proof. This proof is split into two main cases. The first case is the following: There exists a CCC2 for a $(3, \epsilon)$-PR cluster, disregarding a $(3, \epsilon)$-PR cluster. In fact, in this case, we do not need the assumption that two points from different PR clusters are close. If there exists a CCC to a $(3, \epsilon)$-PR cluster, then denote the CCC by c_x and the cluster by C_y. Otherwise, let c_x denote a CCC2 to a $(3, \epsilon)$-PR cluster C_y, disregarding a $(3, \epsilon)$-PR center c_z. Then c_x is at distance $\leq r^*$ to all but ϵn points in C_y. Therefore, $d(c_x, c_y) \leq 2r^*$ and so c_x is at distance $\leq 3r^*$ to all points in C_y. Consider the following perturbation d'':

$$d''(s,t) = \begin{cases} \min(3r^*, 3d(s,t)) & \text{if } s = c_x,\ t \in C_y, \\ 3d(s,t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation, because for all $v \in C_y$, $d(c_x, v) \leq 3r^*$. Define d' as the metric completion of d''. Then, by Lemma 4.1, d' is a 3-metric perturbation with optimal cost $3r^*$. Given any non-center $v \in S$, the set of centers $\{c_\ell\}_{\ell=1}^k \setminus \{c_y\} \cup \{v\}$ achieves the optimal score, since c_x is at distance $3r^*$ from C_y, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $3r^*$). Therefore, from Fact 6.10, one of the centers in $\{c_\ell\}_{\ell=1}^k \setminus \{c_y\} \cup \{v\}$ must be the center for all but ϵn points in C_y under d'. If this center is c_ℓ, $\ell \neq x, y$, then for all but ϵn points $u \in C_y$, $d(c_\ell, u) \leq r^*$, and $d(c_\ell, u) < d(c_z, u)$ for all $z \neq \ell, y$. Then, by definition, c_ℓ is a CCC for the $(3, \epsilon)$-PR cluster, C_y. But then, by construction, ℓ must equal x, so we have a contradiction. Note that if some c_ℓ has for all but ϵn points $u \in C_y$, $d(c_\ell, u) \leq d(c_z, u)$ (non-strict inequality) for all $z \neq \ell, y$, then there is another equally good partition in which c_ℓ is not the center for all but ϵn points in C_y, so we still obtain a contradiction. Therefore, either v or c_x must be the center for all but ϵn points in C_y under d'.

If c_x is the center for all but ϵn points in C_y, then, because C_y is $(3, \epsilon)$-PR, the corresponding cluster must contain fewer than ϵn points from C_x. Furthermore, since for all $\ell \neq x$ and $u \in C_x$, $d(u, c_x) < d(u, c_\ell)$, it follows that v must be the center for all but ϵn points in C_x. Therefore, every non-center $v \in S$ is at distance $\leq r^*$ to all but ϵn points in either C_x or C_y.

Now partition all the non-centers into two sets S_x and S_y, such that

$$S_x = \{p \mid \text{for the majority of points } q \in C_x,\ d(p,q) \leq r^*\}$$

and

$$S_y = \{p \mid p \notin S_x \text{ and for the majority of points } q \in C_y,\ d(p,q) \leq r^*\}.$$

Then given $p, q \in S_x$, there exists an $s \in C_x$ such that $d(p,q) \le d(p,s) + d(s,q) \le 2r^*$ (since both points are close to more than half of points in C_x). Similarly, any two points $p, q \in S_y$ are $\le 2r^*$ apart.

Now, we will find a set of $k+2$ points that $(3,3)$-hits S. For now, assume that S_x and S_y are both nonempty. Given a pair $p \in S_x$, $q \in S_y$, we claim that $\{c_\ell\}_{\ell=1}^k \cup \{p,q\}$ $(3,3)$-hits S. Given a non-center $s \in C_i$ such that $i \neq x, y$, without loss of generality, let $s \in S_x$. Then c_i, p, and c_x are all distance $3r^*$ to s. Furthermore, c_i, c_x, and p are all distance $3r^*$ to c_i. Given a point $s \in C_x$, then c_x, c_y, and p are distance $3r^*$ to s, because $d(c_x, c_y) \le 2r^*$. Finally, c_x, c_y, and p are distance $3r^*$ to c_x, and similar arguments hold for $s \in C_y$ and c_y. Therefore, $\{c_\ell\}_{\ell=1}^k \cup \{p,q\}$ $(3,3)$-hits S.

If $S_x = \emptyset$ or $S_y = \emptyset$, then we can prove a slightly stronger statement: For each pair of non-centers $\{p,q\}$, $\{c_\ell\}_{\ell=1}^k \cup \{p,q\}$ $(3,3)$-hits S. Without loss of generality, let $S_y = \emptyset$. Given a point $s \in C_i$ such that $i \neq x$ and $i \neq y$, then c_i, c_x, and p are all distance $3r^*$ to s. Given a point $s \in C_x$, then p, c_x, and c_y are all distance $\le 3r^*$ to s. Given a point $s \in C_y$, then p, c_x, and c_y are all distance $\le 3r^*$ to s, because $s, p \in S_x$ implies $d(s,p) \le 2r^*$. Thus, we have proven case 1.

Now, we turn to the other case. Assume there does not exist a CCC2 to a PR cluster, disregarding a PR center. In this case, we need to use the assumption that there exist $(3,\epsilon)$-PR clusters C_x and C_y, and $p \in C_x$, $q \in C_y$ such that $d(p,q) \le r^*$. Then, by the triangle inequality, p is distance $\le 3r^*$ to all points in C_x and C_y. Consider the following d'':

$$d''(s,t) = \begin{cases} \min(3r^*, 3d(s,t)) & \text{if } s = p,\ t \in C_x \cup C_y, \\ 3d(s,t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation, because $d(p,v) \le 3r^*$ for all $v \in C_x \cup C_y$. Define d' as the metric completion of d''. Then, by Lemma 4.1, d' is a 3-metric perturbation with optimal cost $3r^*$. Given any non-center $s \in S$, the set of centers $\{c_\ell\}_{\ell=1}^k \setminus \{c_x, c_y\} \cup \{p,s\}$ achieves the optimal cost, since p is distance $3r^*$ from $C_x \cup C_y$, and all other clusters have the same center as in $\mathcal{OPT}$ (achieving radius $3r^*$).

From Fact 6.10, one of the centers in $\{c_\ell\}_{\ell=1}^k \setminus \{c_x, c_y\} \cup \{p,s\}$ must be the center for all but ϵn points in C_x under d'. If this center is c_ℓ for $\ell \neq x, y$, then for all but ϵn points $t \in C_x$, $d(c_\ell, t) \le r^*$ and $d(c_\ell, t) < d(c_z, t)$ for all $z \neq \ell, x, y$. So, by definition, c_ℓ is a CCC2 for C_x disregarding c_y, which contradicts our assumption. Similar logic applies to the center for all but ϵn points in C_y. Therefore, p and s must be the centers for C_x and C_y. Since s was an arbitrary non-center, all non-centers are distance $\le r^*$ to all but ϵn points in either C_x or C_y.

Similar to Case 1, we now partition all the non-centers into two sets S_x and S_y, such that

$$S_x = \{u \mid \text{for the majority of points } v \in C_x,\ d(u,v) \le r^*\}$$

and

$$S_y = \{u \mid u \notin S_x \text{ and for the majority of points } v \in C_y,\ d(u,v) \le r^*\}.$$

As before, each pair of points in S_x are distance $\le 2r^*$ apart and similarly for S_y. It is no longer true that $d(c_x, c_y) \le 2r^*$; however, we can prove that for both S_x and S_y, there exist points from two distinct clusters each. From the previous paragraph, given a non-center $s \in C_i$ for $i \neq x, y$, we know that p and s are centers for C_x and C_y. With an identical argument, given $t \in C_j$ for $j \neq x, y, i$, we can show that q and t are centers for C_x and C_y. It follows that S_x and S_y both contain points from at least two distinct clusters.

Now, we finish the proof by showing that for each pair $u \in S_x$, $v \in S_y$, $\{c_\ell\}_{\ell=1}^k \cup \{u,v\}$ $(3,3)$-hits S. Given a non-center $s \in C_i$, without loss of generality, $s \in S_x$, then there exists $j \neq i$ and $t \in C_j \cap S_x$. Then c_i, c_j, and u are $3r^*$ to s and c_i, c_x, and u are $3r^*$ to c_i. In the case where $i = x$, then c_i, c_j, and u are $3r^*$ to c_i. This concludes the proof. □

Now, we move to the second half of the proof of Theorem 6.8. Recall that the proof from the previous section relied on a ranking argument, in which optimal clusters were mapped to their closest centers from the set C of $k+2$ points from the first half of the proof. This is the basis for the following fact:

Fact 6.12. *Given a k-center clustering instance (S, d) with optimal clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ such that for all $i \in [k]$, $|C_i| > 2\epsilon n$, let d' denote an α-perturbation of d and let $\mathcal{C}'$ denote the set of (α, ϵ)-PR clusters. For each $C_x \in \mathcal{C}'$, there exists a ranking $R_{x,d'}$ of S such that for any set of optimal centers $C = \{c'_1, \ldots, c'_k\}$ under d', the center that is closest in d' to all but ϵn points in C_x is the highest-ranked point in $R_{x,d'}$.*[13]

Proof. Assume the lemma is false. Then there exists an (α, ϵ)-PR cluster C_i, two distinct points $u, v \in S$, and two sets of k centers C and C' both containing u and v, and both sets achieve the optimal score under an α-perturbation d', but u is the center for C_i in C while v is the center for C_i in C'. Then $\mathrm{Vor}_C(u)$ is ϵ-close to C_i; similarly, $\mathrm{Vor}_{C'}(v)$ is ϵ-close to C_i. This implies u is closer to all but ϵn points in C_i than v, and v is closer to all but ϵn points in C_i than u. Since $|C_i| > 2\epsilon n$, this causes a contradiction. □

We also define $R_{x,d',C} : C \to [n']$ as the ranking specific to C. Recall that our goal is to show a contradiction assuming two points from different PR clusters are close. From Lemma 6.6 and Lemma 6.11, we know there is a set of $k+2$ points, and any size k subset is optimal under a suitable perturbation. By Lemma 6.10, each size k subset must have a mapping from PR clusters to centers, and from Fact 6.12, these mappings are derived from a ranking of all possible center points by the PR clusters. In other words, each PR cluster C_x can rank all the points in S, so for any set of optimal centers for an α-perturbation, the top-ranked center is the one whose cluster is ϵ-close to C_x. Now, using Fact 6.12, we can try to give a contradiction by showing that there is no set of rankings for the PR clusters that is consistent with all the optimal sets of centers guaranteed by Lemmas 6.6 and 6.11. The following lemma gives relationships among the possible rankings. These will be our main tools for contradicting PR and thus finishing the proof of Theorem 6.8.

Lemma 6.13. *Given a k-center clustering instance (S, d) such that all optimal clusters are size $> 2\epsilon n$, and given non-centers $p, q \in S$ such that $C = \{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3,3)$-hits S, let the set $\mathcal{C}'$ denote the set of $(3, \epsilon)$-PR clusters. Consider a 3-perturbation d' such that all size k subsets of C are optimal sets of centers under d'. The following are true:*

(1) Given $C_x \in \mathcal{C}'$ and C_i such that $i \neq x$, $R_{x,d'}(c_x) < R_{x,d'}(c_i)$.
(2) There do not exist $s \in C$ and $C_x, C_y \in \mathcal{C}'$ such that $x \neq y$, and $R_{x,d',C}(s) + R_{y,d',C}(s) \leq 4$.
(3) Given C_i and $C_x \in \mathcal{C}'$ such that $x \neq i$, if $R_{x,d',C}(c_i) \leq 3$, then for all $C_y \in \mathcal{C}'$ such that $y \neq x, i$, $R_{y,d',C}(p) \geq 3$ and $R_{y,d',C}(q) \geq 3$.

Proof.

(1) By definition of the optimal clusters, for each $s \in C_x$, $d(c_x, s) < d(c_i, s)$, and therefore, by construction, $d'(c_x, s) < d'(c_i, s)$. It follows that $R_{x,d'}(c_x) < R_{x,d'}(c_i)$.
(2) Assume there exists $s \in C$ and $C_x, C_y \in \mathcal{C}'$ such that $R_{x,d',C}(s) + R_{y,d',C}(s) \leq 4$.
Case 1: $R_{x,d',C}(s) = 1$ and $R_{y,d',C}(s) \leq 3$. Define u and v such that $R_{y,d',C}(u) = 1$ and $R_{y,d',C}(v) = 2$. (If u or v is equal to s, then redefine it to an arbitrary center in $C \setminus \{s, u, v\}$.) Consider the set of centers $C' = C \setminus \{u, v\}$, which is optimal under d' by Lemma 6.6. By Fact 6.12, s is the center for all but ϵn points in both C_x and C_y, causing a contradiction.

[13] Formally, for each $C_x \in \mathcal{C}'$, there exists a bijection $R_{x,d'} : S \to [n]$ such that for all sets of k centers C that achieve the optimal cost under d', then $c = \mathrm{argmin}_{c' \in C} R_{x,d'}(c')$ if and only if $\mathrm{Vor}_C(c)$ is ϵ-close to C_x.

Case 2: $R_{x,d',C}(s) = 2$ and $R_{y,d',C}(s) = 2$. Define u and v such that $R_{x,d',C}(u) = 1$ and $R_{y,d',C}(v) = 1$. (Again, if u or v is equal to s, then redefine it to an arbitrary center in $C \setminus \{s, u, v\}$.) Consider the set of centers $C' = C \setminus \{u, v\}$, which is optimal under d' by Lemma 6.6. However, by Fact 6.12, s is the center for all but ϵn points in both C_x and C_y, causing a contradiction.

(3) Assume $R_{x,d',C}(c_i) \leq 3$.

Case 1: $R_{x,d',C}(c_i) = 2$. Then, by Lemma 6.13 part 1, $R_{x,d',C}(c_x) = 1$. Consider the set of centers $C' = C \setminus \{c_x, p\}$, which is optimal under d'. By Fact 6.12, $\text{Vor}_{C'}(c_i)$ must be ϵ-close to C_x. In particular, $\text{Vor}_{C'}(c_i)$ cannot contain more than ϵn points from C_i. But by definition, for all $j \neq i$ and $s \in C_i$, $d(c_i, s) < d(c_j, s)$. It follows that $\text{Vor}_{C'}(q)$ must contain all but ϵn points from C_i. Therefore, for all but ϵn points $s \in C_i$, for all j, $d'(q, s) < d'(c_j, s)$. If $R_{y,d',C}(q) \leq 2$, then C_y ranks c_y or p number one. Then for the set of centers $C' = C \setminus \{c_y, p\}$, $\text{Vor}_{C'}(q)$ contains more than ϵn points from C_y and C_i, contradicting the fact that C_y is $(3, \epsilon)$-PR. Therefore, $R_{y,d',C}(q) \geq 3$. The argument to show $R_{y,d',C}(p) \geq 3$ is symmetric.

Case 2: $R_{x,d',C}(c_i) = 3$. If there exists $j \neq i, x$ such that $R_{x,d',C}(c_i) = 2$, then without loss of generality, we are back in case 1. By Lemma 6.13 part 1, $R_{x,d',C}(c_x) \leq 2$. Then either p or q are ranked top two, without loss of generality, $R_{x,d',C}(p) \leq 2$. Consider the set $C' = C \setminus \{c_x, p\}$. Then as in the previous case, $\text{Vor}_{C'}(c_i)$ must be ϵ-close to C_x, implying for all but ϵn points $s \in C_i$, for all j, $d'(q, s) < d'(c_j, s)$. If $R_{y,d',C}(q) \leq 2$, again, then C_y ranks c_y or p as number one. Let $C' = C \setminus \{c_y, p\}$, and then $\text{Vor}_{C'}(q)$ contains more than ϵn points from C_y and C_i, causing a contradiction. Furthermore, if $R_{y,d',C}(p) \leq 2$, then we arrive at a contradiction by Lemma 6.13 part 2. □

We are almost ready to bring everything together to give a contradiction. Recall that Lemma 6.11 allows us to choose a pair (p, q) such that $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hits S. For an arbitrary choice of p and q, we may not end up with a contradiction. It turns out, we will need to make sure one of the points comes from a PR cluster and is very high in the ranking list of its own cluster. This motivates the following fact, which is the final piece to the puzzle:

Fact 6.14. *Given a k-center clustering instance (S, d) such that all optimal clusters are size $> 2\epsilon n$, given an (α, ϵ)-PR cluster C_x, and given $i \neq x$, then there are fewer than ϵn points $s \in C_x$ such that $d(c_i, s) \leq \min(r^*, \alpha d(c_x, s))$.*

Proof. Assume the fact is false. Then let $B \subseteq C_x$ denote a set of size ϵn such that for all $s \in B$, $d(c_i, s) \leq \min(r^*, \alpha d(c_x, s))$. Construct the following perturbation d': For all $s \in B$, set $d'(c_x, s) = \alpha d(c_x, s)$. For all other pairs s, t, set $d'(s, t) = d(s, t)$. This is clearly an α-perturbation by construction. Then the original set of optimal centers still achieves cost r^* under d', because for all $s \in B$, $d'(c_i, s) \leq r^*$. Clearly, the optimal cost under d' cannot be $< r^*$. It follows that the original set of optimal centers C is still optimal under d'. However, all points in B are no longer in $\text{Vor}_C(c_x)$ under d', contradicting the fact that C_x is (α, ϵ)-PR. □

Now, we are ready to prove Theorem 6.8.

Proof of Theorem 6.8. Assume towards contradiction that there are two points at distance $\leq r^*$ from different $(3, \epsilon)$-PR clusters. Then, by Lemma 6.11, there exists a partition S_1, S_2 of non-centers of S such that for all pairs $p \in S_1$, $q \in S_2$, $\{c_\ell\}_{\ell=1}^k \cup \{p, q\}$ $(3, 3)$-hit S. Given three $(3, \epsilon)$-PR clusters C_x, C_y, and C_z, let c'_x, c'_y, and c'_z denote the centers in $\{c_1, \ldots, c_k\}$ ranked highest by C_x, C_y, and C_z disregarding c_x, c_y, and c_z, respectively. Define $p = \text{argmin}_{s \in C_x} d(c_x, s)$, and without loss of generality, let $p \in S_1$. Then pick an arbitrary point q from S_2 and define $C = \{c_\ell\}_{\ell=1}^k \cup \{p, q\}$. Define

d' as in Lemma 6.6 (i.e., we define d' so all size k subsets of C are optimal sets of centers under d'). We claim that $R_{x,d',C}(p) < R_{x,d',C}(c'_x)$: From Fact 6.14, there are fewer than ϵn points $s \in C_x$ such that $d(c'_x, s) \leq \min(r^*, 3d(c_x, s))$. Among each remaining point $s \in C_x$, we will show $d'(p, s) \leq d'(c'_x, s)$. Recall that $d(p, s) \leq d(p, c_x) + d(c_x, s) \leq 2r^*$, so $d'(p, s) = \min(3r^*, 3d(p, s))$. There are two cases to consider.

Case 1: $d(c'_x, s) > r^*$. Then, by construction, $d'(c'_x, s) \geq 3r^*$, and so $d'(p, s) \leq d'(c'_x, s)$.

Case 2: $3d(c_x, s) < d(c'_x, s)$. Then

$$\begin{aligned}
d'(p, s) &\leq 3d(p, s) && \text{by construction of } d',\\
&\leq 3(d(p, c_x) + d(c_x, s)) && \text{by triangle inequality,}\\
&\leq 6d(c_x, s) && \text{by definition of } p,\\
&\leq 2d(c'_x, s) && \text{by assumption,}\\
&\leq \min(3r^*, 3d(c'_x, s)) && \text{by construction of } d',\\
&= d'(c'_x, s),
\end{aligned}$$

and this proves our claim.

Because $R_{x,d',C}(p) < R_{x,d',C}(c'_x)$ and $R_{x,d',C}(c_x) < R_{x,d',C}(c'_x)$, it follows that the top two can only be c_x, p, or q. Therefore, either $R_{x,d',C}(p) \leq 2$ or $R_{x,d',C}(q) \leq 2$. The rest of the argument is broken up into cases.

Case 1: $R_{x,d',C}(c'_x) \leq 3$. From Lemma 6.13 part 3, then $R_{y,d',C}(p) \geq 3$ and $R_{y,d',C}(q) \geq 3$. It follows by process of elimination that $R_{y,d',C}(c_y) = 1$ and $R_{y,d',C}(c_{y'}) = 2$. Again, by Lemma 6.13 part 3, $R_{x,d',C}(p) \geq 3$ and $R_{x,d',C}(q) \geq 3$, causing a contradiction.

Case 2: $R_{x,d',C}(c_{x'}) > 3$ and $R_{y,d',C}(c_{y'}) \leq 3$. Then $R_{x,d',C}(p) \leq 3$ and $R_{x,d',C}(q) \leq 3$. From Lemma 6.13 part 3, $R_{x,d',C}(p) \geq 3$ and $R_{x,d',C}(q) \geq 3$, therefore, we have a contradiction. Note, the case where $R_{x,d',C}(c_{x'}) > 3$ and $R_{z,d',C}(c_{z'}) \leq 3$ is identical to this case.

Case 3: The final case is when $R_{x,d',C}(c_{x'}) > 3$, $R_{y,d',C}(c_{y'}) > 3$, and $R_{z,d',C}(c_{z'}) > 3$. So for each $i \in \{x, y, z\}$, the top three for C_i in C is a permutation of $\{c_i, p, q\}$. Then each $i \in \{x, y, z\}$ must rank p or q in the top two, so by the Pigeonhole Principle, either p or q is ranked top two by two different PR clusters, contradicting Lemma 6.13. This completes the proof. □

We note that Case 3 in Theorem 6.8 is the reason why we need to assume there are at least three $(3, \epsilon)$-PR clusters. If there are only two, C_x and C_y, then it is possible that there exist $u \in C_x, v \in C_y$ such that $d(u, v) \leq r^*$. In this case, for p, q, d', and C as defined in the proof of Theorem 6.8, if C_x ranks c_x, p, q as its top three and C_y ranks c_y, q, p as its top three, then there is no contradiction.

6.3 Asymmetric k-center

Now, we consider asymmetric k-center under $(3, \epsilon)$-PR. The asymmetric case is a more challenging setting, and our algorithm does not return the optimal solution; however, our algorithm outputs a clustering that is ϵ-close to the optimal solution.

Recall the definition of the symmetric set A from Section 3, $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$, equivalently, the set of all CCV's. We might first ask whether A respects the structure of $\mathcal{OPT}$, as it did under 2-perturbation resilience. Namely, whether *Condition 1:* all optimal centers are in A and *Condition 2:* $\arg\min_{q \in A} d(q, p) \in C_i \implies p \in C_i$ hold. In fact, we will show that neither conditions hold in the asymmetric case, but both conditions are only slightly violated.

6.3.1 Structure of optimal centers. First, we give upper and lower bounds on the number of optimal centers in A, which will help us construct an algorithm for $(3, \epsilon)$-PR later on. We call a center c_i "bad" if it is not in the set A, i.e., $\exists q$ such that $d(q, c_i) \leq r^*$ but $d(c_i, q) > r^*$. First, we give

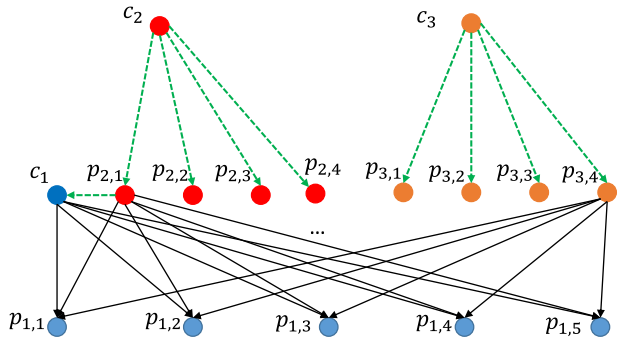


Fig. 8. An (α, ϵ)-perturbation resilient asymmetric k-center instance with one bad center (c_y). The dotted arrows are distance 1, and the solid arrows are distance $\frac{1}{\alpha}$.

an example of a $(3, \epsilon)$-PR instance with at least one bad center, and then we show that all $(3, \epsilon)$-PR instances must have at most six bad centers.

Lemma 6.15. *For all $\alpha, n, k \geq 1$ such that $\frac{n}{k} \in \mathbb{N}$, there exists a clustering instance with one bad center satisfying $(\alpha, \frac{2}{n})$-perturbation resilience.*

Proof. Given $\alpha, n, k \geq 1$, we construct a clustering instance such that all clusters are size $\frac{n}{k}$. Denote the clusters by $C_1, \ldots, C_k$ and the centers by $c_1, \ldots, c_k$. For each i, denote the non-centers in C_i by $p_{i,1}, \ldots, p_{i,L}$. Now, we define the distances as follows: For convenience, set $L = \frac{n}{k} - 1$. For all $2 \leq i \leq k$ and $1 \leq j \leq L$, let $d(c_i, p_{i,j}) = 1$. For all $2 \leq i \leq k$, $1 \leq j, \ell \leq L$, let $d(p_{i,j}, p_{1,\ell}) = \frac{1}{\alpha}$ and $d(c_1, p_{1,\ell}) = \frac{1}{\alpha}$. Finally, let $d(p_{2,1}, c_1) = 1$. All other distances are the maximum allowed by the triangle inequality. In particular, the distance between two points p and q is set to infinity unless there exists a path from p to q with finite distance edges defined above (see Figure 8).

The optimal clusters and centers are $C_1, \ldots, C_k$ and $c_1, \ldots, c_k$, achieving a radius of 1, and c_1 is a bad center, because $d(p_{2,1}, c_1) = 1$ but $d(c_1, p_{2,1}) = \infty$. It is left to show that this instance satisfies $(\alpha, \frac{2}{n})$-perturbation resilience. Given an arbitrary α-perturbation d', we must show that at most $\frac{2}{n} \cdot n = 2$ points switch clusters. By definition of an α-perturbation, for all p, q, we have $d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$ (recall that, without loss of generality, a perturbation only increases the distances). The centers $c_2, \ldots, c_k$ *must* remain optimal centers under d', since for all $2 \leq i \leq k$, $d'(c_i, p_{i,1}) \leq \alpha$ and no other point $q \neq c_i, p_{i,1}$ satisfies $d(q, p_{i,1}) < \infty$. Now, we must determine the final optimal center. Note that for all $2 \leq i, j \leq k$ and $1 \leq \ell, m \leq L$, we have

$$\begin{aligned} d'(p_{i,\ell}, p_{1,m}) &\leq \alpha d(p_{i,\ell}, p_{1,m}), \\ &\leq \alpha \cdot \frac{1}{\alpha}, \\ &< d(c_j, p_{1,m}), \\ &\leq d'(c_j, p_{1,m}). \end{aligned}$$

Therefore, c_j cannot be a center for $p_{i,\ell}$, for all $2 \leq i, j \leq k$ and $1 \leq \ell \leq L$. Therefore, the final optimal center c under d' must be either c_1 or $p_{i,\ell}$ for $2 \leq i \leq k$ and $1 \leq \ell \leq L$. Furthermore, it follows that c's cluster at least contains $C_1 \setminus \{c_1\}$ and for each $2 \leq i \leq k$, c_i's cluster at least contains $C_i \setminus \{c\}$. Therefore, the optimal clustering under d' differs from $\mathcal{OPT}$ by at most two points. This concludes the proof. □

Now, we show there are at most six bad centers for any asymmetric k-center instance satisfying $(3, \epsilon)$-PR.

Lemma 6.16. *Given a $(3,\epsilon)$-perturbation resilient asymmetric k-center instance such that all optimal clusters are size $> 2\epsilon n$, there are at most 6 bad centers, i.e., at most six centers c_i such that $\exists q$ with $d(q,c_i) \le r^*$ and $d(c_i,q) > r^*$.*

Proof. Assume the lemma is false. By assumption, there exists a set B, $|B| \ge 7$, of centers c_i such that $\exists q$ with $d(q,c_i) \le r^*$ and $d(c_i,q) > r^*$. The first step is to use this set of bad centers to construct a set C of $\le k-3$ points that are $\le 3r^*$ from every point in S. Once we find C, we will show how this set cannot exist under $(3,\epsilon)$-perturbation resilience, causing a contradiction.

Given a center $c_i \in B$ and q such that $d(q,c_i) \le r^*$ and $d(c_i,q) > r^*$, note that $d(c_i,q) > r^*$ implies $q \notin C_i$. For each $c_i \in B$, define $a(i)$ as the center of q's cluster. Then $d(a(i),c_i) \le d(a(i),q) + d(q,c_i) \le 2r^*$ and so for all $p \in C_i$, we have $d(a(i),p) \le d(a(i),c_i) + d(c_i,p) \le 3r^*$. If for each $c_i \in B$, $a(i)$ is not in B, then we would be able to remove B from the set of optimal centers, and the remaining centers are still distance $3r^*$ from all points in S (finishing the first half of the proof). However, we need to consider the case where there exist centers c_i in B such that $a(i)$ is also in B. Our goal is to show there exists a subset $B' \subseteq B$ of size 3, such that for each $c_i \in B'$, $a(i) \notin B'$; therefore, the set of optimal centers without B' is still distance $3r^*$ from all points in S.

Construct a directed graph $G = (B,E)$ where $E = \{(c_i,c) \mid c = a(i)\}$. Then every point has out-degree ≤ 1. Finding B' corresponds to finding ≥ 3 points with no edges to one another, i.e., an independent set of G. Consider a connected component $G' = (V',E')$ of G. Since V' is connected, we have $|E'| \ge |V'| - 1$. Since every vertex has out-degree ≤ 1, $|E'| \le |V'|$. Then, we have two cases.

Case 1: $|E'| = |V'| - 1$. Then G' is a tree, and so there must exist an independent set of size $\left\lceil \frac{|V'|}{2} \right\rceil$.

Case 2: $|E'| = |V'|$. Then G' contains a cycle, and so there exists an independent set of size $\left\lfloor \frac{|V'|}{2} \right\rfloor$.

It follows that we can always find an independent set of size $\left\lfloor \frac{|V'|}{2} \right\rfloor$ for the entire graph G. For $|B| \ge 7$, there exists such a set B' of size ≥ 3. Then, we have the property that $c_i \in B' \implies a(i) \notin B'$.

Now let $C = \{c_\ell\}_{\ell=1}^{k} \setminus B'$. By construction, B' is distance $\le 3r^*$ to all points in S. Consider the following 3-perturbation d': Increase all distances by a factor of 3, except $d(a(i),p)$, for i such that $c_i \in B'$ and $p \in C_i$, which we increase to $\min(3r^*, 3d(a(i),p))$. Then, by Lemma 2.8, the optimal radius is $3r^*$. Therefore, the set C achieves the optimal cost over d' even though $|C| \le k-3$. Then, we can pick any combination of three dummy centers, and they must all result in clusterings that are ϵ-close to $\mathcal{OPT}$. We will show this contradicts $(3,\epsilon)$-perturbation resilience.

We pick five arbitrary points $p_1,p_2,p_3,p_4,p_5 \in S \setminus C$ and define $C' = C \cup \{p_1,p_2,p_3,p_4,p_5\}$. From the above paragraph, each size 3 subset $P \subseteq \{p_1,p_2,p_3,p_4,p_5\}$ added to C will result in a set of optimal centers under d'. Then, by Fact 6.2, each point in $C \cup P$ must be the center for the majority of points in exactly one cluster. To obtain a contradiction, we consider the ranking defined by Fact 6.7 of C' over d'.

We start with a claim about the rankings: For each $c' \in C'$, for all pairs x,y such that $x \ne y$, if $\nexists c \in C$ such that $R_{x,d',C'}(c) < R_{x,d',C'}(c')$ or $R_{y,d',C'}(c) < R_{y,d',C'}(c')$, then $R_{x,d',C'}(c') + R_{y,d',C'}(c') \ge 5$. In words, there cannot be two clusters such that c' is ranked first among $C \cup \{c'\}$ and top two (or first and third) among C' for both clusters. Assume this is false. Then there exist $x \ne y$ such that $R_{x,d',C'}(c') + R_{y,d',C'}(c') \le 4$, so there are at most two total points ranked above c' in $R_{x,d',C'}$ and $R_{y,d',C'}$, and these points must be from the set $\{p_1,p_2,p_3,p_4,p_5\}$. Without loss of generality, denote these points by p and p' (if there are one or zero points ranked above c', then let one or both of p and p' be arbitrary). Then consider the set of centers $C' \setminus \{p,p'\}$ that is size k and must be optimal under d' as described earlier. However, the partitioning is not ϵ-close to $\mathcal{OPT}$, since c' is the best center (ranked 1) for both C_x and C_y. This completes the proof of the claim.

Now consider the set $D = \{c_i \in C \mid \exists x \text{ s.t. } R_{x,d',C'}(c_i) = 1\}$, i.e., the set of points in C that are ranked 1 for some cluster. Denote $m = (k-3) - |D|$, which is the number of points in C that are

ALGORITHM 5: $(3, \epsilon)$-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 $G = (A, E)$ where $E = \{(u, v) \mid d(u, v) \leq r^*\}$. Define a new symmetric k-center instance (S, A, d') where $d'(u, v) = dist_G(u, v)$.
(3) For all $k - 6 \leq k' \leq k$, run a symmetric k-center 2-approximation algorithm on (S, A, d'). If the output is a set of centers C achieving cost $\leq 2r^*$, then go to step 4.
(4) For all $C' \subseteq C$ of size $k - 6$ and $S' \subseteq S$ of size 6, return if $\text{cost}(C' \cup S') \leq 3r^*$.

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

not ranked 1 for any cluster. By the claim and since $|C| = k - 3$, there are exactly $m + 3$ clusters whose top-ranked point is not in C. Given one such cluster C_x, again by the claim, the top two ranked points must not be from the set D. Therefore, there are $2(m + 3)$ slots that must be filled by $m + 5$ points, so (for all $m \geq 0$) by the Pigeonhole Principle, there must exist a point $p \in C'$ ranked in the top two by two different clusters. This directly contradicts the claim, so we have a contradiction that completes the proof. □

6.3.2 Algorithm under $(3, \epsilon)$-PR. From the previous lemma, we know that at most a constant number of centers are bad. Essentially, our algorithm runs a symmetric 2-approximation algorithm on A, for all $k - 6 \leq k' \leq k$, to find a 2-approximation for the clusters in A. For instance, iteratively pick an unmarked point, and mark all points distance $2r^*$ away from it [30]. Then, we use brute force to find the remaining six centers, which will give us a 3-approximation for the entire point set. Under $(3, \epsilon)$-perturbation resilience, this 3-approximation must be ϵ-close to $\mathcal{OPT}$. We are not able to output $\mathcal{OPT}$ exactly, since Condition 2 may not be satisfied for up to ϵn points. The asymmetric k-center algorithm runs an approximation algorithm for symmetric k-center as a subroutine. The symmetric k-center instance (S, A, d) is a generalization: The set of allowable centers A is a subset of the points S to be clustered. The classic 2-approximation algorithms for k-center apply to this setting as well.

Theorem 6.17. *Algorithm 5 runs in polynomial time and outputs a clustering that is ϵ-close to $\mathcal{OPT}$ for $(3, \epsilon)$-perturbation resilient asymmetric k-center instances such that all optimal clusters are size $> 2\epsilon n$.*

Proof. We define three types of clusters. A cluster C_i is green if $c_i \in A$, it is yellow if $c_i \notin A$ but $C_i \cap A \neq \emptyset$, and it is red if $C_i \cap A = \emptyset$. Denote the number of yellow clusters by y and the number of red clusters by x. From Lemma 6.16, we know that $x + y \leq 6$. The symmetric k-center instance (S, A, d') constructed in step 2 of the algorithm is a subset of an instance with $k - x$ optimal clusters of cost r^*, so the $(k - x)$-center cost of (S, A, d') is at most r^*. Therefore, step 3 will return a set of centers achieving cost $\leq 2r^*$ for some $k' \leq k - x$. By definition of green clusters, we know that $k - x - y$ clusters have their optimal center in A. For each green cluster C_i, let $c(i) \in C$ denote the center that is distance $\leq 2r^*$ to c_i (if there is more than one point in C, then denote $c(i)$ by one of them arbitrarily). Let $C' = \{c(i) \mid C_i \text{ is green}\}$ and $|C'| \leq k - x - y$. Then the set $C' \cup \{c_x \mid x \text{ is not green}\}$ is cost $\leq 3r^*$, and the algorithm is guaranteed to encounter this set in the final step.

Finally, we explain why $C' \cup \{c_x \mid x \text{ is not green}\}$ must be ϵ-close to $\mathcal{OPT}$. Let $B = \{c_x \mid x \text{ is not green}\}$. Create a 3-perturbation in which we increase all distances by 3, except for the distances from $C' \cup B$ to all points in their Voronoi tile, which we increase up to $3r^*$. Then, the

optimal score is $3r^*$ by Lemma 4.1, and $C' \cup B$ achieves this score. Therefore, by $(3, \epsilon)$-perturbation resilience, the Voronoi tiling of $C' \cup B$ must be ϵ-close to $\mathcal{OPT}$. This completes the proof. □

6.4 APX-hardness under Perturbation Resilience

Now, we show hardness of approximation even when it is guaranteed the clustering satisfies (α, ϵ)-perturbation resilience for $\alpha \geq 1$ and $\epsilon > 0$. The hardness is based on a reduction from the general clustering instances, so the APX-hardness constants match the non-stable APX-hardness results. This shows the condition on the cluster sizes in Theorem 6.1 is tight.[14]

Theorem 6.18. *Given $\alpha \geq 1$, $\epsilon > 0$, it is NP-hard to approximate k-center to 2, k-median to 1.73, or k-means to 3.94, even when it is guaranteed the instance satisfies (α, ϵ)-perturbation resilience.*

Proof. Given $\alpha \geq 1$, $\epsilon > 0$, assume there exists a β-approximation algorithm $\mathcal{A}$ for k-median under (α, ϵ)-perturbation resilience. We will show a reduction to k-median without perturbation resilience. Given a k-median clustering instance (S, d) of size n, we will create a new instance (S', d') for $k' = k + n/\epsilon$ with size $n' = n/\epsilon$ as follows: First, set $S' = S$ and $d' = d$ and then add n/ϵ new points to S', such that their distance to every other point is $2\alpha n \max_{u,v \in S} d(u, v)$. Let $\mathcal{OPT}$ denote the optimal solution of (S, d). Then the optimal solution to (S', d') is to use $\mathcal{OPT}$ for the vertices in S and make each of the n/ϵ added points a center. Note that the cost of $\mathcal{OPT}$ and the optimal clustering for (S', d') are identical, since the added points are distance 0 to their center. Given a clustering C on (S, d), let C' denote the clustering of (S', d') that clusters S as in C and then adds n/ϵ extra centers on each of the added points. Then the cost of C and C' are the same, so it follows that C is a β-approximation to (S, d) if and only if C' is a β-approximation to (S', d'). Next, we claim that (S', d') satisfies (α, ϵ)-perturbation resilience. Given a clustering C' that is an α-approximation to (S', d'), then there must be a center located at all n/ϵ of the added points, otherwise the cost of C' would be $> \alpha\mathcal{OPT}$. Therefore, C' agrees with the optimal solution on all points except for S; therefore, C' must be ϵ-close to the optimal solution. Now that we have established a reduction, the theorem follows from hardness of 1.73-approximation for k-median [32]. The proofs for k-center and k-means are identical, using hardness from Reference [27] and Reference [32], respectively. □

7 CONCLUSION

Our work pushes the understanding of (promise) stability conditions farther in several 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 perturbation resilience. Our work also shows 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. Finally, we initiate the study of clustering under local stability. We define a local notion of perturbation resilience, and we give algorithms that simultaneously output all optimal clusters satisfying local stability, while ensuring the worst-case approximation guarantee. Although $\alpha = 2$ is tight for k-center, the best value of perturbation resilience for symmetric k-median and other center-based objectives is not known. Currently, the best upper bound is $\alpha = 2$ [1], but no lower bounds are known for $\alpha = 1 + \epsilon$, for constant $\epsilon > 0$.

[14] In fact, this hardness holds even under the strictly stronger notion of *approximation stability* [8]; therefore, it generalizes a hardness result by Balcan et al. [8].

REFERENCES

[1] Haris Angelidakis, Konstantin Makarychev, and Yury Makarychev. 2017. Algorithms for stable and perturbation—Resilient problems. In *Proceedings of the Symposium on Theory of Computing (STOC'17)*.

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

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012. Learning topic models—Going beyond SVD. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'12)*. 1–10.

[4] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. 2004. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.* 33, 3 (2004), 544–562.

[5] Pranjal Awasthi, Avrim Blum, and Or Sheffet. 2010. Stability yields a PTAS for k-median and k-means clustering. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'10)*. 309–318.

[6] Pranjal Awasthi, Avrim Blum, and Or Sheffet. 2012. Center-based clustering under perturbation stability. *Inf. Proc. Lett.* 112, 1 (2012), 49–54.

[7] Pranjal Awasthi and Or Sheffet. 2012. Improved spectral-norm bounds for clustering. In *Proceedings of the International Workshop on Approximation, Randomization, and Combinatorial Optimization Algorithms and Techniques (APPROX-RANDOM'12)*. Springer, 37–49.

[8] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. 2013. Clustering under approximation stability. *J. ACM* 60, 2 (2013), 8.

[9] Maria-Florina Balcan and Mark Braverman. 2009. Finding low error clusterings. In *Proceedings of the Conference on Learning Theory (COLT'09)*. 3–4.

[10] Maria-Florina Balcan and Mark Braverman. 2017. Nash equilibria in perturbation-stable games. *Theor. Comput.* 13, 13 (2017), 1–31.

[11] Maria Florina Balcan and Yingyu Liang. 2016. Clustering under perturbation resilience. *SIAM J. Comput.* 45, 1 (2016), 102–155.

[12] Maria Florina Balcan, Heiko Röglin, and Shang-Hua Teng. 2009. Agnostic clustering. In *Proceedings of the International Conference on Algorithmic Learning Theory*. 384–398.

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

[14] Yonatan Bilu, Amit Daniely, Nati Linial, and Michael E. Saks. 2013. On the practically interesting instances of MAX-CUT. In *Proceedings of the 30th International Symposium on Theoretical Aspects of Computer Science (STACS'13)*.

[15] Yonatan Bilu and Nathan Linial. 2012. Are stable instances easy? *Combin. Prob. Comput.* 21, 5 (2012), 643–660.

[16] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. 2015. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the Symposium on Discrete Algorithms (SODA'15)*. 737–756.

[17] Moses Charikar, Sudipto Guha, Éva Tardos, and David B. Shmoys. 1999. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the Symposium on Theory of Computing (STOC'99)*. 1–10.

[18] Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. 2001. Algorithms for facility location problems with outliers. In *Proceedings of the Symposium on Discrete Algorithms (SODA'01)*. 642–651.

[19] Chandra Chekuri and Shalmoli Gupta. 2018. Perturbation resilient clustering for k-center and related problems via LP relaxations. In *Proceedings of the International Workshop on Approximation, Randomization, and Combinatorial Optimization Algorithms and Techniques (APPROX-RANDOM'18)*. 9:1–9:16.

[20] Ke Chen. 2008. A constant factor approximation algorithm for k-median clustering with outliers. In *Proceedings of the Symposium on Discrete Algorithms (SODA'08)*. 826–835.

[21] Julia Chuzhoy, Sudipto Guha, Eran Halperin, Sanjeev Khanna, Guy Kortsarz, Robert Krauthgamer, and Joseph Seffi Naor. 2005. Asymmetric k-center is log* n-hard to approximate. *J. ACM* 52, 4 (2005), 538–551.

[22] Vincent Cohen-Addad and Chris Schwiegelshohn. 2017. On the local structure of stable clustering instances. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'17)*. 49–60.

[23] Amit Deshpande, Anand Louis, and Apoorv Vikram Singh. 2019. On Euclidean k-means clustering with alpha-center proximity. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'19)*.

[24] Martin E. Dyer and Alan M. Frieze. 1985. A simple heuristic for the p-centre problem. *Op. Res. Lett.* 3, 6 (1985), 285–288.

[25] Martin E. Dyer and Alan M. Frieze. 1986. Planar 3DM is NP-complete. *J. Algor.* 7, 2 (1986), 174–184.

[26] Zachary Friggstad, Kamyar Khodamoradi, and Mohammad R. Salavatipour. 2019. Exact algorithms and lower bounds for stable instances of Euclidean k-means. In *Proceedings of the Symposium on Discrete Algorithms (SODA'19)*.

[27] Teofilo F. Gonzalez. 1985. Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.* 38 (1985), 293–306.

[28] Rishi Gupta, Tim Roughgarden, and C. Seshadhri. 2014. Decompositions of triangle-dense graphs. In *Proceedings of the Conference on Innovations in Theoretical Computer Science (ITCS'14)*. 471–482.

[29] Moritz Hardt and Aaron Roth. 2013. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the Symposium on Theory of Computing (STOC'13)*. 331–340.
[30] Dorit S. Hochbaum and David B. Shmoys. 1985. A best possible heuristic for the k-center problem. *Math. Op. Res.* 10, 2 (1985), 180–184.
[31] Harry B. Hunt III, Madhav V. Marathe, Venkatesh Radhakrishnan, and Richard E. Stearns. 1998. The complexity of planar counting problems. *SIAM J. Comput.* 27, 4 (1998), 1142–1167.
[32] Kamal Jain, Mohammad Mahdian, and Amin Saberi. 2002. A new greedy approach for facility location problems. In *Proceedings of the Symposium on Theory of Computing (STOC'02)*. 731–740.
[33] Richard M. Karp. 1972. Reducibility among combinatorial problems. In *Complexity of Computer Computations*. Springer, 85–103.
[34] Jon Kleinberg and Eva Tardos. 2006. *Algorithm Design*. Pearson Education.
[35] Amit Kumar and Ravindran Kannan. 2010. Clustering with spectral norm and the k-means algorithm. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'10)*. 299–308.
[36] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. 2004. A simple linear time (1+ ε)-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS'04)*. 454–462.
[37] Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. 2017. α-expansion is exact on stable instances. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, Vol. 1050. 6.
[38] Euiwoong Lee, Melanie Schmidt, and John Wright. 2017. Improved and simplified inapproximability for k-means. *Inf. Proc. Lett.* 120 (2017), 40–43.
[39] Konstantin Makarychev, Yury Makarychev, Maxim Sviridenko, and Justin Ward. 2016. A bi-criteria approximation algorithm for *k* means. In *Proceedings of the International Workshop on Approximation, Randomization, and Combinatorial Optimization Algorithms and Techniques (APPROX-RANDOM'16)*. 14:1–14:20.
[40] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. 2014. Bilu-linial stable instances of max cut and minimum multiway cut. In *Proceedings of the Symposium on Discrete Algorithms (SODA'14)*. 890–906.
[41] Bodo Manthey and Matthijs B. Tijink. 2018. Perturbation resilience for the facility location problem. *Op. Res. Lett.* 46, 2 (2018), 215–218.
[42] Matúš Mihalák, Marcel Schöngens, Rastislav Šrámek, and Peter Widmayer. 2011. On the complexity of the metric TSP under stability considerations. In *SOFSEM: Theory and Practice of Computer Science*. Springer, 382–393.
[43] Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. 2012. The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM* 59, 6 (2012), 28.
[44] Tim Roughgarden. 2014. Beyond worst-case analysis. Retrieved from: http://theory.stanford.edu/ tim/f14/f14.html.
[45] Daniel A. Spielman and Shang-Hua Teng. 2004. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM* 51, 3 (2004), 385–463.
[46] Leslie G. Valiant and Vijay V. Vazirani. 1986. NP is as easy as detecting unique solutions. *Theoret. Comput. Sci.* 47 (1986), 85–93.
[47] Aravindan Vijayaraghavan, Abhratanu Dutta, and Alex Wang. 2017. Clustering stable instances of Euclidean k-means. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'17)*. 6503–6512.
[48] Sundar Vishwanathan. 1996. An O(Log*N) approximation algorithm for the asymmetric P-center problem. In *Proceedings of the Symposium on Discrete Algorithms (SODA'96)*. 1–5.
[49] Konstantin Voevodski, Maria-Florina Balcan, Heiko Röglin, Shang-Hua Teng, and Yu Xia. 2011. Min-sum clustering of protein sequences with limited distance information. In *Proceedings of the International Workshop on Similarity-based Pattern Recognition*. 192–206.

Received January 2019; revised October 2019; accepted December 2019