Clustering is an umbrella term for a wide variety of unsupervised data processing techniques. A relatively comprehensive description of clustering is that it aims to group together data instances that are similar, while separating dissimilar objects. Most of the common clustering tools output a partitioning of the input data into groups, clusters, that share some form of cohesiveness or between-cluster separation requirement[1]. However, in many cases, real data sets, in particular large ones, have on top of such cohesive separated groups, a significant amount of "background" unstructured data. An obvious example of such a scenario is when the input data set is the set of pixels of an image and the goal of the clustering is to detect groups of pixels that correspond to objects in that image. Clustering in such situations is the focus of this work. Maybe surprisingly, this topic has received relatively little attention in the clustering research community, and even less so when it comes to theoretical work.

The discussion of finding clustering structure in data sets that also contain subsets that do not conform well to that structure usually falls under the terminology of noise robustness (see e.g., [**?**],[**?**],[**?**], [**?**],[**?**]). However, noise robustness, at least in that context, addresses the noisy part of the data as either generated by some specific generative model (like uniform random noise, or Gaussian perturbations) or refers to worst-case adversarially generated noisy data. In this chapter we take a different approach. What distinguishes the noise that we consider form the "clean" part of the input data is that it is *structureless*. The exact meaning of such a notion of structurlessness may vary depending on the type of structure the clustering algorithm is aiming to detect in the data. We focus on defining structurelessness as not having significantly large dense subsets. We believe that such a notion is well suited to address "gray background" contrasting with cohesive subsets of the data that are the objects that the clustering aims to detect.

The distinction between structured and unstructured parts of the data requires, of course, a clear notion of relevant structure. For that, we resort to a relatively large body of recent work proposing notions of clusterable data sets. That work was developed mainly to address the gap between the computational hardness of (the optimization problem of) many common clustering objectives and the apparent feasibility of clustering in practical applications. We refer the reader to [**?**] for a survey of that body of work.

[1]The assignment to clusters can sometimes be probabilistic, and clusters may be allowed to intersect, but these aspects are orthogonal to the discussion in this chapter.

Here, we focus on two such notions, one based on the α-center-proximity introduced by [**?**] and the other, λ-separation, introduced by [**?**].

Our approach diverges from previous discussions of clusterable inputs in yet another aspect. Much of the theoretical research of clustering algorithms views clustering as an optimization problem. For some predetermined objective function (or clustering cost), the algorithm's task is to find the data partitioning that minimizes that objective. In particular, this approach is shared by all the works surveyed in [**?**]. However, in many practical situations the reality is different. Given a large data set to cluster, often-times there is no way a user may know what is the cost of the optimal clustering of that data, or how close to optimal the algorithm's outcome is. Instead, a user might have a notion of meaningful cluster structure, and will be happy with any outcome that meets such a requirement. Consequently, our algorithms aim to provide meaningful clustering solutions (where "meaningful" is defined in a way inspired by the above mentioned notions of clusterability) without reference to any particular optimization objective function. Our algorithms efficiently compute a hierarchical clustering tree that captures all such meaningful solutions. One should notice that all of those notions of clusterability (those under which it can be show that an objective-minimizing clustering can be found efficiently) assume that there exists an optimal solution that satisfies the meaningfulness condition (such as being perturbation robust, or having significantly smaller distances of points to their own cluster centers than to other centers). Under those assumptions, an algorithm that outputs a tree capturing all meaningful solutions, allows efficient detection of the cost-optimal clustering (in fact, the algorithms of [**?**] also yield such trees, for clean, noiseless inputs). Consequently, under the assumptions of those previous works, our algorithms yield an efficient procedure for finding such an optimal solution.

1 Related Work

The goal of clustering is to partition a set of objects into *dissimilar* subsets of *similar* objects. Based on the definition of similarity, the optimal solution to a clustering task is achieved by optimizing an objective function. Although solving this optimization problem is usually NP-hard, the clustering task is routinely and successfully employed in practice. This gap between theory and practice recommends characterizing the real world data sets by defin-

ing mathematical notions of *clusterable* data. As a result, provably efficient clustering algorithms can be found for these so called *nice* data.

In the past few years, there has been a line of work on defining notions of clusterability. The goal of all these methods has been to show that clustering is computationally efficient if the input X enjoys some nice structure. In [**?**], a clustering instance is considered to be *stable* if the optimal solution to a given objective function does not change under small multiplicative perturbations of distances between the points. Using this assumption, they give an efficient algorithm to find the max-cut clustering of graphs which are resilient to $O(\sqrt{|X|})$ perturbations. Using a similar assumption, [**?**] considered additive perturbations of the underlying metric and designed an efficient algorithm that outputs a clustering with near-optimal cost.

In terms of clusterability conditions, the most relevant previous papers are those addressing clsutering under α-*center proximity* condition (see Def. 5). Assuming that the centers belong to X (*proper* setting), [**?**] shows an efficient algorithm that outputs the optimal solution of a given center-based objective assuming that optimal solution satisfies the $(\alpha > 3)$-center proximity. This result was improved to $(\alpha = \sqrt{2} + 1 \approx 2.4)$ when the objective is k-median [**?**]. In [**?**] it was shown that unless P=NP such a result cannot be obtained for $(\alpha < 2)$-center proximal inputs.

However, as mentioned above, these results apply only to the noiseless case. Few methods have been suggested for analyzing clusterability in the presence of noise. [**?**] considers a dataset which has α-center proximity except for an ϵ fraction of the points. They give an efficient algorithm which provides a $1 + O(\epsilon)$-approximation to the cost of the k-median optimal solution when $\alpha > 2 + \sqrt{7} \approx 4.6$. Note that, while this result applies to adversarial noise as well, it only yields an approximation to the desired solution and the approximation guarantee is heavily influenced by the size of noise.

In a different line of work, [**?**] studied the problem of robustifying any center-based clustering objective to noise. To achieve this goal, they introduce the notion of *center separation* (look at Def. 7). Informally, an input has center separation when it can be covered by k well-separated set of balls. Given such an input, they propose a paradigm which converts any center-based clustering algorithm into a clustering algorithm which is robust to small amount of noise. Although this framework works for any objective-based clustering algorithm, it requires a strong restriction on the noise and clusterability of the data. For example, when the size of the noise is $\frac{5}{100}|X|$, their algorithm is able to obtain a robustified version of 2-median, only if X

is covered by k unit balls which are separated with distance 10.

In this work, we consider a natural relaxation of [**?**, **?**], with the goal to capture more realistic domains containing arbitrary amount of noise, assuming that noise is *structureless* (in a precise sense defined below). For example, in [**?**], the size of the noise $|\mathcal{N}| \leq \frac{m(C)}{8}$ (where $m(C)$ is size of the smallest cluster). Our algorithms can handle much larger amount of noise as long as they satisfy the *structureless* condition.

We define a novel notion of "gray background" noise. Informally, we call noise *structureless* if it does not have similar structure to a *nice* cluster at any part of the domain. Under that definition (look at Def. 6), our positive, efficient clustering results, do not depend on any restriction on the size of the noise.

Given a clusterable input X which contains *structureless* noise, we propose an efficient algorithm that outputs a hierarchical clustering tree of X that captures all *nice* clusterings of X. Our algorithm perfectly recovers the underlying *nice* clusterings of the input and its performance is independent of number of noisy points in the domain.

We complement our algorithmic results by proving that under more relaxed conditions, either on the level of clusterability of the clean part of the data, or on the unstructuredness requirements on the noise, such results become impossible.

1.1 Outline

The rest of this chapter is structured as follows. In Section 2, we present our notation and formal definitions. In Section 3 we show that the type of noise that we address in this paper is likely to arise under some natural assumptions on the data generating process. In Section 4, we present an efficient algorithm that, for any input set X which contains structureless noise, recovers all the underlying clusterings of non-noise subset of X that satisfies α-center proximity for $\alpha > 2 + \sqrt{7}$. We complement these results by proving that for $\alpha \leq 2\sqrt{2}+3$ in the case that we have arbitrary noise and for $\alpha \leq \sqrt{2} + 3$ in the case of structureless noise, efficient discovery of all nicely structured subsets is not possible.

In Section 5.1, we describe an efficient algorithm that, for any input X, recovers all the underlying clusterings of X that satisfy λ-center separation for $\lambda \geq 3$. We also prove that it is NP-Hard to improve this to $\lambda \leq 2$. In Section 5.2, we consider a similar problem in the presence of either arbitrary

or structureless noise. We propose an efficient algorithm that, for any input X which contains structureless noise, recovers all the underlying clusterings of non-noise subset of X that satisfy λ-center separation for $\lambda \geq 4$. We will also show that this result is tight for the case of structureless noise. We complement our results by showing that, under arbitrary noise assumption, no similar positive result can be achieved for $\lambda \leq 6$. Note that all our missing proofs can be found in the appendix.

2 Notation and definition

Let $(\mathbf{M}, d)$ be a metric space. Given a data set $X \subseteq \mathbf{M}$ and an integer k. A k-clustering of X denoted by $\mathcal{C}_X$ is a partition of X into k disjoints sets. Given points $c_1, \ldots, c_k \in \mathbf{M}$, we define the clustering induced by these points (or *centers*) by assigning each $x \in X$ to its nearest center. In the *steiner* setting, the centers can be arbitrary points of the metric space $\mathbf{M}$. In the *proper* setting, we restrict our centers to be members of the data set X. In this paper, we will be working in the **proper** setting.

For any set $\mathcal{A} \subseteq X$ with center $c \in \mathbf{M}$, we define the radius of $\mathcal{A}$ as $r_c(\mathcal{A}) = \max_{x \in \mathcal{A}} d(x, c)$. Throughout the paper, we will use the notation $\mathcal{C}_X$ to denote the clustering of the set X and $\mathcal{C}_S$ to denote the clustering of some $S \subseteq X$.

Definition 1 ($r(\mathcal{C}_X)$, $m(\mathcal{C}_X)$)**.** *Given a clustering $\mathcal{C}_X = \{C_1, \ldots, C_k\}$ induced by centers $c_1, \ldots, c_k \in \mathbf{M}$, we define $m(\mathcal{C}_\mathcal{X}) = \min_i |C_i|$ and $r(\mathcal{C}_\mathcal{X}) = \max_i r(C_i)$.*

Definition 2 ($\mathcal{C}_X$ restricted to a set)**.** *Given $S \subseteq X$ and a clustering $\mathcal{C}_X = \{C_1, \ldots, C_k\}$ of the set X. We define $\mathcal{C}_X$ restricted to the set S as $\mathcal{C}_{X|_S} = \{C_1 \cap S, \ldots, C_k \cap S\}$.*

Definition 3 ($\mathcal{C}_X$ respects $\mathcal{C}_S$)**.** *Given $S \subseteq X$, clusterings $\mathcal{C}_X = \{C_1, \ldots, C_k\}$ and $\mathcal{C}_S = \{S_1, \ldots, S_{k'}\}$. We say that $\mathcal{C}_X$ respects $\mathcal{C}_S$ if $\mathcal{C}_{X|_S} = \mathcal{C}_S$.*

Definition 4 ($\mathcal{T}$ or $\mathcal{L}$ captures $\mathcal{C}_S$)**.** *Given a hierarchical clustering tree $\mathcal{T}$ of X and a clustering $\mathcal{C}_S$ of $S \subseteq X$. We say that $\mathcal{T}$ captures $\mathcal{C}_S$ if there exists a pruning $\mathcal{P}$ which respects $\mathcal{C}_S$.*

Similarly, given a list of clusterings $\mathcal{L}$ of X and a clustering $\mathcal{C}_S$ of $S \subseteq X$. We say that $\mathcal{L}$ captures $\mathcal{C}_S$ if there exists a clustering $\mathcal{C}_X \in \mathcal{L}$ which respects $\mathcal{C}_S$.

Definition 5 (α-center proximity [?]). *A clustering $\mathcal{C}_X = \{C_1, \ldots, C_k\}$ satisfies α-center proximity w.r.t X and k if there exist centers $c_1, \ldots, c_k \in \mathbf{M}$ such that the following holds. For all $x \in C_i$ and $i \neq j$, $\alpha d(x, c_i) < d(x, c_j)$*

Next, we formally define our notion of structureless noise. Roughly, such noise should be scattered sparsely, namely, there should be no significant amount of noise in any small enough ball. Note that such a restriction does not impose any upper bound on the number of noise points.

Definition 6 ((α, η)-center proximity). *Given $S \subseteq X$, a clustering $\mathcal{C}_S = \{S_1, \ldots, S_k\}$ has (α, η)-center proximity w.r.t X, S and k if there exists centers $s_1, \ldots, s_k \in \mathbf{M}$ such that the following holds.*

- $\diamond$ **α-center proximity**: *For all $x \in S_i$ and $i \neq j$, $\alpha d(x, s_i) < d(x, s_j)$*
- $\diamond$ **η-sparse noise**: *For any ball B, $r(B) \leq \eta\, r(\mathcal{C}_S) \implies |B \cap (X \setminus S)| < \frac{m(\mathcal{C}_S)}{2}$*

Definition 7 (λ-center separation [?]). *A clustering $\mathcal{C}_X = \{C_1, \ldots, C_k\}$ has λ-center separation w.r.t X and k if there exists centers $c_1, \ldots, c_k \in \mathbf{M}$ such that $\mathcal{C}_X$ is the clustering induced by these centers and the following holds. For all $i \neq j$, $d(c_i, c_j) > \lambda r(\mathcal{C}_X)$*

Definition 8 ((λ, η)-center separation). *Given $S \subseteq X$, a clustering $\mathcal{C}_S$ has (λ, η)-center separation w.r.t X, S and k if there exists centers $s_1, \ldots, s_k \in \mathbf{M}$ such that $\mathcal{C}_X$ is the clustering induced by these centers and the following holds.*

- $\diamond$ **λ-center separation**: *For all $i \neq j$, $d(s_i, s_j) > \lambda r(\mathcal{C}_S)$*
- $\diamond$ **η-sparse noise**: *For any ball B, $r(B) \leq \eta\, r(\mathcal{C}_S) \implies |B \cap (X \setminus S)| < \frac{m(\mathcal{C}_S)}{2}$*

We denote a ball of radius x at center c by $B(c, x)$. We denote by $P_i(c)$ a collection of i many points sitting on the same location c. If the location is clear from the context, we will use the notation P_i.

3 Justification of sparse noise

In this section, we examine our sparseness condition. We will show that if the set of points $\mathcal{N}$ are generated by a non concentrated distribution in a ball in $\mathbf{R}^d$ then with high probability, as long as $\mathcal{N}$ is not too large (so as to" drown" the original data set), it will satisfy the sparse noise condition. The

proof is based on the epsilon approximation theorem for classes of finite VC-dimension, applied to the set of balls in $\mathbf{R}^d$. The following, rather natural, definition of non concentrated distribution was introduced in [**?**].

Definition 9. *A probability distribution over the d-dimensional unit ball is* non-concentrated *if, for some constant c, the probability density of any point x is at most c times its density under the uniform distribution over that ball.*

Theorem 10 (Noise by non concentrated distribution is sparse)**.** *Let X be a ball of radius R in $\mathbf{R}^d$ and $S \subseteq X$. Let $\mathcal{C}$ be a clustering of S which satisfies α-center proximity (or λ-center separation). Given parameters $\epsilon, \delta \in (0,1)$. Let $\mathcal{N} \subseteq X$ be picked i.i.d according to a non concentrated probability distribution. If $|\mathcal{N}| < c\left(\frac{R}{r(C)\eta}\right)^d m(C)$ then with high probability, $S \cup \mathcal{N}$ satisfies (α, η)-center proximity (the (λ, η)-center separation, respectively).*

Proof. Let $H = \{B \text{ is a ball } : B \subseteq X\}$. Observe that VC-Dim$(H) = d + 1$. Let $\gamma := \frac{r(C)}{R}$. Since the noise-generating distribution P is c-concentrarted, for every ball B, $P(B) \leq c\frac{\text{vol}(B)}{\text{vol}(X)} = c\gamma^d$. Now, the fundamental ϵ-approximation theorem (Theorem **??**) establishes the result. □ □

Note that Theorem 10 shows that the cardinality of the noise set, $|\mathcal{N}|$, can be much bigger than the size of the smallest cluster $m(\mathcal{C})$.

4 Center Proximity

In this section, we study the problem of recovering (α, η)-center proximal clusterings of a set X, in the presence of noise. The goal of our algorithm is to produce an efficient representation (hierarchical clustering tree) of all possible (α, η)-center proximal nice clusterings rather than to output a single clustering or to optimize an objective function. Here is a more precise overview of the results of this section:

- *Positive result under sparse noise* - In Section 4.1, we give our main result under sparse noise. If $\alpha \geq 2 + \sqrt{7} \approx 4.6$ and $\eta \geq 1$; for any value of t, Alg. 1 outputs a tree which captures all clusterings $\mathcal{C}^*$ (of a subset of X) which satisfy (α, η)-center proximity and $m(C^*) = t$.
- *Lower bound under sparse noise* - In Section 4.2, we show that if $\alpha \leq 2 + \sqrt{3} \approx 3.7$ and $\eta \leq 1$ then there is no tree and no list of 'small' size ($< 2^{k/2}$) which can capture all clusterings $\mathcal{C}$ (of a subset of X) which satisfy

(α, η)-center proximity even for a fixed value of the size of the smallest cluster ($m(C) = t$).

- *Lower bound with arbitrary noise* - In Section 4.3, we show that for a given value of a parameter t, if $\alpha \leq 2\sqrt{2} + 3 \approx 5.8$ and the number of noisy points exceeds $\frac{3}{2}t$ then no tree can capture all clusterings $\mathcal{C}$ (of a subset of X) which satisfy α-center proximity even for fixed $m(\mathcal{C}) = t$. Identical result holds for 'small' ($< 2^{k/2}$) lists if the number of noisy points exceeds $\frac{3k}{2}t$.

4.1 Positive result under sparse noise

Given a clustering instance (X, d) and a parameter t, we introduce an efficient algorithm which outputs a hierarchical clustering tree $\mathcal{T}$ of X with the following property. For every k, for every $S \subseteq X$ and for every k-clustering $\mathcal{C}_S$ which satisfies (α, η)-center proximity (for $\alpha \geq 2 + \sqrt{7}$ and $\eta \geq 1$) and $m(\mathcal{C}_S) = t$, $\mathcal{T}$ captures $\mathcal{C}_S$. It is important to note that our algorithm only knows X and has no knowledge of the set S.

Our algorithm has a linkage based structure similar to [**?**]. However, our method benefits from a novel *sparse distance condition*. We introduce the algorithm in Alg. 1 and prove its efficiency and correctness in Theorem 13 and Theorem 12 respectively.

Definition 11 (Sparse distance condition)**.** *Given a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ of the set X and a parameter t. We say that the ball $B \subseteq X$ satisfies the sparse distance condition w.r.t clustering $\mathcal{C}$ when the following holds.*

- $|B| \geq t$.
- *For any $C_i \in \mathcal{C}$, if $C_i \cap B \neq \emptyset$, then $C_i \subseteq B$ or $|B \cap C_i| \geq t/2$.*

Intuitively, Alg. 1 works as follows. It maintains a clustering $\mathcal{C}^{(l)}$, which is initialized so that each point is in its own cluster. It then goes over all pairs of points p, q in increasing order of their distance $d(p, q)$. If $B(p, d(p, q))$ satisfies the sparse distance condition w.r.t $\mathcal{C}^{(l)}$, then it merges all the clusters which intersect with this ball into a single cluster and updates $\mathcal{C}^{(l)}$. Furthermore, the algorithm builds a tree with the nodes corresponding to the merges performed so far. We will show that for all $S \subseteq X$ which are (α, η)-proximal t-min nice and for all clusterings $\mathcal{C}_S$ which have (α, η)-center proximity, Alg. 1 outputs a tree which captures $\mathcal{C}_S$.

Theorem 12. *Given a clustering instance (X, d) and a parameter t. Alg. 1 outputs a tree $\mathcal{T}$ with the following property. For all k, $S \subseteq X$ and for all*

Algorithm 1: Alg. for (α, η)-center proximity with parameter t

Input: (X, d) and t
Output: A hierarchical clustering tree T of X.

1 Let $\mathcal{C}^{(l)}$ denote the clustering X after l merge steps have been performed. Initialize $\mathcal{C}^{(0)}$ so that all points are in their own cluster. That is, $\mathcal{C}^{(0)} = \{\{x\} : x \in X\}$.
2 Go over all pairs of points p, q in increasing order of the distance $d(p, q)$. If $B = B(p, d(p, q))$ satisfies the sparse distance condition then
3 Merge all the clusters which intersect with B into a single cluster.
4 Output clustering tree T. The leaves of T are the points in dataset X. The internal nodes correspond to the merges performed.

k-clusterings $\mathcal{C}^*_S = \{S^*_1, \ldots, S^*_k\}$ which satisfy $(2 + \sqrt{7}, 1)$-center proximity the following holds. If $m(\mathcal{C}^*_S) = t$ then $\mathcal{T}$ captures $\mathcal{C}_S$.

Theorem 13. *Given clustering instance (X, d) and t. Alg. 1 runs in poly$(|X|)$.*

Proof. Let $n = |X|$. Checking if B satisfies the sparse-distance condition takes $O(n)$ time and hence the algorithm runs in $O(n^3)$ time. □

4.2 Lower bound under sparse noise

Theorem 14. *Given the number of clusters k and parameter t. For all $\alpha \leq 2 + \sqrt{3}$ and $\eta \leq 1$ there exists a clustering instance (X, d) such that any clustering tree $\mathcal{T}$ of X has the following property. There exists $S \subseteq X$ and clustering $\mathcal{C}_S$ which satisfies (α, η)-center proximity and $m(\mathcal{C}_S) = t$ but $\mathcal{T}$ doesn't capture $\mathcal{C}_S$.*

Theorem 15. *Given the number of clusters k and parameter t. For all $\alpha \leq 2 + \sqrt{3}$, $\eta \leq 1$ there exists (X, d) such that any list $\mathcal{L}$ (of clusterings of X) has the following property. If $|\mathcal{L}| < 2^{\frac{k}{2}}$ then there exists $S \subseteq X$ and clustering $\mathcal{C}_S$ which satisfies (α, η)-center proximity and $m(\mathcal{C}_S) = t$ but $\mathcal{L}$ doesn't capture $\mathcal{C}_S$.*

4.3 Lower bound under arbitrary noise

Theorem 16. *Given the number of clusters k and a parameter t. For all $\alpha < 2\sqrt{2}+3$ there exists (X,d) such that any clustering tree $\mathcal{T}$ of X has the following property. There exists $S \subseteq X$ and there exists clustering $\mathcal{C}_S$ which satisfies α-center proximity such that $m(\mathcal{C}_S) = t$ and the following holds. If $|X \setminus S| \geq \frac{3t(\mathcal{C}_S)}{2} + 5$, then $\mathcal{T}$ doesn't capture $\mathcal{C}_S$.*

Theorem 17. *Given the number of clusters k and parameter t. For all $\alpha \leq 2+\sqrt{2}+3$ there exists (X,d) such that any list $\mathcal{L}$ (of clusterings of X) has the following property. There exists $S \subseteq X$ and there exists clustering $\mathcal{C}_S$ which satisfies α-center proximity such that $m(\mathcal{C}_S) = t$ and the following holds. If $|\mathcal{L}| < 2^{\frac{k}{2}}$ and $|X \setminus S| \geq \frac{k}{2}(\frac{3t(\mathcal{C}_S)}{2} + 5)$, then $\mathcal{L}$ doesn't capture $\mathcal{C}_S$.*

5 Center Separation

5.1 Center Separation without noise

In this section, we study the problem of recovering λ-center separated clusterings of a set X, in the absence of noise. We do not want to output a single clustering but to produce an efficient representation (hierarchal clustering tree) of all possible λ-center separated nice clusterings. In Section 5.1.1 we give an algorithm that generates a tree of all possible λ-center separated clusterings of X for $\lambda > 3$. In Section 5.1.2, we prove that for $\lambda < 2$, it is NP-Hard to find any such clustering.

5.1.1 Positive result under no noise

Given a clustering instance (X,d), our goal is to output a hierarchical clustering tree T of X which has the following property. For every k and for every k-clustering $\mathcal{C}_X$ which satisfies λ-center separation, there exists a pruning $\mathcal{P}$ of the tree which equals $\mathcal{C}_X$. Our algorithm (Alg. 2) uses single-linkage to build a hierarchical clustering tree of X. We will show that when $\lambda \geq 3$ our algorithm achieves the above mentioned goal.

Theorem 18. *Given (X,d). For all $\lambda \geq 3$, Alg. 2 outputs a tree $\mathcal{T}$ with the following property. For all k and for all k-clusterings $\mathcal{C}^*_X = \{C^*_1, \ldots, C^*_k\}$ which satisfy λ-center separation w.r.t X and k, the following holds. For every $1 \leq i \leq k$, there exists a node N_i in the tree T such that $C^*_i = N_i$.*

Algorithm 2: Alg. for λ-center separation

Input: (X, d)
Output: A hierarchical clustering tree T of X.

1 Initialize the clustering so that each point is in its own cluster.
2 Run single-linkage till only a single cluster remains. Output clustering tree T.

5.1.2 Lower bound with no noise

We will prove that for $\lambda \leq 2$, finding any solution for λ-center separation is NP-Hard. [**?**] proved that finding any solution for α-center proximity is NP-Hard for $\alpha < 2$. Our reduction is same as the reduction used in Theorem 1 in [**?**] and hence we omit the proof.

Theorem 19. *Given a clustering instance (X, d) and the number of clusters k. For $\lambda < 2$, finding a clustering which satisfies λ-center separation is NP-Hard.*

5.2 Center Separation in the presence of noise

In this section, we study the problem of recovering (λ, η)-center separated clusterings of a set X, in the presence of noise. Here is a more precise overview of the results of this section:

- *Positive result under sparse noise* - In Section 5.2.1, we show that if $\lambda \geq 4$ and $\eta \geq 1$; for any value of parameters r and t, Alg. 3 outputs a clustering which respects all clusterings $\mathcal{C}^*$ (of a subset of X) which satisfies (λ, η)-center proximity and $m(C^*) = t$ and $r(C^*) = r$.
- *Lower bound under sparse noise* - In Section 5.2.2, we show that, if $\lambda < 4$ and $\eta \leq 1$ then there is no tree and no list of 'small' size ($< 2^{k/2}$) which can capture all clusterings $\mathcal{C}$ (of subset of X) which satisfy (λ, η)-center proximity even for fixed values of the size of the smallest cluster ($m(C) = t$) and maximum radius ($r(C) = r$).
- *Lower bound with arbitrary noise* - In Section 5.2.3, we show that for a given value of parameters r and t, if $\lambda \leq 6$ and the number of noisy points exceeds $\frac{3}{2}t$ then no tree can capture all clusterings $\mathcal{C}$ (of a subset of X) which satisfy λ-center separation even for fixed $m(\mathcal{C}) = t$ and $r(\mathcal{C}) = r$. Identical result holds for 'small' ($< 2^{k/2}$) lists if the number of noisy points exceeds $\frac{3k}{2}t$.

5.2.1 Positive result under sparse noise

We are given a clustering instance (X, d) and parameters r and t. Our goal is to output a clustering $\mathcal{C}_X$ which has the following property. For every k, for every $S \subseteq X$ and for every k-clustering $\mathcal{C}_S$ which satisfies (λ, η)-center separation, the clustering $\mathcal{C}_X$ restricted to S equals $\mathcal{C}_S$.

In the next section, we propose a clustering algorithm (Alg. 3) and prove (Theorem 20) that our algorithm indeed achieves the above mentioned goal (under certain assumptions on the parameters λ and η). It is important to note that our algorithm only knows X and has no knowledge of the set S.

Intuitively, Alg. 3 works as follows. In the first phase, it constructs a list of balls which have radius at most r and contain at least t points. It then constructs a graph as follows. Each ball found in the first phase is represented by a vertex. If two balls have a 'large' intersection then there is an edge between the corresponding vertices in the graph. We then find the connected components in the graph which correspond to the clustering of the original set X.

Theorem 20. *Given a clustering instance (X, d) and parameters r and t. For every k, for every $S \subseteq X$ and for all k-clusterings $\mathcal{C}_S^* = \{S_1^*, \ldots, S_k^*\}$ which satisfy $(4, 1)$-center separation such that $m(\mathcal{C}_S^*) = t$ and $r(\mathcal{C}_S^*) = r$, the following holds. Alg. 3 outputs a clustering $\mathcal{C}_X$ such that $\mathcal{C}_X|_S = \mathcal{C}_S^*$.*

Theorem 21. *Given (X, d) and parameters r and t. Alg. 3 runs in poly$(|X|)$.*

Proof. Let $n = |X|$. Phase 1 of Alg. 3 runs in $O(n^2)$ time. Phase 2 gets a list of size l. Constructing G and finding connected components takes $O(l^2)$ time. Hence, the algorithm runs in $O(n^2)$ time. □

5.2.2 Lower bound under sparse noise

Theorem 22. *Given the number of clusters k and parameters r and t. For all $\lambda < 4$ and $\eta \leq 1$, there exists a clustering instance (X, d) such that any clustering tree $\mathcal{T}$ of X has the following property. There exists $S \subseteq X$ and a k-clustering $\mathcal{C}_S = \{S_1, \ldots, S_k\}$ which satisfies (λ, η)-center separation such that $m(\mathcal{C}_S) = t$ and $r(\mathcal{C}_S) = r$, but $\mathcal{T}$ doesn't capture $\mathcal{C}_S$.*

Theorem 23. *Given the number of clusters k and parameters r and t. For all $\lambda \leq 4$ and $\eta \leq 1$ there exists a clustering instance (X, d) such that any list $\mathcal{L}$ (of clusterings of X) has the following property. If $|\mathcal{L}| < 2^{\frac{k}{2}}$ then there exists $S \subseteq X$ and clustering $\mathcal{C}_S$ which satisfies (λ, η)-center separation and $m(\mathcal{C}_S) = t$ and $r(\mathcal{C}_S) = r$, but $\mathcal{L}$ doesn't capture $\mathcal{C}_S$.*

Algorithm 3: Alg. for (λ, η)-center separation with parameters t and r

Input: $(X, d), t$ and r
Output: A clustering $\mathcal{C}$ of the set X.

Phase 1
Let $\mathcal{L}$ denote the list of balls found so far. Initialize $\mathcal{L}$ to be the empty set. $\mathcal{L} = \emptyset$.
Go over all pairs of points $p, q \in X$ in increasing order of the distance $d(p, q)$. Let $B := B(p, d(p, q))$. If $|B| \geq t$ and $r(B) \leq r$ then
    $\mathcal{L} = \mathcal{L} \cup B$
Output the list of balls $\mathcal{L} = \{B_1, \ldots, B_l\}$ to the second phase of the algorithm.

Phase 2
Construct a graph $G = (V, E)$ as follows. $V = \{v_1, v_2, \ldots, v_l\}$. If $|B_i \cap B_j| \geq t/2$ then construct an edge between v_i and v_j.
Find connected components $(G_1, \ldots, G_k)$ in the graph G.
Merge all the points in the same connected component together to get a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ of the set X.
Assign $x \in X \setminus \cup_i B_i$ to the closest cluster C_i. That is, $i := \arg\min_{j \in [k]} \min_{y \in C_j} d(x, y)$. Output $\mathcal{C}$.

5.2.3 Lower bound with arbitrary noise

Theorem 24. *Given the number of clusters k and parameters r and t. For all $\lambda < 6$, there exists a clustering instance (X, d) such that any clustering tree $\mathcal{T}$ of X has the following property. There exists $S \subseteq X$ and there exists k-clustering $\mathcal{C}_S$ which satisfies λ-center separation such that $m(\mathcal{C}_S) = t$, $r(\mathcal{C}_S) = r$ and the following holds. If $|X \setminus S| \geq \frac{3t}{2} + 5$, then $\mathcal{T}$ doesn't capture $\mathcal{C}_S$.*

Theorem 25. *Given the number of clusters k and parameters r and t. For all $\lambda \leq 6$ there exists (X, d) such that any list $\mathcal{L}$ (of clusterings of X) has the following property. There exists $S \subseteq X$ and there exists clustering $\mathcal{C}_S$ which satisfies λ-center separation such that $m(\mathcal{C}_S) = t$, $r(\mathcal{C}_S) = r$ and the following holds. If $|\mathcal{L}| < 2^{\frac{k}{2}}$ and $|X \setminus S| \geq \frac{k}{2}(\frac{3t(\mathcal{C}_S)}{2} + 5)$, then $\mathcal{L}$ doesn't capture $\mathcal{C}_S$.*