The Price of Hierarchical Clustering

Anna Arutyunova & Heiko Röglin

Suggested Citation:
Arutyunova, A., & Röglin, H. (2025). The Price of Hierarchical Clustering. Algorithmica, 87(10), 1420–1452. https://doi.org/10.1007/s00453-025-01327-7

Algorithmica (2025) 87:1420–1452
https://doi.org/10.1007/s00453-025-01327-7

The Price of Hierarchical Clustering

Anna Arutyunova[1] **· Heiko Röglin**[2]

Received: 23 November 2023 / Accepted: 13 May 2025 / Published online: 2 July 2025

Abstract
Hierarchical Clustering is a popular tool for understanding the hereditary properties of a data set. Such a clustering is actually a sequence of clusterings that starts with the trivial clustering in which every data point forms its own cluster and then successively merges two existing clusters until all points are in the same cluster. A hierarchical clustering achieves an approximation factor of α if the costs of each k-clustering in the hierarchy are at most α times the costs of an optimal k-clustering. We study as cost functions the maximum (discrete) radius of any cluster (k-center problem) and the maximum diameter of any cluster (k-diameter problem). In general, the optimal clusterings do not form a hierarchy and hence an approximation factor of 1 cannot be achieved. We call the smallest approximation factor that can be achieved for any instance the *price of hierarchy*. For the k-diameter problem we improve the upper bound on the price of hierarchy to $3 + 2\sqrt{2} \approx 5.83$. Moreover we significantly improve the lower bounds for k-center and k-diameter, proving a price of hierarchy of exactly 4 and $3 + 2\sqrt{2}$, respectively.

Keywords Hierarchical clustering · Approximation algorithms · k-center · k-diameter

Previously appeared on ESA 2022 [1].

✉ Anna Arutyunova
anna.arutyunova@hhu.de

Heiko Röglin
roeglin@cs.uni-bonn.de

[1] Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

[2] Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

1 Introduction

Clustering is an ubiquitous task in data analysis and machine learning. In a typical clustering problem, the goal is to partition a set of objects into different clusters such that only similar objects belong to the same cluster. There are numerous ways how clustering can be modeled formally and many different models have been studied in the literature in the last decades. In many theoretical models, one assumes that the data comes from a metric space and that the desired number of clusters is given. Then the goal is to optimize some objective function like *k*-center, *k*-median, or *k*-means. In most cases the resulting optimization problems are NP-hard and hence approximation algorithms have been studied extensively.

One aspect of real-world clustering problems that is not captured by these models is that it is often already a non-trivial task to determine for a given data set the right or most reasonable number of clusters. One particularly appealing way to take this into account is hierarchical clustering. A hierarchical clustering of a data set is actually a sequence of clusterings, one for each possible number of clusters. It starts with the trivial clustering in which every data point forms its own cluster and then successively merges two existing clusters until all points are in the same cluster. This way for every possible number of clusters, a clustering is obtained. These clusterings help to understand the hereditary properties of the data and they provide information at different levels of granularity.

While hierarchical clustering is successfully used in many applications, it is not as well understood from a theoretical point of view as the models in which the number of clusters is given as part of the input. One reason for this is that it is not obvious how the quality of a hierarchical clustering should be measured. A possibility that has been explored in the literature is to define the quality of a hierarchical clustering based on its worst level. To be precise, let $(\mathcal{X}, d)$ be a metric space and $\mathcal{P} \subset \mathcal{X}$ a set of n points. Furthermore let $\mathscr{H} = (\mathcal{H}_n, \ldots, \mathcal{H}_1)$ be a hierarchical clustering of $\mathcal{P}$, where $\mathcal{H}_k$ denotes a k-clustering, i.e., a clustering with at most k non-empty clusters. Then $\mathcal{H}_{k-1}$ arises from $\mathcal{H}_k$ by merging some of the existing clusters. We assume that some objective function like k-center, k-median, or k-means is selected and denote by $\mathsf{cost}(\mathcal{H}_k)$ the objective value of $\mathcal{H}_k$ with respect to the selected objective function. Furthermore, let $\mathcal{O}_k$ denote an optimal k-clustering and let $\mathsf{cost}(O_k)$ denote its objective value. Then we say that $\mathscr{H}$ achieves an approximation factor of $\alpha \geq 1$ if $\mathsf{cost}(\mathcal{H}_k) \leq \alpha \cdot \mathsf{cost}(\mathcal{O}_k)$ for every k, assuming that cost is an objective that is to be minimized. In this work we consider the radius objective, which is well-known from the k-center problem. Here the cost is defined as the maximum radius of a cluster. Furthermore we consider the diameter objective, where the cost is defined as the maximum distance between any two points lying in the same cluster.

An α-approximation for small α yields a strong guarantee for the hierarchical clustering on every level. However, in general there do not exist optimal clusterings $\mathcal{O}_n, \ldots, \mathcal{O}_1$ that form a hierarchy. So even with unlimited computational resources, a 1-approximation usually cannot be achieved. In the literature different algorithms for computing hierarchical clusterings with respect to different objective functions have been developed and analyzed. Dasgupta and Long [2] and Charikar et al. [3] initiated this line of research and presented both independently from each other an algorithm

that computes efficiently an 8-approximate hierarchical clustering with respect to the radius and diameter objective. That is, for every level k, the maximal radius or diameter of any cluster in the k-clustering computed by their algorithms is at most 8 times the maximal radius or diameter in an optimal k-clustering. Inspired by [2], Plaxton [4] proposed a constant-factor approximation for the k-median and k-means objective. Later a general framework that also leads constant approximation guarantees for many objective functions including in particular k-median and k-means has been proposed by Lin et al. [5].

Despite these articles and other related work, which we discuss below in detail, many questions in the area of hierarchical clustering are not yet resolved. We find it particularly intriguing to find out which approximation factors can be achieved for different objectives. This question comes in two flavors depending on the computational resources available. Of course it is interesting to study which approximation factors can and cannot be achieved in polynomial time, assuming P $\neq$ NP. Since in general there do not exist hierarchical clusterings that are optimal on each level, it is also interesting to study which approximation factors can and cannot be achieved in general without the restriction to polynomial-time algorithms.

For an objective function like radius or diameter we define its *price of hierarchy* as the smallest α such that for any instance there exists an α-approximate hierarchical clustering. Hence, the price of hierarchy is a measure for how much quality one has to sacrifice for the hierarchical structure of the clusterings.

Our main results are tight bounds for the price of hierarchy for the radius, discrete radius and diameter objective. Here the difference between radius and discrete radius lies in the choice of centers. For the radius objective we allow to choose the center of a cluster $C \subset \mathcal{P}$ from the whole metric space $\mathcal{X}$, while for the discrete radius objective the center must be contained in C itself. We will see that this has an impact on the price of hierarchy. For all three objectives the algorithms in [2, 3] compute an 8-approximate hierarchical clustering in polynomial time. Until recently this was also the best known upper bound for the price of hierarchy in the literature for hierarchical radius and diameter. For discrete radius, Großwendt [6] shows an upper bound for the price of hierarchy of 4. The best known lower bounds are 2, proven by Das and Kenyon-Mathieu [7] for diameter and by Großwendt [6] for (discrete) radius. We improve the framework in [5] for radius and diameter and show an upper bound on the price of hierarchy of $3 + 2\sqrt{2} \approx 5.83$. The upper bound of $3 + 2\sqrt{2}$ for the radius was also recently proved by Bock [8] in independent work. However our main contribution lies in the design of clustering instances to prove a lower bound of 4 for discrete radius and $3 + 2\sqrt{2}$ for radius and diameter.

1.1 Related Work

Gonzales [9] presents a simple and elegant incremental algorithm for k-center. The algorithm exhibits the following nice property: given a set $\mathcal{P}$ which has to be clustered, it returns an ordering of the points, such that the first k points constitute the centers of the k-center solution, and this solution is a 2-approximation for every $1 \leq k \leq |\mathcal{P}|$. However the resulting clusterings are usually not hierarchically compatible. Dasgupta and Long [2] use the ordering computed by Gonzales' algorithm

to compute a hierarchical clustering. The authors present an 8-approximation for the objective functions (discrete) radius and diameter. In an independent work Charikar et al. [3] also present an 8-approximation for the three objectives which outputs the same clustering as the algorithm in [2] under some reasonable conditions [7]. In a recent work, Mondal [10] gives a 6-approximation for hierarchical (discrete) radius. In Appendix A we present an instance where this algorithm computes only a 7-approximation contradicting the claimed guarantee.

Plaxton [4] shows that a similar approach as in [2] yields a hierarchical clustering with constant approximation guarantee for the *k*-median and *k*-means objectives. Later a general framework for a variety of incremental and hierarchical problems was introduced by Lin et al. [5]. Their framework can be applied to compute hierarchical clusterings for any cost function which satisfies a certain nesting property, especially those of *k*-median and *k*-means. This yields a 20.71α-approximation for *k*-median and a 576β-approximation for *k*-means. Here $\alpha = 2.67059$ and $\beta = 5.912$ are the currently best approximation guarantees for *k*-median [11] and *k*-means [12]. The algorithms presented in [2–5] run in polynomial time. Unless P=NP there is no polynomial-time α-approximation for $\alpha < 2$ for hierarchical (discrete) radius and diameter. For (discrete) radius this is an immediate consequence of the reduction from dominating set presented by [13]. A similar reduction from clique cover yields the statement for hierarchical diameter.

However even without time constraints it is not clear what approximation guarantee can be achieved for hierarchical clustering. It is easy to find examples, where the approximation guarantee of any hierarchical clustering for all three objectives is greater than one. Das and Kenyon-Mathieu [7] and Großwendt [6] present instances for diameter and (discrete) radius, where no hierarchical clustering has an approximation guarantee smaller than 2. On the other hand Großwendt [6] proves an upper bound of 4 on the approximation guarantee of hierarchical discrete radius by using the framework of Lin et al. [5]. In recent independent work Bock [8] improved the bound for hierarchical radius to $3 + 2\sqrt{2}$. While his approach is inspired by Dasgupta and Long [2], the resulting algorithm is similar to the algorithm we present in this paper as an improvement of [5].

Aside from the theoretical results, there also exist greedy heuristics, which are more commonly used in applications. One very simple bottom up, also called agglomerative, algorithm is the following: starting from the clustering where every point is separate, it merges in every step the two clusters whose merge results in the smallest increase of the cost function. For (discrete) radius and diameter this algorithm is known as complete linkage and for the *k*-means cost this is Ward's method [14]. Ackermann et al. [15] analyze the approximation guarantee of complete linkage in the Euclidean space. They show an approximation guarantee of $O(\log(k))$ for all three objectives assuming the dimension of the Euclidean space to be constant. This was later improved by Großwendt and Röglin [16] to $O(1)$. In arbitrary metric spaces complete linkage does not perform well. There Arutyunova et al. [17] prove a lower bound of $\Omega(k)$ for all three objectives. For Ward's method Großwendt et al. [18]

show an approximation guarantee of 2 under the strong assumption that the optimal clusters are well separated.

Recently other cost functions for hierarchical clustering were proposed, which do not compare to the optimal clustering on every level. Dasgupta [19] defines a new cost function for similarity measures and presents an $O(\alpha \log(n))$-approximation for the respective problem. This was later improved to $O(\alpha)$ independently by Charikar and Chatziafratis [20] and Cohen-Addad et al. [21]. Here α is the approximation guarantee of sparsest cut. However Cohen-Addad et al. [21] prove that every hierarchical clustering is an $O(1)$-approximation to the corresponding cost function for dissimilarity measures when the dissimilarity measure is a metric. A cost function more suitable for Euclidean spaces was developed by Wang and Moseley [22]. They prove that a randomly generated hierarchical clustering performs poorly for this cost function and show that bisecting k-means computes an $O(1)$-approximation.

2 Results

We define the *price of hierarchy* ρ_{cost} with respect to an objective function cost as the smallest number such that for every clustering instance there exists a hierarchical clustering which is a ρ_{cost}-approximation with respect to cost. Observe that the results [2, 3, 6, 7] imply that the price of hierarchy for radius and diameter is between 2 and 8 and for discrete radius between 2 and 4. We close these gaps and prove that the price of hierarchy for radius and diameter is exactly $3 + 2\sqrt{2}$ and for discrete radius exactly 4. Notice that this does not imply the existence of polynomial-time algorithms with approximation guarantee ρ_{cost}. Especially our algorithm which computes a $3 + 2\sqrt{2}$-approximation for radius and diameter does not run in polynomial time. This is also the case for the $3 + 2\sqrt{2}$-approximation for radius presented by Bock [8] in independent work. Our upper bound of $3 + 2\sqrt{2}$ can be achieved by a small improvement in the framework of Lin et al. [5]. However our most technically demanding contribution is the design of a clustering instance for every $\epsilon > 0$ such that every hierarchical clustering has approximation guarantee at least $3 + 2\sqrt{2} - \epsilon$ for radius and diameter and $4 - \epsilon$ for discrete radius. It requires a careful analysis of all possible hierarchical clusterings, which is highly non-trivial for complex clustering instances.

3 Preliminaries

A clustering instance $(\mathcal{X}, \mathcal{P}, d)$ consists of a metric space $(\mathcal{X}, d)$ and a finite subset $\mathcal{P} \subset \mathcal{X}$. For a set (or cluster) $C \subset \mathcal{P}$ we denote by

$$\mathsf{diam}(C) = \max_{p,q \in C} d(p, q)$$

the *diameter* of C. By $\mathsf{rad}(C, c) = \max_{p \in C} d(c, p)$ we denote the radius of C with respect to a center $c \in \mathcal{X}$. This is the largest distance between c and a point in C. The *radius* of C is defined as the smallest radius of C with respect to a center $c \in \mathcal{X}$, i.e.,

$$\mathsf{rad}(C) = \min_{c \in \mathcal{X}} \mathsf{rad}(C, c)$$

while the *discrete radius* of C is defined as the smallest radius of C with respect to a center $c \in C$, i.e.,

$$\mathsf{drad}(C) = \min_{c \in C} \mathsf{rad}(C, c).$$

A k-clustering of $\mathcal{P}$ is a partition of $\mathcal{P}$ into at most k non-empty subsets. We consider three closely related clustering problems.

The k-diameter problem asks to minimize $\mathsf{diam}(\mathcal{C}_k) = \max_{C \in \mathcal{C}_k} \mathsf{diam}(C)$, i.e., the maximum diameter of a k-clustering $\mathcal{C}_k$. In the k-center problem we want to minimize the maximum radius $\mathsf{rad}(\mathcal{C}_k) = \max_{C \in \mathcal{C}_k} \mathsf{rad}(C)$, and in the discrete k-center problem we want to minimize the maximum discrete radius $\mathsf{drad}(\mathcal{C}_k) = \max_{C \in \mathcal{C}_k} \mathsf{drad}(C)$.

Definition 1 Given an instance $(\mathcal{X}, \mathcal{P}, d)$, let $n = |\mathcal{P}|$. We call two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of $\mathcal{P}$ with $|\mathcal{C}| \geq |\mathcal{C}'|$ hierarchically compatible if for all $C \in \mathcal{C}$ there exists $C' \in \mathcal{C}'$ with $C \subset C'$. A *hierarchical clustering* of $\mathcal{P}$ is a sequence of clusterings $\mathscr{H} = (\mathcal{H}_n, \ldots, \mathcal{H}_1)$, such that

1. $\mathcal{H}_i$ is an i-clustering of $\mathcal{P}$
2. for $1 < i \leq n$ the two clusterings $\mathcal{H}_{i-1}$ and $\mathcal{H}_i$ are hierarchically compatible.

For $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}, \mathsf{drad}\}$ let $\mathcal{O}_i$ denote the optimal i-clustering with respect to cost. We say that $\mathscr{H}$ is an α-approximation with respect to cost if for all $i = 1, \ldots, n$ we have

$$\mathsf{cost}(\mathcal{H}_i) \leq \alpha \cdot \mathsf{cost}(\mathcal{O}_i).$$

Since optimal clusterings are generally not hierarchically compatible, there is usually no hierarchical clustering with approximation guarantee $\alpha = 1$. We have to accept that the restriction on hierarchically compatible clusterings comes with an unavoidable increase in the cost compared to an optimal solution.

Definition 2 For $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}, \mathsf{drad}\}$ the *price of hierarchy* $\rho_{\mathsf{cost}} \geq 1$ is defined as follows.

1. For every instance $(\mathcal{X}, \mathcal{P}, d)$, there exists a hierarchical clustering $\mathscr{H}$ of $\mathcal{P}$ that is a ρ_{cost}-approximation with respect to cost.

2. For any $\alpha < \rho_{\mathsf{cost}}$ there exists an instance $(\mathcal{X}, \mathcal{P}, d)$, such that there is no hierarchical clustering of $\mathcal{P}$ that is an α-approximation with respect to cost.

Thus ρ_{cost} is the smallest possible number such that for every clustering instance there is a hierarchical clustering with approximation guarantee ρ_{cost}.

4 An Upper Bound on the Price of Hierarchy

The framework by by Lin et al. [5] can be applied to compute incremental and hierarchical solutions to a large class of minimization problems. We first discuss their framework in the context of hierarchical clustering for (discrete) radius and diameter. In the second part we then present an improved version of their algorithm for radius and diameter.

First we introduce the notion of a *hierarchical sequence*, which is a relaxation of a hierarchical clustering in the sense that it does not have to contain a *k*-clustering for every $1 \leq k \leq |\mathcal{P}|$.

Definition 3 Given an instance $(\mathcal{X}, \mathcal{P}, d)$, with $n = |\mathcal{P}|$. We call a sequence $\mathscr{C} = (\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)})$ of clusterings a hierarchical sequence if it satisfies

1. $|\mathcal{C}^{(t)}| = n$ and $|\mathcal{C}^{(1)}| = 1$
2. for $1 \leq i \leq t$ either $\mathcal{C}^{(i-1)} = \mathcal{C}^{(i)}$ or $\mathcal{C}^{(i-1)}$ is obtained from $\mathcal{C}^{(i)}$ by merging some of its clusters.

Such a hierarchical sequence can be extended to a hierarchical clustering of $\mathcal{P}$ as follows. We define the respective hierarchical clustering $h(\mathscr{C})$ by assigning every $1 \leq i \leq n$ the clustering among $\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)}$ of smallest cost and size at most *i*. We say that $\mathscr{C}$ is an α-approximation iff $h(\mathscr{C})$ is an α-approximation.
Before we are able to define the algorithm we need one important definition from [5].

Definition 4 Given an instance $(\mathcal{X}, \mathcal{P}, d)$. For $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}, \mathsf{drad}\}$ we say that the (γ, δ)-nesting property holds for reals $\gamma, \delta \geq 0$, if for any two clusterings $\mathcal{C}, \mathcal{D}$ of $\mathcal{P}$ with $|\mathcal{C}| > |\mathcal{D}|$ there exists a clustering $\mathcal{C}'$ with

1. $|\mathcal{C}'| \leq |\mathcal{D}|$
2. $\mathcal{C}'$ is hierarchically compatible with $\mathcal{C}$ and
3. $\mathsf{cost}(\mathcal{C}') \leq \gamma \mathsf{cost}(\mathcal{C}) + \delta \mathsf{cost}(\mathcal{D})$.

We say that $\mathcal{C}'$ is a *nesting* of $\mathcal{C}$ at $\mathcal{D}$. Let $\mathsf{Augment}_{\mathsf{cost}}(\mathcal{C}, \mathcal{D}, \gamma, \delta)$ denote the subroutine that computes such a clustering $\mathcal{C}'$.

Algorithm 1 (Lin et al. [5])

Require: Clustering instance $(\mathcal{X}, \mathcal{P}, d)$, with $d(x, y) > 2$ for all $x, y \in \mathcal{P}$, optimal clusterings $\mathcal{O}_{|\mathcal{P}|}, \ldots, \mathcal{O}_1$ of $\mathcal{P}$ with respect to cost
Ensure: A hierarchical clustering of $\mathcal{P}$
1: Set $\Delta = \mathsf{cost}(\mathcal{O}_1), t = \lceil \log_{2\gamma}(\Delta) \rceil + 1$ and $\mathcal{C}^{(t)} = \mathcal{O}_{|\mathcal{P}|}$
2: **for** $i = t - 1$ **to** 1 **do**
3: Let $1 \leq n_i \leq |\mathcal{P}|$ be the smallest number such that $\mathsf{cost}(\mathcal{O}_{n_i}) \in ((2\gamma)^{t-i-1}, (2\gamma)^{t-i}]$
4: **if** such a number exists **then**
5: set $\mathcal{C}^{(i)} = \mathsf{Augment}_{\mathsf{cost}}(\mathcal{C}^{(i+1)}, \mathcal{O}_{n_i}, \gamma, \delta)$
6: **else**
7: set $\mathcal{C}^{(i)} = \mathcal{C}^{(i+1)}$
8: **end if**
9: **end for**
10: **return** $h((\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)}))$

The algorithm of Lin et al. [5] is shown as Algorithm 1. It computes a hierarchical sequence $\mathscr{C} = (\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)})$ of clusterings as follows. Starting with $\mathcal{C}^{(t)} = \mathcal{O}_{|\mathcal{P}|}$ the algorithm builds the i-th clustering $\mathcal{C}^{(i)}$ as nesting of $\mathcal{C}^{(i+1)}$ at an optimal clustering $\mathcal{O}_{n_i}$. This guarantees that the clusterings are hierarchically compatible.

Theorem 1 [5] *For* $\mathsf{cost} \in \{\mathsf{drad}, \mathsf{rad}, \mathsf{diam}\}$, *if the* (γ, δ)*-nesting property holds for reals* $\gamma \geq 1, \delta > 0$, *then Algorithm* 1 *computes a hierarchical clustering of* $\mathcal{P}$ *with approximation guarantee* $4\gamma\delta$ *with respect to* cost.

Großwendt [6] proved the existence of such a nesting property for $\mathsf{diam}, \mathsf{rad}$, and drad.

Lemma 2 [6] *For* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ *there exists a* (2, 1)*-nesting and for* $\mathsf{cost} = \mathsf{drad}$ *there exists a* (1, 1)*-nesting.*

In combination with Theorem 1 this yields $\rho_{\mathsf{drad}} \leq 4$. However, for the other two objectives we obtain an upper bound of only 8. We improve Algorithm 1 to obtain the claimed upper bound of $3 + 2\sqrt{2}$.

In the definition of the (γ, δ)-nesting property we require a nesting of $\mathcal{C}$ at $\mathcal{D}$ for arbitrary clusterings $\mathcal{C}, \mathcal{D}$ with $|\mathcal{C}| > |\mathcal{D}|$. However, in Algorithm 1 we know more about the structure of $\mathcal{C}$. This clustering is obtained by repeatedly nesting at optimal clusterings of increasing cost. In Algorithm 2 we define a nesting subroutine for this type of clusterings that eventually leads to a better approximation-guarantee.

Algorithm 2

Require: Step size $\alpha > 1$. Clustering instance $(\mathcal{X}, \mathcal{P}, d)$, with $d(x, y) > 2$ for all $x, y \in \mathcal{P}$, optimal clusterings $\mathcal{O}_{|\mathcal{P}|}, \ldots, \mathcal{O}_1$ of $\mathcal{P}$ with respect to cost

Ensure: A hierarchical clustering of $\mathcal{P}$

Set $\Delta = \mathsf{cost}(\mathcal{O}_1), t = \lceil \log_\alpha(\Delta) \rceil + 1$ and $\mathcal{C}^{(t)} = \mathcal{O}_{|\mathcal{P}|}$
For all $C \in \mathcal{C}^{(t)}$ we set $\mathsf{parent}_t(C) = C$
for $i = t - 1$ **to** 1 **do**
  Let $1 \leq n_i \leq |\mathcal{P}|$ be the smallest number such that $\mathsf{cost}(\mathcal{O}_{n_i}) \in (\alpha^{t-i-1}, \alpha^{t-i}]$
  if such a number exists **then**
    For $C \in \mathcal{C}^{(i+1)}$ let $O \in \mathcal{O}_{n_i}$ be a cluster with $\mathsf{parent}_{i+1}(C) \cap O \neq \emptyset$ and set $\mathsf{Nest}_i(C) = O$
    Set $\mathcal{C}^{(i)} = \{\bigcup_{C \in \mathsf{Nest}_i^{(-1)}(O)} C \mid O \in \mathcal{O}_{n_i}\}$
    Set $\mathsf{parent}_i(\bigcup_{C \in Nest_i^{(-1)}(O)} C) = O$ for all $O \in \mathcal{O}_{n_i}$
  else
    Set $\mathcal{C}^{(i)} = \mathcal{C}^{(i+1)}, \mathsf{parent}_i = \mathsf{parent}_{i+1}$
  end if
end for
return $h((\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)}))$

The main difference between Algorithm 1 and Algorithm 2 is the replacement of the function $\mathsf{Augment}_{\mathsf{cost}}(\mathcal{C}_{i+1}, \mathcal{O}_{n_i}, \gamma, \delta)$, which computes the nesting of $\mathcal{C}^{(i+1)}$ at $\mathcal{O}_{n_i}$, by a more explicit approach to compute such a nesting. We use the fact that $\mathcal{C}^{(i+1)}$ is obtained by a nesting at $\mathcal{O}_{n_{i+1}}$. This is reflected in the function parent_{i+1} which assigns every cluster in $C^{(i+1)}$ a cluster from $\mathcal{O}_{n_{i+1}}$. In iteration i we then use the $(i+1)$-st parent function to determine which clusters of $\mathcal{C}^{(i+1)}$ will be merged to obtain $\mathcal{C}^{(i)}$. We are allowed to merge clusters $C, D \in \mathcal{C}^{(i+1)}$ if there is a cluster $O \in \mathcal{O}_{n_i}$ which has a non-empty intersection with both, $\mathsf{parent}_{i+1}(C)$ and $\mathsf{parent}_{i+1}(D)$. The parent of the merged cluster in $\mathcal{C}^{(i)}$ is then set to O.

Lemma 3 *For* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ *and any* $\alpha > 1$ *Algorithm 2 computes a hierarchical clustering with approximation guarantee* $\alpha\left(\frac{2}{\alpha-1} + 1\right)$.

Proof Let n denote the cardinality of $\mathcal{P}$. Notice first that $(\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)})$ is indeed a hierarchical sequence. The first property of a hierarchical sequence is satisfied: We define $\mathcal{C}^{(t)} = \mathcal{O}_n$ and since $\mathsf{cost}(\mathcal{O}_1) = \Delta \in (\alpha^{t-2}, \alpha^{t-1}]$ we obtain $|\mathcal{C}^{(1)}| \leq n_1 = 1$. The second property is satisfied since $\mathcal{C}^{(i)}$ either equals $\mathcal{C}^{(i+1)}$ or is obtained by merging clusters from $\mathcal{C}^{(i+1)}$. Thus Algorithm 2 indeed computes a hierarchical clustering.

Diameter ($\mathsf{cost} = \mathsf{diam}$): Let $1 \leq i \leq t$. We claim

1. for every cluster $C \in \mathcal{C}^{(i)}$ and every point $p \in \mathsf{parent}_i(C)$ that $\max_{q \in C} d(p, q) \leq \sum_{l=1}^{t-i} \alpha^l$,
2. that $\mathsf{diam}(\mathcal{C}^{(i)}) \leq \alpha^{t-i} + 2\sum_{l=1}^{t-i-1} \alpha^l$.

We prove this by induction over *i*, starting with $i = t$ in decreasing order. Observe that $\mathcal{C}^{(t)}$ consists only of clusters of size one so these claims are true for $i = t$.

Let $1 \leq i \leq t-1$. If $\mathcal{C}^{(i)} = \mathcal{C}^{(i+1)}$ both claims are true by induction hypothesis. Thus we assume from now on that $\mathcal{C}^{(i)} \neq \mathcal{C}^{(i+1)}$. For the first claim, we fix a cluster $C \in C^{(i)}$ and two points $p \in \mathsf{parent}_i(C)$ and $q \in C$. Let $D \in \mathcal{C}^{(i+1)}$ be the cluster which contains q. Since $\mathcal{C}^{(i)}$ is obtained by merging clusters from $\mathcal{C}^{(i+1)}$, we know that $D \subset C$ and thus $\mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C) \neq \emptyset$. Let $x \in \mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C)$. By the induction hypothesis

$$d(x,q) \leq \max_{y \in D} d(x,y) \leq \sum_{l=1}^{t-i-1} \alpha^l.$$

Since p and x lie both in $\mathsf{parent}_i(C)$ we obtain $d(p,x) \leq \mathsf{diam}(\mathcal{O}_{n_i}) \leq \alpha^{t-i}$. Using the triangle inequality we conclude

$$d(p,q) \leq d(p,x) + d(x,q) \leq \sum_{l=1}^{t-i} \alpha^l.$$

For the second claim we again fix a cluster $C \in \mathcal{C}^{(i)}$ and two points $p, q \in C$. Let $B, D \in \mathcal{C}^{(i+1)}$ such that $p \in B$ and $q \in D$. Observe that $B \cup D \subset C$ and thus $\mathsf{parent}_{i+1}(B) \cap \mathsf{parent}_i(C) \neq \emptyset \neq \mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C)$. Let $x_p \in \mathsf{parent}_{i+1}(B) \cap \mathsf{parent}_i(C)$ and $x_q \in \mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C)$. Since x_p and x_q lie both in $\mathsf{parent}_i(C)$ we obtain $d(x_p, x_q) \leq \mathsf{diam}(\mathcal{O}_{n_i}) \leq \alpha^{t-i}$. We apply the triangle inequality and the induction hypothesis to obtain

$$d(p,q) \leq d(p,x_p) + d(x_p,x_q) + d(x_q,q) \leq \alpha^{t-i} + 2\sum_{l=1}^{t-i-1} \alpha^l.$$

Radius ($\mathsf{cost} = \mathsf{rad}$): Let $1 \leq i \leq t$. We claim that for every cluster $C \in \mathcal{C}^{(i)}$ and the center c of cluster $\mathsf{parent}_i(C)$ holds $\max_{q \in C} d(c,q) \leq \alpha^{t-i} + 2\sum_{l=1}^{t-i-1} \alpha^l$. Notice that this immediately implies

$$\mathsf{rad}(\mathcal{C}^{(i)}) \leq \alpha^{t-i} + 2\sum_{l=1}^{t-i-1} \alpha^l.$$

We prove this by induction over *i*. Observe that $\mathcal{C}^{(t)}$ consists only of clusters of size one. So this claim is true for $i = t$. Let $1 \leq i \leq t-1$. If $\mathcal{C}^{(i)} = \mathcal{C}^{(i+1)}$ the claim is true by induction hypothesis. Thus we assume from now on that $\mathcal{C}^{(i)} \neq \mathcal{C}^{(i+1)}$. We fix a cluster $C \in C^{(i)}$ a point $q \in C$ and denote by c the center of $\mathsf{parent}_i(C)$. Let $D \in \mathcal{C}^{(i+1)}$ be the cluster which contains q. Since $\mathcal{C}^{(i)}$ is obtained by merging clusters from $\mathcal{C}^{(i+1)}$, we know that $D \subset C$ and thus $\mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C) \neq \emptyset$. Let $x \in \mathsf{parent}_{i+1}(D) \cap \mathsf{parent}_i(C)$. By induction hypothesis the following holds for the center d of $\mathsf{parent}_{i+1}(D)$

$$\max_{v \in D} d(d, v) \leq \alpha^{t-i-1} + 2 \sum_{l=1}^{t-i-2} \alpha^l.$$

Together with the triangle inequality this implies

$$d(x, q) \leq d(x, d) + d(d, q) \leq \mathsf{rad}(\mathcal{O}_{n_{i+1}}) + \alpha^{t-i-1} + 2 \sum_{l=1}^{t-i-2} \alpha^l \leq 2 \sum_{l=1}^{t-i-1} \alpha^l.$$

This yields the claim, as

$$d(c, q) \leq d(c, x) + d(x, q) \leq \mathsf{rad}(\mathcal{O}_{n_i}) + 2 \sum_{l=1}^{t-i-1} \alpha^l \leq \alpha^{t-i} + 2 \sum_{l=1}^{t-i-1} \alpha^l.$$

Finally we can bound the approximation factor for both radius and diameter. Let $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$. Since $d(x, y) > 2$ for all $x, y \in \mathcal{P}$ we get that $\mathsf{cost}(\mathcal{O}_{n-1}) > 1$. Thus for every $1 \leq m < n$ there is $1 \leq i \leq t-1$ such that $\mathsf{cost}(\mathcal{O}_m) \in (\alpha^{t-i-1}, \alpha^{t-i}]$. Thus the clustering $h((\mathcal{C}^{(t)}, \ldots, \mathcal{C}^{(1)}))$ is an $\alpha\left(\frac{2}{\alpha-1} + 1\right)$-approximation iff for all $1 \leq i \leq t$

$$\mathsf{cost}(\mathcal{C}^{(i)}) \leq \alpha\Big(\frac{2}{\alpha - 1} + 1\Big)\mathsf{cost}(\mathcal{O})$$

for all optimal clusterings $\mathcal{O}$ with $\mathsf{cost}(\mathcal{O}) \in (\alpha^{t-i-1}, \alpha^{t-i}]$. We obtain

$$\begin{aligned}\mathsf{cost}(\mathcal{C}^{(i)}) \leq & \alpha^{t-i} + 2 \sum_{l=1}^{t-i-1} \alpha^l < \alpha^{t-i} + 2 \cdot \frac{\alpha^{t-i}}{\alpha - 1} = \alpha^{t-i}\Big(\frac{2}{\alpha - 1} + 1\Big) \\ \leq & \alpha\Big(\frac{2}{\alpha - 1} + 1\Big)\mathsf{cost}(\mathcal{O}).\end{aligned}$$

□

Theorem 4 *For* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ *we have* $\rho_{\mathsf{cost}} \leq 3 + 2\sqrt{2} \approx 5.828$.

Proof Let $(\mathcal{X}, \mathcal{P}, d)$ be a clustering instance. We can assume without loss of generality that $d(x, y) > 2$ for all $x, y \in \mathcal{P}$, otherwise we scale the metric d accordingly. For $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ we then use Algorithm 2 with $\alpha = 1 + \sqrt{2}$ to compute a hierarchical clustering. By Lemma 3 we obtain a hierarchical clustering that is an $3 + 2\sqrt{2}$ approximation and thus $\rho_{\mathsf{cost}} \leq 3 + 2\sqrt{2}$. □

5 A Lower Bound on the Price of Hierarchy

The most challenging contributions of this article are matching lower bounds on the price of hierarchy for diameter, radius, and discrete radius.

Theorem 5 *For* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ *we have* $\rho_{\mathsf{cost}} \geq 3 + 2\sqrt{2}$ *and for* $\mathsf{cost} = \mathsf{drad}$ *we have* $\rho_{\mathsf{cost}} \geq 4$.

There is already existing work in this area by Das and Kenyon-Mathieu [7] for the diameter and Großwendt [6] for the radius. Both show a lower bound of 2 for the respective objective. To improve upon these results we have to construct much more complex instances which differ significantly from those in [6, 7].

For every $\epsilon > 0$ we will construct a clustering instance $(\mathcal{X}, \mathcal{P}, d)$ such that for any hierarchical clustering $\mathscr{H} = (\mathcal{H}_{|\mathcal{P}|}, \ldots, \mathcal{H}_1)$ of $\mathcal{P}$ there is $1 \leq i \leq |\mathcal{P}|$ such that $\mathsf{cost}(\mathcal{H}_i) \geq \alpha \cdot \mathsf{cost}(\mathcal{O}_i)$, where $\mathcal{O}_i$ is an optimal i-clustering of $\mathcal{P}$ with respect to cost and $\alpha = (3 + 2\sqrt{2} - \epsilon)$ for $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ and $\alpha = 4 - \epsilon$ for $\mathsf{cost} = \mathsf{drad}$.

The proof is divided in three parts. First we introduce the clustering instance $(\mathcal{X}, \mathcal{P}, d)$ and determine its optimal clusterings. In the second part we develop the notion of a *bad* cluster. We prove that any hierarchical clustering contains such bad clusters and develop a lower bound on their cost. In the third part we compare the lower bound to the cost of optimal clusterings and prove Theorem 5.

5.1 Definition of the Clustering Instance

For $n \in \mathbb{N}$ we denote by $[n]$ the set of numbers from 1 to n.

Let $k \in \mathbb{N}$ and $\Gamma = k + 1$. For $0 \leq \ell \leq k$ we define point sets $\mathcal{Q}_\ell$ and $\mathcal{P}_\ell$ recursively as follows

1. For $\ell = 0$ let $\mathcal{P}_0 = \mathcal{Q}_0 = [1]$ and denote by N_0 the cardinality of $\mathcal{P}_0$.
2. For $\ell > 0$ let $\mathcal{Q}_\ell = [\Gamma \cdot N_{\ell-1}]^{N_{\ell-1}}$ and $\mathcal{P}_\ell = \prod_{i=0}^{\ell} \mathcal{Q}_i$. Furthermore set $N_\ell = |\mathcal{P}_\ell|$.

Moreover let $\phi_\ell : \mathcal{P}_\ell \to [N_\ell]$ be a bijection for $0 \leq \ell \leq k$.

We refer to a point $X \in \mathcal{P}_k$ as a matrix with $k + 1$ rows and $N_{\ell-1}$ entries in the ℓ -th row. Thus we write

$$X = (x_{01} \mid \ldots \mid x_{\ell 1}, \ldots, x_{\ell N_{\ell-1}} \mid \ldots \mid x_{k1}, \ldots, x_{kN_{k-1}}).$$

Let $X_\ell = (x_{\ell 1}, \ldots, x_{\ell N_{\ell-1}}) \in \mathcal{Q}_\ell$ for $0 \leq \ell \leq k$. For a shorter representation we can replace the ℓ-th row directly by X_ℓ and for $0 \leq i \leq j \leq k$ we can replace the i-th up to j-th row by $X_{[i:j]} = (X_i \mid \ldots \mid X_j)$.

Let $X \in \mathcal{P}_k$ and $1 \leq \ell \leq k$. Notice that $X_{[0:\ell-1]} \in \mathcal{P}_{\ell-1}$ and let $m = \phi_{\ell-1}(X_{[0:\ell-1]})$, we define

$$A_\ell^X = \{(X_{[0:\ell-1]} \mid x_{\ell 1}, \ldots, x_{\ell m-1}, \star, x_{\ell m+1}, \ldots, x_{\ell N_{\ell-1}} \mid X_{[\ell+1:k]}) \mid \star \in [\Gamma \cdot N_{\ell-1}]\}.$$

Thus all coordinates of points in A_ℓ^X are fixed and agree with those of X except one which is variable. Here $X_{[0:\ell-1]}$ serves as prefix which indicates through $\phi_{\ell-1}$ which coordinate of X_ℓ can be changed.

We define $\mathcal{A}_\ell = \{A_\ell^X \mid X \in \mathcal{P}_k\}$ as the set containing all subsets of this form. It is clear that $\mathcal{A}_\ell$ is a partition of $\mathcal{P}_k$ and that it contains only sets of size $\Gamma \cdot N_{\ell-1}$. Furthermore we set $\mathcal{A}_0 = \{\{X\} \mid X \in \mathcal{P}_k\}$.

Example 1 If we perform the first three steps of the construction we get $\mathcal{Q}_0 = [1], \mathcal{Q}_1 = [\Gamma], Q_2 = [\Gamma^2]^\Gamma$ and

$$\mathcal{P}_1 = \{(1 \mid x_{11}) \mid x_{11} \in [\Gamma]\},$$
$$\mathcal{P}_2 = \{(1 \mid x_{11} \mid x_{21}, \ldots, x_{2\Gamma}) \mid x_{11} \in [\Gamma], x_{2i} \in [\Gamma^2] \text{ for } 1 \leq i \leq \Gamma\}$$

Since ϕ_0 is a map between two sets of cardinality one this map is always unique. Now suppose that we picked ϕ_1 such that $\phi_1((x_{01} \mid x_{11})) = x_{11}$ for all $(x_{01} \mid x_{11}) \in \mathcal{P}_1$. Then the partition $\mathcal{A}_1$ consists of the sets

$$\{(1 \mid \star \mid x_{21}, \ldots, x_{2\Gamma}) \mid \star \in [\Gamma]\}$$

with $x_{2i} \in [\Gamma^2]$ for all $1 \leq i \leq \Gamma$. The partition $\mathcal{A}_2$ consists of the sets

$$\{(1 \mid x_{11} \mid x_{21}, \ldots, x_{2x_{11}-1}, \star, x_{2x_{11}+1}, \ldots, x_{2\Gamma}) \mid \star \in [\Gamma^2]\}$$

with $x_{11} \in [\Gamma]$ and $x_{2i} \in [\Gamma^2]$ for all $1 \leq i \leq \Gamma$ with $i \neq x_{11}$.
Let $G = (V, E, w)$ denote the weighted hyper-graph with $V = \mathcal{P}_k$ and $E = \bigcup_{i=1}^k \mathcal{A}_i$. The weight of a hyper-edge $e \in E$ is set to ℓ iff $e \in \mathcal{A}_\ell$. For $0 \leq \ell \leq k$, the sub-graph $G_\ell = (V_\ell, E_\ell, w_\ell)$ is given by $V_\ell = \mathcal{P}_k, E_\ell = \bigcup_{i=0}^\ell \mathcal{A}_i$ and $w_\ell = w_{|E_\ell}$.

We extend G to a hyper-graph $H = (V', E', w')$ as follows. Let $V' = V \cup \bigcup_{i=0}^k \{v_A \mid A \in \mathcal{A}_i\}$ and $E' = E \cup \bigcup_{i=0}^k \{\{v, v_A\} \mid A \in \mathcal{A}_i, v \in A\}$. Thus H contains one vertex for every $A \in \bigcup_{i=0}^k \mathcal{A}_i$ and this vertex is connected by edges to every vertex $v \in A$. For $e \in E$ we set $w'(e) = w(e)$ and for $e = \{v, v_A\}$ for some $A \in \mathcal{A}_\ell$ and $v \in A$ we set $w'(e) = \ell/2$.

The clustering instance $(\mathcal{X}, \mathcal{P}, d)$ is given by $\mathcal{X} = V', \mathcal{P} = V$, and d as the shortest path metric on H. Observe that the extension of G to H is only necessary for the lower bound for the radius but not for the diameter and the discrete radius. This is because the additional points $V' \setminus V$ do not belong to $\mathcal{P}$ and are hence irrelevant for the clustering instance for the diameter and discrete radius. In the lower bound for the radius they will be used as centers, however.

Lemma 6 *Let $p, q \in V$, then $d(p, q)$ is the length of a shortest path between p and q in G.*

Proof By definition $d(p, q)$ is the length of a shortest path between p and q in H. Suppose the shortest path contains a vertex v_A for some $A \in \bigcup_{i=0}^k \mathcal{A}_i$ with $v \in A$ as

predecessor and $w \in A$ as ancestor. Since v and w are connected in H by the hyperedge A we can delete v_A from the path and the length of the path does not change. The resulting path is also a path in G, so $d(p, q)$ is also the length of a shortest path between p and q in G.

Next we state some structural properties of the graph G and the clustering instance $(\mathcal{X}, \mathcal{P}, d)$. To establish a lower bound on the approximation factor of a hierarchical clustering we first focus on the optimal clusterings of the instance $(\mathcal{X}, \mathcal{P}, d)$. One can already guess that $\mathcal{A}_\ell$ is an optimal clustering with $\frac{N_k}{\Gamma N_{\ell-1}}$ clusters with respect to $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}, \mathsf{drad}\}$ and we will prove this in this section. First we need the following statement about the connected components of G_ℓ.

Lemma 7 *The vertex set of every connected component in $G_\ell$has cardinality $N_\ell$and is of the form $V_\ell^X = \{(X' \mid X) \mid X' \in \mathcal{P}_\ell\}$for a given $X = (X_{\ell+1} \mid \ldots \mid X_k) \in \prod_{i=\ell+1}^k \mathcal{Q}_i$.*

Proof Notice that $|V_\ell^X| = N_\ell$ and that $\{V_\ell^X \mid X \in \prod_{i=\ell+1}^k \mathcal{Q}_i\}$ is a partition of V. Furthermore since $E_\ell = \bigcup_{i=0}^{\ell} \mathcal{A}_i$ any edge $e \in E_\ell$ is either completely contained in or disjoint to V_ℓ^X.

It is left to show that V_ℓ^X is connected. We prove this via induction over ℓ. For $\ell = 0$ this is clear because $|V_0^X| = 1$. For $\ell > 0$ let $Y = (Y_\ell \mid X), Z = (Z_\ell \mid X) \in \prod_{i=\ell}^k \mathcal{Q}_i$. By the induction hypothesis we know that the sets $V_{\ell-1}^Y, V_{\ell-1}^Z$ are connected. To prove that V_ℓ^X is connected it is sufficient to show that there is a path from a point in $V_{\ell-1}^Y$ to a point in $V_{\ell-1}^Z$. We show this claim by induction over the number m of coordinates in which Y and Z differ. For $m = 0$ there is nothing to show. If $m > 0$ pick $1 \leq s \leq N_{\ell-1}$ such that $y_{\ell s} \neq z_{\ell s}$ and let $P = \phi_{\ell-1}^{-1}(s) \in \prod_{i=0}^{\ell-1} \mathcal{Q}_i$. Consider the point $(P \mid Y_\ell \mid X)$ which is contained in $V_{\ell-1}^Y$. This point is also contained in the set

$$\{(P \mid y_{\ell 1}, \ldots, y_{\ell s-1}, \star, y_{\ell s+1}, \ldots, y_{\ell N_{\ell-1}} \mid X) \mid \star \in [\Gamma \cdot N_{\ell-1}]\} \in E_\ell.$$

Thus there is an edge in G_ℓ connecting a point in $V_{\ell-1}^Y$ to a point in $V_{\ell-1}^{Y'}$ with $Y' = (y_{\ell 1}, \ldots, y_{\ell s-1}, z_{\ell s}, y_{\ell s+1}, \ldots, y_{N_{\ell-1}} \mid X)$. Now Y' and Z differ in $m-1$ coordinates, thus there is a path between two points in $V_{\ell-1}^{Y'}$ and $V_{\ell-1}^Z$ by induction hypothesis. If we combine this with the induction hypothesis that $V_{\ell-1}^{Y'}$ is connected this yields the claim (see Fig. 1 for an illustration). □

Lemma 8 *Any clustering of $(\mathcal{X}, \mathcal{P}, d)$ with less than $\frac{N_k}{N_{\ell-1}}$ clusters costs at least ℓ if $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{drad}\}$ and $\ell/2$ if $\mathsf{cost} = \mathsf{rad}$.*

Proof The shortest path in G between any two points which lie in different connected components of $G_{\ell-1}$ must contain an edge of weight $\geq \ell$. Thus any set of points $M \subset V$ which is disconnected in $G_{\ell-1}$ has diameter $\geq \ell$. Remember that the

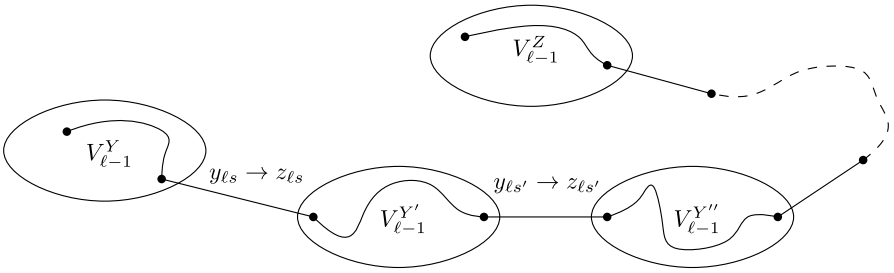

Fig. 1 Here we see the construction of the path. It corresponds to changing the coordinates of Y successively until they match Z. We use an edge in $\mathcal{A}_\ell$ to change y_{ls} to z_{ls}, next we change $y_{ls'}$ to $z_{ls'}$ and proceed like this until we obtain Z. The respective edges are then connected to a path from $V^X_{\ell-1}$ to $V^Z_{\ell-1}$

discrete radius of M is given by $\mathsf{drad}(M) = \min_{c\in M} \max_{p\in M} d(p,c)$. For every possible choice of $c \in M$ there exists a point $p \in M$ which is not in the same connected component of $G_{\ell-1}$ as c, thus $d(c,p) \geq \ell$ and therefore $\mathsf{drad}(M) \geq \ell$ and $\mathsf{rad}(M) \geq \mathsf{diam}(M)/2 \geq \ell/2$.

We conclude that if $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{drad}\}$ any cluster of cost smaller than ℓ is contained in one of the sets $V^X_{\ell-1}$ for some $X \in \prod_{i=\ell}^k \mathcal{Q}_i$ by Lemma 7 and any clustering with less than $\left|\prod_{i=\ell}^k \mathcal{Q}_i\right|$ clusters costs at least ℓ. By the same argument if $\mathsf{cost} = \mathsf{rad}$ any cluster of cost smaller than $\ell/2$ is contained in one of the sets $V^X_{\ell-1}$ for some $X \in \prod_{i=\ell}^k \mathcal{Q}_i$ by Lemma 7 and any clustering with less than $\left|\prod_{i=\ell}^k \mathcal{Q}_i\right|$ clusters costs at least $\ell/2$. Since

$$\left|\prod_{i=\ell}^{k} \mathcal{Q}_i\right| = \frac{\left|\prod_{i=0}^{k} \mathcal{Q}_i\right|}{\left|\prod_{i=0}^{\ell-1} \mathcal{Q}_i\right|} = \frac{N_k}{N_{\ell-1}}$$

this proves the lemma. □

Corollary 9 *For* $1 \leq \ell \leq k$ *and* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}, \mathsf{drad}\}$ *the clustering* $\mathcal{A}_\ell$ *is an optimal* $\frac{N_k}{\Gamma N_{\ell-1}}$*-clustering for the instance* $(\mathcal{X}, \mathcal{P}, d)$. *Furthermore* $\mathsf{diam}(\mathcal{A}_\ell) = \mathsf{drad}(\mathcal{A}_\ell) = \ell$ *and* $\mathsf{rad}(\mathcal{A}_\ell) = \ell/2$.

Proof If $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{drad}\}$ we obtain by definition of $(\mathcal{X}, \mathcal{P}, d)$ that $\mathsf{cost}(\mathcal{A}_\ell) \leq \ell$. If $\mathsf{cost} = \mathsf{rad}$ we obtain that $\mathsf{cost}(\mathcal{A}) \leq \ell/2$ by picking $v_A \in \mathcal{X}\backslash\mathcal{P}$ as center for $A \in \mathcal{A}_\ell$. On the other hand $|\mathcal{A}_\ell| = \frac{N_k}{\Gamma N_{\ell-1}} < \frac{N_k}{N_{\ell-1}}$ and thus $\mathsf{cost}(\mathcal{A}_\ell) \geq \ell$ if $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{drad}\}$ and $\mathsf{cost}(\mathcal{A}_\ell) \geq \ell/2$ for $\mathsf{cost} = \mathsf{rad}$ by Lemma 8. □

5.2 Characterization of Hierarchical Clusterings

Let from now on $\mathscr{H} = (\mathcal{H}_{N_k}, \ldots, \mathcal{H}_1)$ denote a hierarchical clustering of $(\mathcal{X}, \mathcal{P}, d)$. We introduce the notion of *bad clusters* in $\mathcal{H}_{\frac{N_k}{\Gamma N_{\ell-1}}}$ which are clusters whose cost increases repeatedly, as we will see later. In this section we prove the existence of such clusters in $\mathscr{H}$ and we give a lower bound on their cost.

Definition 5 We call all clusters $C \in \mathcal{H}_{N_k}$ bad at time 0 and denote by $\mathsf{Ker}_0(C) = C$ the kernel of C at time 0 and set $\mathsf{Bad}(0) = \mathcal{H}_{N_k}$.

For $1 \leq \ell \leq k$ we say that a cluster $C \in \mathcal{H}_{\frac{N_k}{\Gamma N_{\ell-1}}}$ is anchored at $\ell \leq \ell' \leq k$ if the set $\bigcup_{D \in \mathsf{Bad}(\ell-1): D \subset C} \mathsf{Ker}_{\ell-1}(D)$ is

1. connected in $G_{\ell'}$,
2. disconnected in $G_{\ell'-1}$.

We call C bad at time ℓ if C is anchored at some $\ell' \geq \ell$. We denote by $\mathsf{Bad}(\ell) \subset \mathcal{H}_{\frac{N_k}{\Gamma N_{\ell-1}}}$ the set of all bad clusters at time ℓ. If C is bad we define the kernel of C as the union of all kernels of bad clusters at time $\ell - 1$ contained in C, i.e.,

$$\mathsf{Ker}_\ell(C) = \bigcup_{D \in \mathsf{Bad}(\ell-1): D \subset C} \mathsf{Ker}_{\ell-1}(D).$$

All clusters in $\mathcal{H}_{\frac{N_k}{\Gamma N_{\ell-1}}} \setminus \mathsf{Bad}(\ell)$ are called good.

The example in Fig. 2 shows that a bad cluster at time ℓ can contain clusters which are good at time $\ell - 1$. However we are only interested in points that are contained exclusively in bad clusters at any time $t < \ell$. The set $\mathsf{Ker}_\ell(C)$ contains exactly such points.

We will use two crucial properties to prove the final lower bound on the approximation factor of any hierarchical clustering $\mathscr{H}$ of $(\mathcal{X}, \mathcal{P}, d)$. We first observe that bad clusters exist in $\mathscr{H}$ for every time-step $1 \leq \ell \leq k$ and second that these clusters have a large cost compared to the optimal clustering.

Lemma 10 *LetCbe a good cluster at time* $1 \leq \ell \leq k$ *and*

$$W = \bigcup_{D \in \mathsf{Bad}(\ell-1): D \subset C} \mathsf{Ker}_{\ell-1}(D),$$

thenWis connected in $G_{\ell-1}$ *and thus* $|W| \leq N_{\ell-1}$.

Proof Suppose W is disconnected in $G_{\ell-1}$. Since $G_k = G$ is connected, there must be a time $\ell' \geq \ell$ such that W is connected in $G_{\ell'}$ and disconnected in $G_{\ell'-1}$. But then C is a bad cluster at time ℓ which is anchored at ℓ' in contradiction to our assumption. Thus W is connected in $G_{\ell-1}$. By Lemma 7 we know that every connected component in $G_{\ell-1}$ is of size $N_{\ell-1}$. □

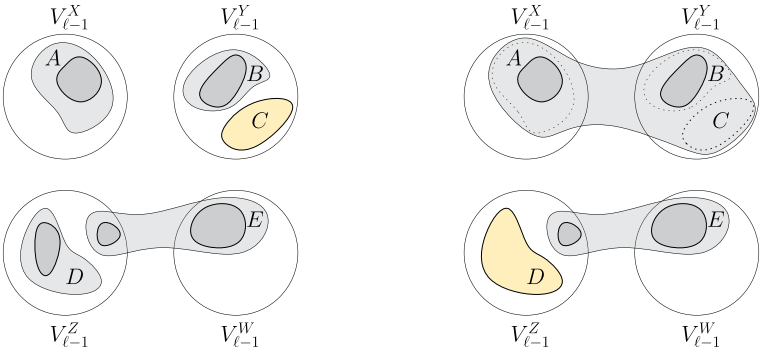

Fig. 2 An illustration of the evolution of good and bad clusters: In the example, we see five clusters at time $\ell - 1$. The clusters A, B, D, E are assumed to be bad, with their kernels depicted in dark gray, while C is assumed to be a good cluster. At time ℓ, clusters A, B and C are merged. The resulting cluster is bad because the kernels of A and B lie in different connected components of $G_{\ell-1}$. Clusters D and E are still present at time ℓ, but now D is a good cluster since its kernel is completely contained in $V^Z_{\ell-1}$, while E is still bad, since its kernel is disconnected in $G_{\ell-1}$

Lemma 11 *For all* $0 \leq \ell \leq k$ *we have* $\sum_{C \in \mathsf{Bad}(\ell)} |\mathsf{Ker}_\ell(C)| \geq \frac{\Gamma-\ell}{\Gamma} N_k$.

Proof We prove this via induction over ℓ. For $\ell = 0$ this is clear since $\bigcup_{C \in \mathsf{Bad}(0)} \mathsf{Ker}_0(C) = \mathcal{P}_k$.

Now suppose that $\ell > 0$ and that

$$\sum_{C \in \mathsf{Bad}(\ell)} |\mathsf{Ker}_\ell(C)| < \frac{\Gamma - \ell}{\Gamma} N_k.$$

By induction hypothesis we know that

$$\sum_{C \in \mathsf{Bad}(\ell-1)} |\mathsf{Ker}_{\ell-1}(C)| \geq \frac{\Gamma - \ell + 1}{\Gamma} N_k.$$

Thus the number of points which are in the kernel of a bad cluster at time $\ell - 1$ but not at time ℓ is larger than

$$\frac{\Gamma - \ell + 1}{\Gamma} N_k - \frac{\Gamma - \ell}{\Gamma} N_k = \frac{N_k}{\Gamma}.$$

In other words these are points that are in the kernel of a bad cluster at time $\ell - 1$ but contained in a good cluster at time ℓ. Now we use that any good cluster at time ℓ can contain only $N_{\ell-1}$ such points by Lemma 10. Thus the number of good clusters is greater than

$$\frac{N_k}{\Gamma} \cdot \frac{1}{N_{\ell-1}} = \frac{N_k}{\Gamma N_{\ell-1}}.$$

We obtain that $\mathcal{H}_{\frac{N_k}{\Gamma N_{\ell-1}}}$ contains more than $\frac{N_k}{\Gamma N_{\ell-1}}$ clusters, which is not possible. □

An immediate consequence of Lemma 11 is the existence of bad clusters at time ℓ for any $0 \leq \ell \leq k$. To prove that their (discrete) radius and diameter is indeed large we need a lower bound on the distance between two points $X, Y \in \mathcal{P}$ that lie in different connected components of G_{j-1} for some $1 \leq j \leq k$.

Suppose that the points X and Y only differ in one coordinate, i.e., there is a $1 \leq s \leq N_{j-1}$ such that $x_{js} \neq y_{js}$, while X and Y agree in all other coordinates. There is only one edge in G_j connecting $V_{j-1}^{X_{[j:k]}}$ with $V_{j-1}^{Y_{[j:k]}}$. Let $P = \phi_{j-1}^{-1}(s)$, then this edge connects the points $(P \mid X_{[j:k]})$ and $(P \mid Y_{[j:k]})$. If we connect X to $(P \mid X_{[j:k]})$ and $(P \mid Y_{[j:k]})$ to Y via a shortest path, this results in a path from X to Y, see Fig. 3. We show that this path is indeed a shortest path between X and Y and generalize this to arbitrary X and Y which are disconnected in G_{j-1}.

Lemma 12 *Let $X, Y \in \mathcal{P}$ be two points and suppose there is $1 \leq j \leq k$ and $1 \leq s \leq N_{j-1}$ such that $x_{js} \neq y_{js}$. Let $P = \phi_{j-1}^{-1}(s) \in \prod_{i=0}^{j-1} \mathcal{Q}_i$. Then*

$$d(X, Y) \geq d\big(X, (P \mid X_{[j:k]})\big) + j + d\big(Y, (P \mid Y_{[j:k]})\big).$$

Proof Observe that if two points in G are connected by an edge they differ in exactly one coordinate. Since $x_{js} \neq y_{js}$ any shortest path connecting X and Y must contain two consecutive points Z, Z' with $Z = (P \mid Z_j \mid \ldots \mid Z_k)$ and $Z' = (P \mid Z'_j \mid \ldots \mid Z'_k)$ such that $z_{js} \neq z'_{js}$ and Z agrees with Z' in all remaining coordinates. We obtain

$$d(X, Y) = d\big(X, Z\big) + d(Z, Z') + d\big(Z', Y\big) = d\big(X, Z\big) + j + d\big(Z', Y\big).$$

It is now left to show that $d(X, Z) \geq d\big(X, (P \mid X_{[j:k]})\big)$ and $d(Y, Z') \geq d\big(Y, (P \mid Y_{[j:k]})\big)$. To prove this we consider a shortest path $V^1, \ldots, V^t$ connecting $V^1 = X$ with $V^t = Z$. Let $W^i = (V^i_{[0:j-1]} \mid X_{[j:k]})$ for $i = 1, \ldots, t$. We claim that W^i is connected to W^{i+1} by an edge in G and that $d(V^i, V^{i+1}) \geq d(W^i, W^{i+1})$ for all $1 \leq i \leq t-1$. So let $1 \leq i \leq t-1$, we know that V^i and V^{i+1} differ in exactly one coordinate. If they differ at a coordinate in row $r \geq j$ we have $W^i = W^{i+1}$ and

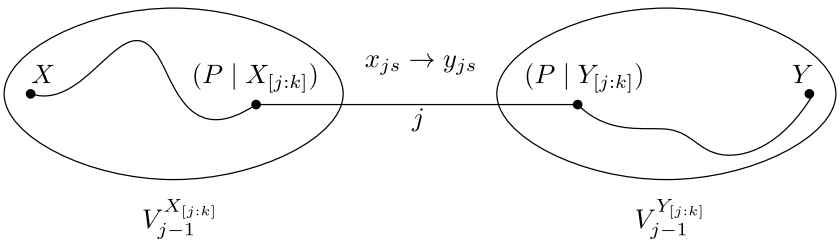

Fig. 3 A shortest path between X and Y. It consists of two shortest paths inside the connected components of G_{j-1} and the unique edge of weight j between these components

thus the claim holds. Otherwise let $u = \phi_{r-1}(V^i_{[0:r-1]})$ then V^i and V^{i+1} satisfy $v^i_{ru} \neq v^{i+1}_{ru}$ and $d(V^i, V^{i+1}) = r$. Since $r \leq j-1$ we obtain that W^i is connected to W^{i+1} by the edge

$$\{(V^i_{[0:r-1]} \mid v^i_{r1}, \ldots, v^i_{ru-1}, \star, v^i_{ru+1}, \ldots, v^i_{rN_{r-1}} \mid W^i_{[r+1:k]}) \mid \star \in [\Gamma N_{r-1}]\},$$

which has weight r. This yields the claim.

Observe that $W^1 = X$ and $W^t = (P \mid X_{[j:k]})$ and that

$$d\big(X, (P \mid X_{[j:k]})\big) \leq \sum_{i=1}^{t-1} d(W^i, W^{i+1}) \leq \sum_{i=1}^{t-1} d(V^i, V^{i+1}) = d(X, Z).$$

Analogously one can show $d(Y, Z') \geq d\big(Y, (P \mid Y_{[j:k]})\big)$ and obtains

$$d(X, Y) = d\big(X, Z\big) + j + d\big(Z', Y\big) \geq d\big(X, (P \mid X_{[j:k]})\big) + j + d\big(Y, (P \mid Y_{[j:k]})\big).$$

□

We now define the so called *anchor set* $\mathsf{Anc}_\ell(C)$ of a bad cluster C at time ℓ. If C is anchored at ℓ' then $\mathsf{Anc}_\ell(C)$ is the union of ℓ' and the anchor set of some bad cluster $D \subset C$ at time $\ell - 1$. If we choose D appropriately the sum of anchors in $\mathsf{Anc}_\ell(C)$ is a lower bound on the discrete radius of C, as we show later. It is clear that ℓ' itself is a lower bound on the discrete radius since $\mathsf{Ker}_\ell(C)$ is disconnected in $G_{\ell'-1}$ by definition. If we additionally assume that the discrete radius of D is large, e.g., lower bounded by the sum of anchors in $\mathsf{Anc}_{\ell-1}(D)$, then it is reasonable to assume that the discrete radius of C is lower bounded by some function in ℓ' and the sum of anchors in $\mathsf{Anc}_{\ell-1}(D)$. Before proving this we give a formal definition of $\mathsf{Anc}_\ell(C)$ and how to choose D.

Definition 6 Let $1 \leq \ell \leq k$ and C be a bad cluster at time ℓ which is anchored at $\ell' \geq \ell$. If $\ell = 1$ we define the anchor set of C as $\mathsf{Anc}_1(C) = \{\ell'\}$ and set $\mathsf{prev}(C) = \{X\}$ for some $X \in C$.

For $\ell > 1$ we distinguish two cases.

Case 1: C contains a bad cluster D which is bad at time $\ell - 1$ and anchored at ℓ'. We then set $\mathsf{Anc}_\ell(C) = \mathsf{Anc}_{\ell-1}(D)$ and $\mathsf{prev}(C) = D$.

Case 2: C does not contain such a cluster. Then let $D \subset C$ be a bad cluster at time $\ell - 1$ minimizing

$$\sum_{a \in \mathsf{Anc}_{\ell-1}(D)} a$$

among all clusters $D' \in \mathsf{Bad}(\ell - 1)$ with $D' \subset C$. We set $\mathsf{Anc}_\ell(C) = \mathsf{Anc}_{\ell-1}(D) \cup \{\ell'\}$and $\mathsf{prev}(C) = D$.

Observe that in Case 2 of the previous definition, the bad cluster D must be anchored at some $\ell_D < \ell'$.

Lemma 13 *Let $1 \leq \ell \leq k$ and C be a bad cluster at time ℓ. If C contains a cluster D which is bad at time $\ell - 1$ then $\mathsf{Ker}_{\ell-1}(D) \subset \mathsf{Ker}_\ell(C)$.*

Proof Since $D \in \mathsf{Bad}(\ell - 1)$ and $D \subset C$, we get

$$\mathsf{Ker}_{\ell-1}(D) \subset \bigcup_{D' \subset \mathsf{Bad}(\ell-1): D' \subset C} \mathsf{Ker}_{\ell-1}(D') = \mathsf{Ker}_\ell(C).$$

□

With the help of Lemma 12 we are able to show how the discrete radius and diameter of a bad cluster, depends on the sum of anchors.

Lemma 14 *Let $1 \leq \ell \leq k$ and C be a bad cluster at time ℓ anchored at ℓ'. Then for any point $Z \in \mathcal{P}$ there is $X \in \mathsf{Ker}_\ell(C)$ such that*

$$d(Z, X) \geq \sum_{a \in \mathsf{Anc}_\ell(C)} a.$$

Proof Let $Z \in \mathcal{P}$ and suppose that C is a bad cluster at time ℓ anchored at ℓ'. We prove the lemma via induction over ℓ. For $\ell = 1$ we know that $\mathsf{Ker}_\ell(C)$ is disconnected in $G_{\ell'-1}$ by definition. Thus there is a point $X \in \mathsf{Ker}_\ell(C)$ which is disconnected from Z in $G_{\ell'-1}$ yielding

$$d(Z, X) \geq \ell' = \sum_{a \in \mathsf{Anc}_1(C)} a.$$

Let $\ell > 1$. If $D = \mathsf{prev}(C)$ is anchored at ℓ' we apply Lemma 13 to observe that $\mathsf{Ker}_{\ell-1}(D) \subset \mathsf{Ker}_\ell(C)$. By induction hypothesis the lemma holds for D. Since $\mathsf{Anc}_\ell(C) = \mathsf{Anc}_{\ell-1}(D)$ the lemma also holds for C.

Otherwise let $D = \mathsf{prev}(C)$ be anchored at $\ell_D < \ell'$. We know that $\mathsf{Ker}_\ell(C)$ is disconnected in $G_{\ell'-1}$. On the other hand $\mathsf{Ker}_{\ell-1}(D)$ is connected in $G_{\ell'-1}$ since $\ell_D < \ell'$. Thus there is $V \in \mathsf{Ker}_\ell(C)$ which is disconnected from $\mathsf{Ker}_{\ell-1}(D)$ in $G_{\ell'-1}$. Let $E \subset C$ be the cluster at time $\ell - 1$ which contains V. Since $V \in \mathsf{Ker}_\ell(C)$ we know that E is a bad cluster at time $\ell - 1$ anchored at $\ell_E < \ell'$. We know that $\mathsf{Ker}_{\ell-1}(E)$ is connected in $G_{\ell'-1}$ and lies in a different connected component than $\mathsf{Ker}_{\ell-1}(D)$. Thus Z is disconnected from $\mathsf{Ker}_{\ell-1}(D)$ or $\mathsf{Ker}_{\ell-1}(E)$ in $G_{\ell'-1}$.

We assume without loss of generality that Z is disconnected from $\mathsf{Ker}_{\ell-1}(E)$ in $G_{\ell'-1}$. Since $\mathsf{Ker}_{\ell-1}(E)$ is connected in $G_{\ell'-1}$ we know by Lemma 7 that $(P \mid Y_{[\ell':k]}) = (P \mid Y'_{[\ell':k]})$ for all $Y, Y' \in \mathsf{Ker}_{\ell-1}(E)$. Also by

Lemma 7 there is $\ell' \le r \le k$ and $1 \le s \le N_{r-1}$ such that $z_{rs} \neq y_{rs}$ for all $Y \in \mathsf{Ker}_{\ell-1}(E)$. Let $P = \phi_{r-1}^{-1}(s)$. Thus we know by induction hypothesis that there is a point $X \in \mathsf{Ker}_{\ell-1}(E) \subset \mathsf{Ker}_{\ell}(C)$ with

$$d(X, (P \mid X_{[r:k]})) \ge \sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a.$$

Figure 4 shows an exemplary path between X and Z.

We apply Lemma 12 to see that

$$\begin{aligned} d(Z, X) &\ge d\big(Z, (P \mid Z_{[r:k]})\big) + r + d\big(X, (P \mid X_{[r:k]})\big) \\ &\ge r + \sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a \\ &\ge \ell' + \sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a \\ &\ge \sum_{a \in \mathsf{Anc}_{\ell}(C)} a \end{aligned}$$

Here the last inequality follows from the minimality of $\sum_{a \in \mathsf{Anc}_{\ell-1}(D)} a$ among all clusters $D' \in \mathsf{Bad}(\ell-1)$ with $D' \subset C$.

If Z is disconnected from $\mathsf{Ker}_{\ell-1}(D)$ in $G_{\ell'-1}$ our argument still works after replacing E by D. □

Lemma 15 *Let $1 \le \ell \le k$ and C be a bad cluster at time ℓ anchored at ℓ'. Then there are two points $X, Y \in \mathsf{Ker}_{\ell}(C)$ such that*

$$d(X, Y) \ge \ell' + 2 \sum_{a \in \mathsf{Anc}_{\ell}(C) \setminus \{\ell'\}} a.$$

Proof Suppose that C is a bad cluster at time ℓ anchored at ℓ'. We prove the lemma via induction over ℓ. For $\ell = 1$ we know that $\mathsf{Ker}_{\ell}(C)$ is disconnected in $G_{\ell'-1}$ by

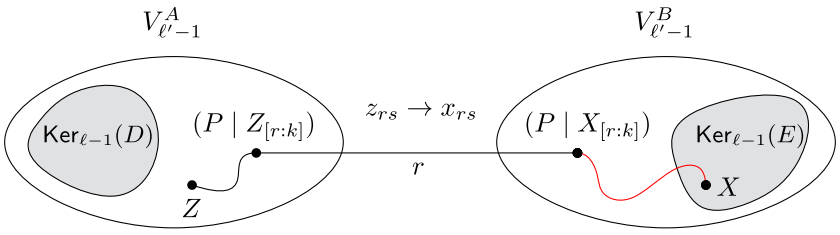

Fig. 4 Shows the special case where $Z_{[r:k]}$ and $Y_{[r:k]}$ only differ in the rs-coordinate. The length of the right path is lower bounded by $\sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a$

definition. Thus there are two points $X, Y \in \mathsf{Ker}_\ell(C)$ that are disconnected in $G_{\ell'-1}$ yielding

$$d(X,Y) \geq \ell' = \ell' + 2 \sum_{a \in \mathsf{Anc}_1(C) \setminus \{\ell'\}} a.$$

Let $\ell > 1$. If $D = \mathsf{prev}(C)$ is anchored at ℓ' we apply Lemma 13 to observe that $\mathsf{Ker}_{\ell-1}(D) \subset \mathsf{Ker}_\ell(C)$. By induction hypothesis the lemma holds for D. Since $\mathsf{Anc}_\ell(C) = \mathsf{Anc}_{\ell-1}(D)$ the lemma also holds for C.

Otherwise let $D = \mathsf{prev}(C)$ be anchored at $\ell_D < \ell'$. We know that $\mathsf{Ker}_\ell(C)$ is disconnected in $G_{\ell'-1}$ and $\mathsf{Ker}_{\ell-1}(D)$ is connected in $G_{\ell'-1}$. Thus there is $V \in \mathsf{Ker}_\ell(C)$ which is disconnected from $\mathsf{Ker}_{\ell-1}(D)$ in $G_{\ell'-1}$. Let $E \subset C$ be the cluster at time $\ell - 1$ which contains V. We know that E is a bad cluster at time $\ell - 1$ anchored at $\ell_E < \ell'$. Furthermore $\mathsf{Ker}_{\ell-1}(E)$ is connected in $G_{\ell'-1}$ and lies in a different connected component than $\mathsf{Ker}_{\ell-1}(D)$.

Since $\mathsf{Ker}_{\ell-1}(D)$ and $\mathsf{Ker}_{\ell-1}(E)$ are disconnected in $G_{\ell'-1}$ but connected in $G_{\ell'}$, there must be $1 \leq s \leq N_{\ell'-1}$ such that for all $U \in \mathsf{Ker}_{\ell-1}(D)$ and $T \in \mathsf{Ker}_{\ell-1}(E)$ we have $u_{\ell' s} \neq t_{\ell' s}$ by Lemma 7. Let $P = \phi^{-1}_{\ell'-1}(s)$, we know by Lemma 12 that

$$d(U,T) \geq d(U, (P \mid U_{[\ell':k]})) + \ell' + d(T, (P \mid T_{[\ell':k]})).$$

Let $U \in \mathsf{Ker}_{\ell-1}(D)$ and $T \in \mathsf{Ker}_{\ell-1}(E)$. We know by Lemma 14 that for any two points $Z = (P \mid U_{[\ell':k]})$ and $Z' = (P \mid T_{[\ell':k]})$ there must be $X \in \mathsf{Ker}_{\ell-1}(D)$ and $Y \in \mathsf{Ker}_{\ell-1}(E)$ such that

$$d(X,Z) \geq \sum_{a \in \mathsf{Anc}_{\ell-1}(D)} a$$

and

$$d(Y,Z') \geq \sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a.$$

We use Lemma 7 to observe that $Z = (P \mid X_{[\ell':k]})$ and $Z' = (P \mid Y_{[\ell':k]})$ because X is connected to U and Y is connected to T in $G_{\ell'-1}$. Figure 5 shows an exemplary path between X and Y. Thus

$$\begin{aligned} d(X,Y) &\geq d(X, (P \mid X_{[\ell':k]})) + \ell' + d(Y, (P \mid Y_{[\ell':k]})) \\ &\geq d(X,Z) + \ell' + d(Y,Z') \\ &\geq \ell' + \sum_{a \in \mathsf{Anc}_{\ell-1}(D)} a + \sum_{a \in \mathsf{Anc}_{\ell-1}(E)} a \\ &\geq \ell' + 2 \sum_{a \in \mathsf{Anc}_\ell(C) \setminus \{\ell'\}} a \end{aligned}$$

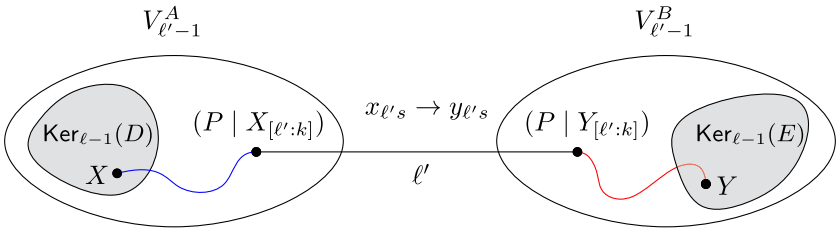

Fig. 5 Shows the special case where $X_{[\ell':k]}$ and $Y_{[\ell':k]}$ only differ in the $\ell's$-coordinate. The length of the left path is lower bounded by $\sum_{a\in\mathsf{Anc}_{\ell-1}(D)} a$, while the length of the right path is lower bounded by $\sum_{a\in\mathsf{Anc}_{\ell-1}(E)} a$

Here the last inequality follows from the minimality of $\sum_{a\in\mathsf{Anc}_{\ell-1}(D)} a$ among all clusters $D' \in \mathsf{Bad}(\ell-1)$ with $D' \subset C$. □

5.3 Comparison to Optimal Clusterings

Our initial motivation was to construct an instance where any hierarchical clustering has a high approximation ratio. If we consider an arbitrary time $1 \le \ell \le k$ then the hierarchical clustering $\mathscr{H}$ on $(\mathcal{X}, \mathcal{P}, d)$ may be even optimal at time ℓ. Thus the bounds which we develop in Lemmas 14 and 15 on the discrete radius and diameter of bad clusters are useless without linking the cost of a bad cluster at time ℓ to the cost of bad clusters at other time steps. Therefore we construct a sequence of clusters $C_1 \subset C_2 \ldots \subset C_k$ where C_i is a bad cluster at time i such that $\mathsf{Anc}_1(C_1) \subset \mathsf{Anc}_2(C_2) \subset \ldots \subset \mathsf{Anc}_k(C_k)$. We then show with the help of Lemmas 14 and 15 that at least one of these clusters has a high discrete radius and diameter compared to the optimal cost.

Lemma 16 *Let C_k be a bad cluster at time k. For $1 \le i \le k-1$ we define $C_i = \mathsf{prev}(C_{i+1})$. For all $1 \le i \le k-1$ cluster C_i is bad at timeiand one of the following two cases occurs:*

1. $\mathsf{Anc}_i(C_i) = \mathsf{Anc}_{i+1}(C_{i+1})$,
2. $\mathsf{Anc}_{i+1}(C_{i+1})\backslash\{\ell\} = \mathsf{Anc}_i(C_i)$, *where* $\ell = \max \mathsf{Anc}_{i+1}(C_{i+1})$.

Proof For $i = k$ cluster C_k is bad at time k by assumption. If C_{i+1} is a bad cluster at time $i+1$ then $C_i = \mathsf{prev}(C_{i+1})$ is a bad cluster at time i, by definition of prev.

Let C_i be anchored at $\ell' \ge i$ and C_{i+1} be anchored at $\ell \ge i+1$. Since $\mathsf{Ker}_i(C_i) \subset \mathsf{Ker}_{i+1}(C_{i+1})$ by Lemma 13, we know that $\ell' \le \ell$. If $\ell' = \ell$ we obtain by Definition 6, that $\mathsf{Anc}_i(C_i) = \mathsf{Anc}_{i+1}(C_{i+1})$, so the lemma holds in this case.

If $\ell' < \ell$ we know by Definition 6 that $\mathsf{Anc}_i(C_i) = \mathsf{Anc}_{i+1}(C_{i+1})\backslash\{\ell\}$. So the lemma also holds in this case. □

Corollary 17 *Let C_k be a bad cluster at time k. For $1 \le i \le k-1$ we define $C_i = \mathsf{prev}(C_{i+1})$. Let $\mathsf{Anc}_k(C_k) = \{\ell_1, \ldots, \ell_s\}$ such that $\ell_{t-1} < \ell_t$ for all $2 \le t \le s$*

and let $\ell_0 = 0$. *Then for any* $1 \leq t \leq s$ *and for anyiwith* $\ell_{t-1} < i \leq \ell_t$, *we have* $\{\ell_1, \ldots, \ell_t\} \subset \mathsf{Anc}_i(C_i)$.

Proof We prove this via induction over i, starting from $i = k$ in decreasing order. There is nothing to show for $i = k$. For $i < k$ we distinguish two cases. If $\mathsf{Anc}(C_i) = \mathsf{Anc}_{i+1}(C_{i+1})$, the lemma follows from the induction hypothesis.

Otherwise remember that $\mathsf{Anc}_i(C_i) \subset \mathsf{Anc}_k(C_k)$ and $\ell_{t-1} < i$. Thus we know that $\max \mathsf{Anc}_i(C_i) \in \{\ell_t, \ldots, \ell_s\}$ and therefore $\ell_t \leq \max \mathsf{Anc}_i(C_i)$.

By Lemma 16 we know that $\mathsf{Anc}_i(C_i) = \mathsf{Anc}_{i+1}(C_{i+1}) \backslash \{\ell\}$, where $\ell = \max \mathsf{Anc}_{i+1}(C_{i+1})$. Thus $\ell_t \leq \max \mathsf{Anc}_i(C_i) < \max \mathsf{Anc}_{i+1}(C_{i+1}) = \ell$ and by induction hypothesis we obtain

$$\{\ell_1, \ldots, \ell_t\} \subset \mathsf{Anc}_{i+1}(C_{i+1}) \backslash \{\ell\} = \mathsf{Anc}_i(C_i).$$

□

Before we are able to prove the theorem we need some final lemma.

Lemma 18 *For every* $\epsilon > 0$ *there exists* $k \in \mathbb{N}$ *such that for every* $s \in \mathbb{N}$ *any sequence of* $s + 1$ *numbers* $(\ell_0, \ldots, \ell_s) \in \mathbb{R}_{\geq 0}^{s+1}$ *with* $\ell_0 = 0$ *and* $\ell_s = k$ *satisfies the following.*

1. *There exists* $1 \leq t \leq s$ *such that for* $\alpha_1 = 4 - \epsilon$ *and* $\Delta_1 = 1$ *we have*

$$\frac{\ell_t + \Delta_1 \sum_{i=0}^{t-1} \ell_i}{\ell_{t-1} + 1} > \alpha_1.$$

2. *There exists* $1 \leq t \leq s$ *such that for* $\alpha_2 = 3 + 2\sqrt{2} - \epsilon$ *and* $\Delta_2 = 2$ *we have*

$$\frac{\ell_t + \Delta_2 \sum_{i=0}^{t-1} \ell_i}{\ell_{t-1} + 1} > \alpha_2.$$

Proof Let $k, s \in \mathbb{N}$ and $j \in \{1, 2\}$.

We call a sequence $(a_0, \ldots, a_s) \in \mathbb{R}_{\geq 0}^{s+1}$ *feasible* if $a_0 = 0, a_s = k$ and for all $1 \leq t \leq s$ we have

$$\frac{a_t + \Delta_j \sum_{i=0}^{t-1} a_i}{a_{t-1} + 1} \leq \alpha_j. \tag{1}$$

Our proof is divided in two parts. In the first part we argue that for all $k, s \in \mathbb{N}$ the existence of a feasible sequence $(\ell_0, \ldots, \ell_s)$ yields the existence of a feasible sequence $(b_0, \ldots, b_s)$ which satisfies (1) for all $u + 1 \leq t \leq s$ with equality, where u is the smallest number such that $b_u \neq 0$. In the second part we observe that there exists $k \in \mathbb{N}$ such that for all $s \in \mathbb{N}$ there is no feasible sequence $(a_0, \ldots, a_s) \in \mathbb{R}_{\geq 0}^{s+1}$

which satisfies (1) for all $u+1 \leq t \leq s$ with equality, where u is the smallest number such that $a_u \neq 0$. In combination both parts yield the lemma.

Part 1: Let $k, s \in \mathbb{N}$ and suppose that there exists a feasible sequence $(\ell_0, \ldots, \ell_s)$. We consider the set

$$M = \{(a_0, \ldots, a_s) \in \mathbb{R}_{\geq 0}^{s+1} \mid (a_0, \ldots, a_s) \text{ is feasible}\}$$

of all feasible sequences.

For $(a_0, \ldots, a_s) \in M$, we claim that $a_t \leq (\alpha_j + 1)^{t+1}$ for all $0 \leq t \leq s$. We show this via a simple induction over t. If $t = 0$ there is nothing to show since $a_0 = 0$. For $t > 0$ we obtain

$$a_t \leq \alpha_j(a_{t-1}+1) - \Delta_j \sum_{i=0}^{t-1} a_i \leq \alpha_j(a_{t-1}+1) \leq \alpha_j((\alpha_j+1)^t + 1) \leq (\alpha_j+1)^{t+1}.$$

Here the first inequality follows from the feasibility of the sequence. As a consequence we see that M is a bounded set. Furthermore M is also closed since $a_0 = 0, a_t = k$ are both linear inequalities and (1) is a linear inequality for all $1 \leq t \leq s$. Thus M is compact.

We consider the function $F : M \to \mathbb{R}$ with $F(a_0, \ldots, a_s) = \sum_{i=0}^{s} a_i$. Since F is continuous and M is compact and non-empty we know that F attains a minimum on M, i.e., there is $(b_0, \ldots, b_s) \in M$ with $F(b_0, \ldots, b_s) \leq F(a_0, \ldots, a_s)$ for all $(a_0, \ldots, a_s) \in M$. We claim that $(b_0, \ldots, b_s)$ satisfies (1) with equality for all $u+1 \leq t \leq s$, where u is the smallest number such that $b_u \neq 0$. Suppose this is not the case and let $u+1 \leq t \leq s$ be a number such that

$$\frac{b_t + \Delta_j \sum_{i=0}^{t-1} b_i}{b_{t-1}+1} < \alpha_j.$$

If $b_{t-1} = 0$, then $(0, \ldots, 0, b_t, \ldots, b_s)$ is also feasible and moreover

$$F(0, \ldots, 0, b_t, \ldots, b_s) = \sum_{i=t}^{s} b_i < b_u + \sum_{i=t}^{s} b_i \leq F(b_0, \ldots, b_s)$$

in contradiction to $(b_0, \ldots, b_s)$ being a minimum. Thus we must have $b_{t-1} > 0$ and therefore by continuity there exists an $\epsilon \in (0, b_{t-1})$, such that

$$\frac{b_t + \Delta_j(b_{t-1} - \epsilon) + \Delta_j \sum_{i=0}^{t-2} b_i}{b_{t-1} - \epsilon + 1} \leq \alpha_j.$$

Observe that the sequence $(c_0, \ldots, c_s) = (b_0, \ldots, b_{t-2}, b_{t-1} - \epsilon, b_t, \ldots, b_s)$ is still feasible. The t-th inequality is satisfied by choice of ϵ. All other inequalities are satisfied, since for all $1 \leq r \leq s$ with $r \neq t$ we have

$$\frac{c_r + \Delta_j \sum_{i=0}^{r-1} c_i}{c_{r-1}+1} \leq \frac{b_r + \Delta_j \sum_{i=0}^{r-1} b_i}{b_{r-1}+1} \leq \alpha_j.$$

On the other hand

$$F(c_0, \ldots, c_s) = \sum_{i=0}^{s} c_i = -\epsilon + \sum_{i=0}^{s} b_i < F(b_0, \ldots, b_s),$$

which again stands in contradiction to $(b_0, \ldots, b_s)$ being the minimum. Thus $(b_0, \ldots, b_s)$ is of the desired form.

Part 2: Let $k, s \in \mathbb{N}$ and $(a_0, \ldots, a_s) \in \mathbb{R}_{\geq 0}^{s+1}$ be a feasible sequence which satisfies (1) for all $u+1 \leq t \leq s$ with equality, where u is the smallest number such that $a_u \neq 0$. Thus we know that $a_1 = \ldots = a_{u-1} = 0$ and $a_u \in (0, \alpha_j]$. Furthermore

$$a_{u+1} = \alpha_j(a_u + 1) - \Delta_j \sum_{i=0}^{u} a_i = \alpha_j(a_u+1) - \Delta_j a_u$$

and for $u+2 \leq t \leq s$ we have

$$\begin{aligned} a_t &= \alpha_j(a_{t-1}+1) - \Delta_j \sum_{i=0}^{t-1} a_i \\ &= \alpha_j(a_{t-1}+1) - \Delta_j a_{t-1} - \Delta_j \sum_{i=0}^{t-2} a_i \\ &= \alpha_j(a_{t-1}+1) - \Delta_j a_{t-1} - (\alpha_j(a_{t-2}+1) - a_{t-1}) \\ &= \alpha_j(a_{t-1} - a_{t-2}) - (\Delta_j - 1)a_{t-1}. \end{aligned}$$

Here we use that (1) is satisfied with equality for t and $t-1$.

Let

$$\Psi = \frac{\alpha_j - \Delta_j + 1 + \sqrt{(\alpha_j - \Delta_j + 1)^2 - 4\alpha_j}}{2}$$

and

$$\Theta = \frac{\alpha_j - \Delta_j + 1 - \sqrt{(\alpha_j - \Delta_j + 1)^2 - 4\alpha_j}}{2}$$

be the two roots of the polynomial $X^2 - (\alpha_j - \Delta_j + 1)X + \alpha_j$. We observe later that $\Phi \neq \Theta$. Let $x = \frac{\Theta a_u - a_{u+1}}{\Theta - \Phi}$ and $y = \frac{a_{u+1} - \Phi a_u}{\Theta - \Phi}$.

Claim: It holds that $a_t = \Phi^{t-u}x + \Theta^{t-u}y$ for all $u \leq t \leq s$.

We prove this claim by induction over t. For $t = u$ we obtain

$$x + y = \frac{\Theta a_u - a_{u+1} + a_{u+1} - \Phi a_u}{\Theta - \Phi} = a_u.$$

For $t = u + 1$ we obtain

$$\Phi x + \Theta y = \frac{\Phi\Theta a_u - \Phi a_{u+1} + \Theta a_{u+1} - \Theta\Phi a_u}{\Theta - \Phi} = a_{u+1}.$$

For $t > u + 1$ we obtain

$$\begin{aligned}
&\Phi^{t-u}x + \Theta^{t-u}y \\
&\quad = \Phi^{t-u-2}x((\alpha_j - \Delta_j + 1)\Phi - \alpha_j) + \Theta^{t-u-2}y((\alpha_j - \Delta_j + 1)\Theta - \alpha_j) \\
&\quad = \alpha_j((\Phi^{t-u-1}x + \Theta^{t-u-1}y) - (\Phi^{t-u-2}x + \Theta^{t-u-2}y)) \\
&\qquad - (\Delta_j - 1)(\Phi^{t-u-1}x + \Theta^{t-u-1}y) \\
&\quad = \alpha_j(a_{t-1} - a_{t-2}) - (\Delta_j - 1)a_{t-1} \\
&\quad = a_t.
\end{aligned}$$

For the first equality we used that Φ and Θ are roots of $X^2 - (\alpha_j - \Delta_j + 1)X + \alpha_j$, i.e., $\Phi^2 = (\alpha_j - \Delta_j + 1)\Phi - \alpha_j$ and $\Theta^2 = (\alpha_j - \Delta_j + 1)\Theta - \alpha_j$. For the third equality we used the induction hypothesis. This proves the claim.

We argue that if k is large enough, there must be $u \le t \le s$ with $a_t < 0$ in contradiction to our assumption that $(a_0, \ldots, a_s)$ is feasible. For this we observe that by choice of α_j and Δ_j, we get $(\alpha_j - \Delta_j + 1)^2 - 4\alpha_j < 0$ and thus Φ and Θ are complex numbers. Furthermore Φ and Θ are complex conjugates and so are x and y. Thus there exists $r > 0$ such that the real part of $\Phi^r x$ and $\Theta^r y$ is negative and thus $\Phi^r x + \Theta^r y$ is negative, see Fig. 6.

Observe that $a_t \le (\alpha_j + 1)^{t-u+1}$ for $u \le t \le s$. One can prove this similar to the bound in Part 1. Thus if $k \ge (\alpha_j + 1)^r$ we obtain $s \ge r + u$ and thus $a_{r+u} = \Phi^r x + \Theta^r y$ is negative. Therefore $(a_0, \ldots, a_s)$ is not feasible in contradiction to our assumption.

Let now $k \ge (\alpha_j + 1)^r$ and suppose there exists $s \in \mathbb{N}$ and a feasible sequence $(\ell_0, \ldots, \ell_s)$. By the first part we know that there also exists a feasible sequence $(a_0, \ldots, a_s)$ which satisfies (1) for all $u + 1 \le t \le s$ with equality, where u is the smallest number such that $a_u \ne 0$. This is in contradiction with the second part, where we prove that for $k \ge (\alpha_j + 1)^r$ such a sequence cannot exist. □

Theorem 5 *For* $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ *we have* $\rho_{\mathsf{cost}} \ge 3 + 2\sqrt{2}$ *and for* $\mathsf{cost} = \mathsf{drad}$ *we have* $\rho_{\mathsf{cost}} \ge 4$.

Proof Let $\epsilon > 0$ and k be the respective number from Lemma 18. We claim that the approximation factor of any hierarchical clustering $\mathscr{H} = (\mathcal{H}_{N_k}, \ldots, \mathcal{H}_1)$ on the instance $(\mathcal{X}, \mathcal{P}, d)$ is larger than $3 + 2\sqrt{2} - \epsilon$ if $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{rad}\}$ and larger than $4 - \epsilon$ if $\mathsf{cost} = \mathsf{drad}$. First we use Lemma 11 to observe that there is a cluster $C_k \in \mathcal{H}_{\frac{N_k}{\Gamma N_{k-1}}}$ that is bad at time k. For $1 \le i \le k - 1$ we define

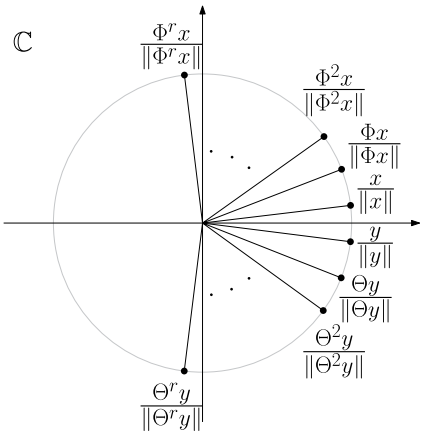

Fig. 6 Here we see the normalized numbers on the complex plane

$C_i = \mathsf{prev}(C_{i+1})$. Let $\mathsf{Anc}_k(C_k) = \{\ell_1, \ldots, \ell_s\}$ with $\ell_{t-1} < \ell_t$ for $2 \le t \le s$ and let $\ell_0 = 0$. We know by Corollary 17, that for any $1 \le t \le s$ and for $i = \ell_{t-1} + 1$ we have $\{\ell_1, \ldots, \ell_t\} \subset \mathsf{Anc}_i(C_i)$. Let $\ell' = \max \mathsf{Anc}_i(C_i)$, we obtain by Lemmas 15 and 14 that

$$\begin{aligned}
\mathsf{diam}(C_i) &\ge \ell' + 2 \sum_{a \in \mathsf{Anc}_i(C_i) \setminus \{\ell'\}} a \ge \ell_t + 2 \sum_{u=1}^{t-1} \ell_u, \\
\mathsf{rad}(C_i) &\ge \frac{\mathsf{diam}(C_i)}{2} \ge \frac{\ell_t + 2\sum_{u=1}^{t-1} \ell_u}{2}, \\
\mathsf{drad}(C_i) &\ge \sum_{a \in \mathsf{Anc}_i(C_i)} a \ge \sum_{u=1}^{t} \ell_u.
\end{aligned}$$

Remember that by Corollary 9 $\mathcal{A}_i$ is an optimal $\frac{N_k}{\Gamma N_{i-1}}$-clustering with $\mathsf{cost}(\mathcal{A}_i) = i$ if $\mathsf{cost} \in \{\mathsf{diam}, \mathsf{drad}\}$ and $\mathsf{cost}(\mathcal{A}_i) = i/2$ if $\mathsf{cost} = \mathsf{rad}$. We obtain

$$\begin{aligned}
\frac{\mathsf{rad}(C_i)}{\mathsf{rad}(\mathcal{A}_i)} = \frac{2\mathsf{rad}(C_i)}{2\mathsf{rad}(\mathcal{A}_i)} \ge \frac{\mathsf{diam}(C_i)}{\mathsf{diam}(\mathcal{A}_i)} \ge \frac{\ell_t + 2\sum_{u=1}^{t-1} \ell_u}{\ell_{t-1} + 1} \\
\frac{\mathsf{drad}(C_i)}{\mathsf{drad}(\mathcal{A}_i)} \ge \frac{\sum_{u=1}^{t} \ell_u}{\ell_{t-1} + 1}
\end{aligned}$$

which are lower bounds on the approximation factor of $\mathscr{H}$.

We apply Lemma 18 on $(\ell_0, \ldots, \ell_s)$ to observe that there is $1 \le t' \le s$ such that

$$\frac{\ell_{t'} + 2\sum_{u=1}^{t'-1} \ell_u}{\ell_{t'-1} + 1} > 3 + 2\sqrt{2} - \epsilon$$

and an $1 \le t'' \le s$ such that

$$\frac{\sum_{u=1}^{t''} \ell_u}{\ell_{t''-1}+1} > 4-\epsilon.$$

This proves the theorem. □

6 Conclusions and Open Problems

We have proved tight bounds for the price of hierarchy with respect to the diameter and (discrete) radius. It would be interesting to also obtain a better understanding of the price of hierarchy for other important objective functions like *k*-median and *k*-means. The best known upper bound is 16 for *k*-median [23] and 32 for *k*-means [6] but no non-trivial lower bounds are known. Closing this gap also for these objectives is a challenging problem for further research.

Another natural question is which approximation factors can be achieved by polynomial-time algorithms. The algorithm we used in this article to prove the upper bounds is not a polynomial-time algorithm because it assumes that for each level *k* an optimal *k*-clustering is given. The approximation factors worsen if only approximately optimal clusterings are used instead. It is known that 8-approximate hierarchical clusterings can be computed efficiently with respect to the diameter and (discrete) radius [2]. It is not clear whether or not it is NP-hard to obtain better hierarchical clusterings. The only NP-hardness results come from the problems with given *k*. Since computing a $(2-\epsilon)$-approximation for *k*-clustering with respect to the diameter and (discrete) radius is NP-hard, this is also true for the hierarchical versions. However, this is obsolete due to our lower bound, which shows that in general there does not even exist a $(2-\epsilon)$-approximate hierarchical clustering.

Appendix A: Counterexample for Mondal's Algorithm

The algorithm by Dasgupta and Long [2] computes a hierarchical clustering which is an 8-approximation with respect to the discrete radius objective and the diameter objective. Mondal's algorithm is a modification of this algorithm and should compute a 6-approximation for the discrete radius objective [10, Theorem 3.7]. We claim that this is not correct and present an example where the approximation factor is 7. First we give a brief summary of Mondal's algorithm.

Let $(\mathcal{X}, \mathcal{P}, d)$ be the clustering instance. In the beginning we compute a numbering of the points in $\mathcal{P}$ by running Gonzales' algorithm [9]. The numbering is computed as follows. We pick the first point $x_1 \in \mathcal{P}$ arbitrarily and set $R_1 = \infty$. For $2 \leq k \leq |\mathcal{P}|$ we set

$$x_k = argmax_{x \in \mathcal{P} \setminus \{x_1, \ldots, x_{k-1}\}} \min_{1 \leq i \leq k-1} d(x, x_i)$$

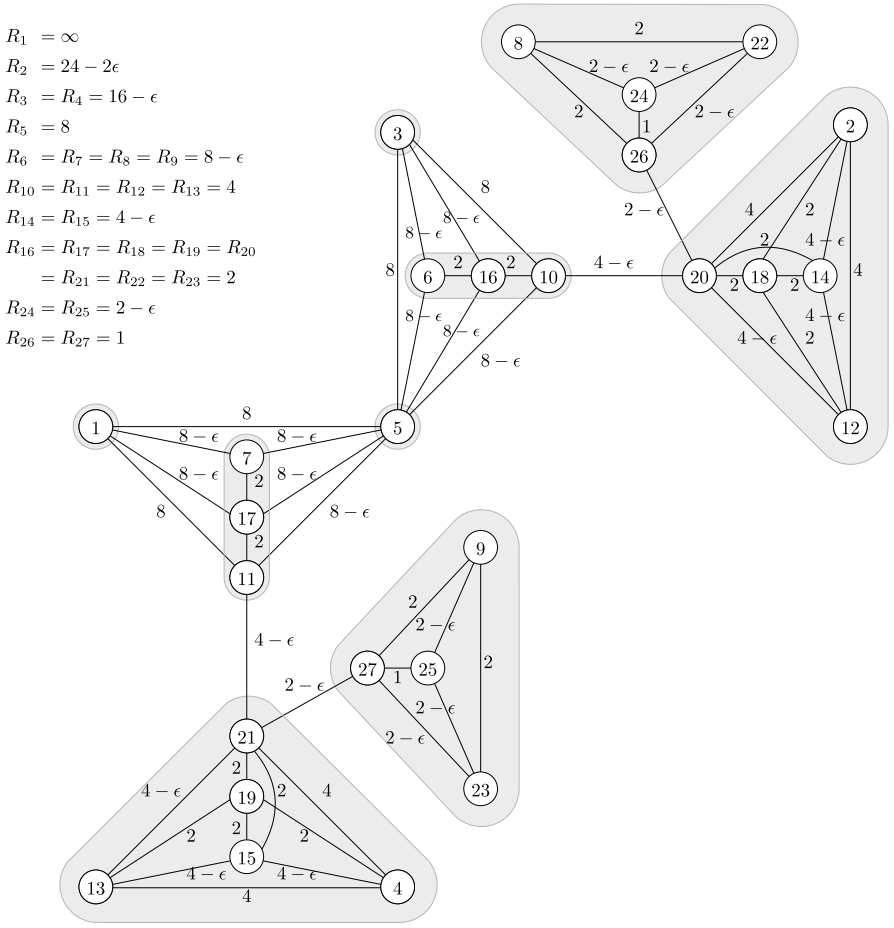

Fig. 7 Here we see the clustering instance and the numbering obtained from Gonzales' algorithm as well as the optimal 9-clustering with radius 2 depicted in gray

and $R_k = \min_{1 \leq i \leq k-1} d(x_k, x_i)$. In other words the *k*-th point is picked as far as possible from the points $x_1, \ldots, x_{k-1}$ and we denote by R_k the distance of x_k to $x_1, \ldots, x_{k-1}$.

Based on the *R*-values we define the parent of a point $x \in \mathcal{P} \backslash \{x_1\}$. Let $N(x) = argmin\{d(x, y) \mid y \in \mathcal{P}, R_x \leq \frac{R_y}{2}\}$ denote the parent of *x*. In other words *N*(*x*) is the point nearest to *x* that satisfies $R_x \leq \frac{R_{N(x)}}{2}$. Notice that every point in $\mathcal{P} \backslash \{x_1\}$ has a properly defined parent, as $R_1 = \infty$.

We build a tree on $\mathcal{P}$ as follows. For every point $x \in \mathcal{P}$ we simply add an edge between *x* and *N*(*x*). The resulting graph is cycle free, since $R_x < R_{N(x)}$ for all $x \in \mathcal{P}$, and contains $|\mathcal{P}| - 1$ edges. Thus it is indeed a tree.

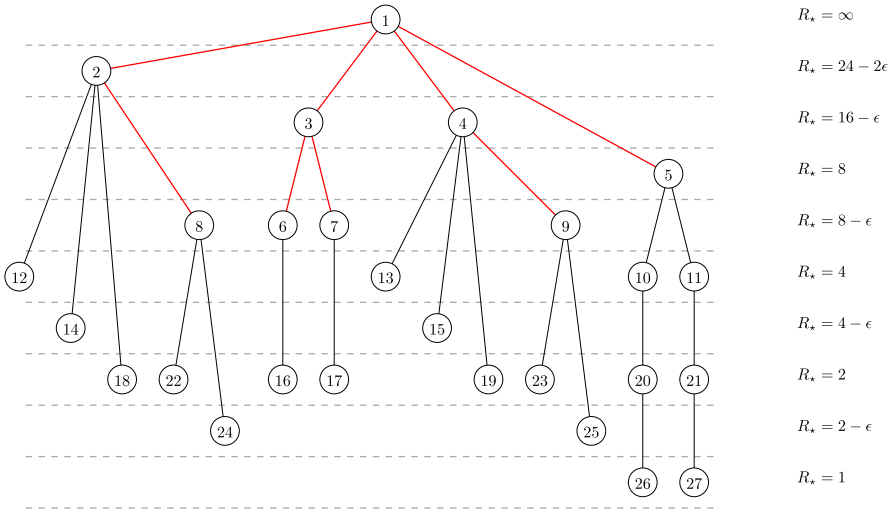

Fig. 8 Here we see the final tree. To obtain the 9-clustering we cut the red edges. The resulting clustering contains the cluster $\{x_5, x_{10}, x_{11}, x_{20}, x_{21}, x_{26}, x_{27}\}$ of radius $14 - 3\epsilon$

For any given $1 \leq k \leq |\mathcal{P}|$ we observe that by deleting the edges $\{x_i, N(x_i)\}$ for all $2 \leq i \leq k$ the tree decomposes into k connected components with vertex sets $H_k^1, \ldots, H_k^k$. We define the k-clustering on $\mathcal{P}$ to be $\mathcal{H}_k = (H_k^1, \ldots, H_k^k)$. Then $\mathscr{H} = (\mathcal{H}_{|\mathcal{P}|}, \ldots, \mathcal{H}_1)$ is a hierarchical clustering of $\mathcal{P}$.

We believe that the algorithm by Mondal does not differ significantly from the algorithm by Dasgupta and Long. Since we already know that the analysis of the approximation guarantee of Dasgupta and Long's algorithm is tight [7] the significant improvement on the approximation guarantee seems surprising. We present an example where Mondal's algorithm in fact computes a $7 - \epsilon$ approximation for some arbitrarily small $\epsilon > 0$, contradicting the claimed approximation guarantee of 6. We believe that this example can be generalized to prove that the approximation guarantee of Mondal's algorithm is at least 8.

Let $\epsilon \in (0, \frac{1}{2})$, Fig. 7 shows a graph with 27 points which need to be clustered. The metric is given by the shortest path metric in the graph. We perform Mondal's algorithm on this instance under the assumption that we can decide how to break ties, whenever they occur.

In Fig. 7 we see the numbering of the points which is computed by Gonzales' algorithm as well as all R-values. Figure 8 shows the resulting tree. We obtain the 9-clustering by cutting all edges $\{x_i, N(x_i)\}$ with $2 \leq i \leq 9$. This clustering contains the cluster $\{x_5, x_{10}, x_{11}, x_{20}, x_{21}, x_{26}, x_{27}\}$, whose radius is $14 - 3\epsilon$, while the radius of the optimal 9-clustering is 2 (see Fig. 7).

Acknowledgements This work has been supported by the German Research Foundation (DFG)—Project Number 416767905.

Author Contributions All authors contributed equally to this work.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—390685813 and 416767905 and by the LamarrInstitute for Machine Learning and Artificial Intelligence (lamarr-institute.org).

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Arutyunova, A., Röglin, H.: The price of hierarchical clustering. In: Chechik, S., Navarro, G., Rotenberg, E., Herman, G. (eds.) 30th Annual European Symposium on Algorithms, ESA 2022, September 5–9, 2022, Berlin/Potsdam, Germany. LIPIcs, vol. 244, pp. 10–11014. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Berlin/Potsdam, Germany (2022). https://doi.org/10.4230/LIPICS.ESA.2022.10
2. Dasgupta, S., Long, P.M.: Performance guarantees for hierarchical clustering. J. Comput. Syst. Sci. **70**(4), 555–569 (2005). https://doi.org/10.1016/j.jcss.2004.10.006
3. Charikar, M., Chekuri, C., Feder, T., Motwani, R.: Incremental clustering and dynamic information retrieval. SIAM J. Comput. **33**(6), 1417–1440 (2004). https://doi.org/10.1137/S0097539702418498
4. Plaxton, C.G.: Approximation algorithms for hierarchical location problems. J. Comput. Syst. Sci. **72**(3), 425–443 (2006). https://doi.org/10.1016/j.jcss.2005.09.004
5. Lin, G., Nagarajan, C., Rajaraman, R., Williamson, D.P.: A general approach for incremental approximation and hierarchical clustering. SIAM J. Comput. **39**(8), 3633–3669 (2010). https://doi.org/10.1137/070698257
6. Großwendt, A.-K.: Theoretical analysis of hierarchical clustering and the shadow vertex algorithm. Ph.D. thesis, University of Bonn (2020). http://hdl.handle.net/20.500.11811/8348

7. Das, A., Kenyon-Mathieu, C.: On hierarchical diameter-clustering and the supplier problem. Theory Comput. Syst. **45**(3), 497–511 (2009). https://doi.org/10.1007/s00224-009-9186-6
8. Bock, F.: Hierarchy cost of hierarchical clusterings. J. Comb. Optim. (2022). https://doi.org/10.1007/s10878-022-00851-4
9. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoret. Comput. Sci. **38**, 293–306 (1985). https://doi.org/10.1016/0304-3975(85)90224-5
10. Mondal, S.A.: An improved approximation algorithm for hierarchical clustering. Pattern Recognit. Lett. **104**, 23–28 (2018). https://doi.org/10.1016/j.patrec.2018.01.015
11. Cohen-Addad, V., Grandoni, F., Lee, E., Schwiegelshohn, C.: Breaching the 2 LMP approximation barrier for facility location with applications to k-median. In: Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023, Florence, Italy, January 22–25, 2023, pp. 940–986. SIAM, Florence, Italy (2023). https://doi.org/10.1137/1.9781611977554.ch37
12. Cohen-Addad, V., Esfandiari, H., Mirrokni, V.S., Narayanan, S.: Improved approximations for Euclidean k-means and k-median, via nested quasi-independent sets. In: STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20–24, 2022, pp. 1621–1628. ACM, Rome, Italy (2022). https://doi.org/10.1145/3519935.3520011
13. Hochbaum, D.S., Shmoys, D.B.: A unified approach to approximation algorithms for bottleneck problems. J. ACM **33**(3), 533–550 (1986). https://doi.org/10.1145/5925.5933
14. Ward, J.H., Jr.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963). https://doi.org/10.1080/01621459.1963.10500845
15. Ackermann, M.R., Blömer, J., Kuntze, D., Sohler, C.: Analysis of agglomerative clustering. Algorithmica **69**(1), 184–215 (2014). https://doi.org/10.1007/s00453-012-9717-4
16. Großwendt, A., Röglin, H.: Improved analysis of complete-linkage clustering. Algorithmica **78**(4), 1131–1150 (2017). https://doi.org/10.1007/s00453-017-0284-6
17. Arutyunova, A., Großwendt, A., Röglin, H., Schmidt, M., Wargalla, J.: Upper and lower bounds for complete linkage in general metric spaces. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM), pp. 18–11822 (2021). https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2021.18
18. Großwendt, A., Röglin, H., Schmidt, M.: Analysis of ward's method. In: Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 2939–2957 (2019). https://doi.org/10.1137/1.9781611975482.182
19. Dasgupta, S.: A cost function for similarity-based hierarchical clustering. In: Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC), pp. 118–127 (2016). https://doi.org/10.1145/2897518.2897527
20. Charikar, M., Chatziafratis, V.: Approximate hierarchical clustering via sparsest cut and spreading metrics. In: Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 841–854 (2017). https://doi.org/10.1137/1.9781611974782.53
21. Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., Mathieu, C.: Hierarchical clustering: objective functions and algorithms. In: Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 378–397 (2018). https://doi.org/10.1137/1.97816119750031.26
22. Wang, Y., Moseley, B.: An objective for hierarchical clustering in Euclidean space and its connection to bisecting k-means. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 6307–6314 (2020). https://doi.org/10.1609/aaai.v34i04.6099
23. Dai, W.: A 16-competitive algorithm for hierarchical median problem. Sci. China Inf. Sci. **57**(3), 1–7 (2014). https://doi.org/10.1007/s11432-014-5065-0