New Approximability Results for the Robust k-Median Problem

Sayan Bhattacharya, Parinya Chalermsook,
Kurt Mehlhorn, and Adrian Neumann

Max-Planck Institut für Informatik
{bsayan,parinya,mehlhorn,aneumann}@mpi-inf.mpg.de

Abstract. We consider a variant of the classical k-median problem, introduced by Anthony et al. [1]. In the *Robust k-Median problem*, we are given an n-vertex metric space (V, d) and m client sets $\{S_i \subseteq V\}_{i=1}^{m}$. We want to open a set $F \subseteq V$ of k facilities such that the worst case connection cost over all client sets is minimized; that is, minimize $\max_i \sum_{v \in S_i} d(F, v)$. Anthony et al. showed an $O(\log m)$ approximation algorithm for any metric and APX-hardness even in the case of uniform metric. In this paper, we show that their algorithm is nearly tight by providing $\Omega(\log m / \log\log m)$ approximation hardness, unless $\mathsf{NP} \subseteq \bigcap_{\delta>0} \mathsf{DTIME}(2^{n^\delta})$. This result holds even for uniform and line metrics. To our knowledge, this is one of the rare cases in which a problem on a line metric is hard to approximate to within logarithmic factor. We complement the hardness result by an experimental evaluation of different heuristics that shows that very simple heuristics achieve good approximations for realistic classes of instances.

1 Introduction

In the classical k-median problem, we are given a set of clients located on a metric space with distance function $d : V \times V \to \mathbb{R}$. The goal is to open a set of facilities $F \subseteq V$, $|F| = k$, so as to minimize the sum of the connection costs of the clients in V, i.e., their distances from their nearest facilities in F. This is a central problem in approximation algorithms, and has received a large amount of attention in the past two decades [4, 6, 7, 11, 12].

At SODA 2008 Anthony et al. [1] introduced a generalization of the k-median problem. In their setting, the set of clients that are to be connected to some facility is not known in advance, and the goal is to perform well in spite of this uncertainty about the future. They formulated the problem as follows.

Definition 1 (Robust k-Median). *An instance of this problem is a triple $(V, \mathcal{S}, d)$. This defines a set of* locations *V, a collection of m sets of* clients *$\mathcal{S} = \{S_1, \ldots, S_m\}$, where $S_i \subseteq V$ for all $i \in \{1, \ldots, m\}$, and a metric distance function $d : V \times V \to \mathbb{R}$. We have to open a set of k facilities $F \subseteq V$, $|F| = k$, and the goal is to minimize the cost of the most expensive set of clients, i.e. minimize $\max_{i=1}^{m} \sum_{v \in S_i} d(v, F)$. Here, $d(v, F)$ denotes the minimum distance of the client v from any location in F, i.e. $d(v, F) = \min_{u \in F} d(u, v)$.*

R Ravi and I.L. Gørtz (Eds.): SWAT 2014, LNCS 8503, pp. 50–61, 2014.

Robust k-Median is a natural generalization of the classical k-median problem (for $m = 1$). Additionally, we can think of it as capturing a notion of *fairness*. To see this, interpret each set S_i as a *group* of clients who pay $\sum_{v \in S_i} d(v, F)$ for connecting to a facility. The objective ensures that no single group pays too much, while minimizing the cost. Anthony et al. [1] gave an $O(\log m)$-approximation algorithm for this problem, and a lower bound of $(2 - \epsilon)$ by a reduction from Vertex Cover. The lower bound was improved to $\log^{\alpha} n$ for small constant $\alpha > 0$ in [5]. Note that their lower bound does not hold in the line metric.

Our Results. We prove nearly tight hardness of approximation for Robust k-Median. We show that, unless $\mathsf{NP} \subseteq \cap_{\delta>0}\mathsf{DTIME}(2^{n^{\delta}})$, it admits no poly-time $o(\log m/\log\log m)$-approximation, *even on uniform and line metrics*.

Our first hardness result is tight up to a constant factor, as a simple rounding scheme gives a matching upper bound on uniform metrics (Sect. 3.1). Our second result shows that Robust k-Median is a rare problem with super-constant hardness of approximation even on line metrics. This surprising result puts Robust k-Median in sharp contrast to most other geometric optimization problems which admit polynomial time approximation schemes, e.g. [2, 10].

Experimentally we show that simple heuristics provide good performance on a realistic class of instances. The details appear in the full paper.

Our Techniques. First, we note that Robust k-Median on uniform metrics is equivalent to the following variant of the set cover problem: Given a set U of ground elements, a collection of sets $\mathcal{X} = \{X \subseteq U\}$, and an integer $t \leq |\mathcal{X}|$, our goal is to select t sets from $\mathcal{X}$ in order to minimize the number of times an element from U is hit (Lemma 2). We call this problem Minimum Congestion Set Packing (MCSP). This characterization allows us to focus on proving the hardness of MCSP, and to employ the tools developed for the set cover problem.

We now revisit the reduction used by Feige [8], building on results of Lund and Yannakakis [13], to prove the hardness of the set cover problem and discuss how our approach differs. Intuitively, they compose the Label Cover instance with a set system that has some desirable properties. Informally speaking, in the Label Cover problem, we are given a graph where each vertex v can be assigned a label from a set L, and each edge e is equipped with a constraint $\Pi_e \subseteq L \times L$ specifying the accepting pairs of labels for e. Our goal is to find a labeling of vertices that maximizes the number of accepting edges. This problem is known to be hard to approximate to within a factor of $2^{\log^{1-\epsilon}|E|}$ [3, 14], where $|E|$ is the number of edges. Thus, if we manage to reduce Label Cover to MCSP, we would hopefully obtain a large hardness of approximation factor for MCSP as well.

From the Label Cover instance, [13] creates an instance of Set Cover by having sets of the form $S(v, \ell)$ for each vertex v and each label $\ell \in L$. Intuitively the set $S(v, \ell)$ means choosing label ℓ for vertex v in the Label Cover instance. Now, if we assume that the solution is well behaved, in the sense that for each vertex v, only one set of the form $S(v, \ell)$ is chosen in the solution, we would be immediately done (because each set indeed corresponds to a label). However, solutions need

not have this form, e.g. choosing sets $S(v, \ell)$ and $S(v, \ell')$ translates to having two labels ℓ, ℓ' for the Label Cover instance. To prevent an ill-behaved solution, *partition systems* were introduced and used in both [13] and [8]. Feige considers the hypergraph version of Label Cover to obtain a sharper hardness result of $\ln n - O(\ln \ln n)$ instead of $\frac{1}{4} \ln n$ in [13]; here n denotes the size of the universe.

Now we highlight how our reduction is different. The high level idea stays the same, i.e. we have sets of the form $S(v, \ell)$ that represent assigning label ℓ to vertex v. However, we need a different partition system and a totally different analysis. Moreover, while a reduction from standard Label Cover gives nearly tight $O(\log n)$ hardness for Set Cover, it can (at best) only give a $2 - \epsilon$ hardness for MCSP. For our results, we do need a reduction from Hypergraph Label Cover. This suggests another natural distinction between MCSP and Set Cover.

Finally, to obtain the hardness result for the line metric, we embed the instance created from the MCSP reduction onto the line while preserving values of optimal solutions. This way we get the same hardness gap for line metrics.

2 Preliminaries

We will show that Robust k-Median is $\Omega(\log m / \log \log m)$ hard to approximate, even for the special cases of *uniform metrics* (Sect. 3) and *line metrics* (Sect. 4). Recall that d is a uniform metric iff we have $d(u, v) \in \{0, 1\}$ for all locations $u, v \in V$. Further, d is a line metric iff the locations in V can be embedded into a line in such a way that $d(u, v)$ equals the euclidean distance between u and v, for all $u, v \in V$. Throughout this paper, we will denote any set of the form $\{1, 2, \ldots, i\}$ by $[i]$. Our hardness results will rely on a reduction from the *r-Hypergraph Label Cover* (HGLC) problem, which is defined as follows.

Definition 2 (r-Hypergraph Label Cover (HGLC)). *An instance of this problem is a triple (G, π, r), where $G = (\mathcal{V}, \mathcal{E})$ is a r-partite hypergraph with vertex set $\mathcal{V} = \bigcup_{j=1}^{r} \mathcal{V}_j$ and edge set $\mathcal{E}$. Each edge $h \in \mathcal{E}$ contains one vertex from each part of $\mathcal{V}$, i.e. $|h \cap \mathcal{V}_j| = 1$ for all $j \in [r]$. Every set $\mathcal{V}_j$ has an associated set of* labels *L_j. Further, for all $h \in \mathcal{E}$ and $j \in [r]$, there is a mapping $\pi_h^j : L_j \to C$ that projects the labels from L_j to a common set of* colors *C.*

The problem is to assign to every vertex $v \in \mathcal{V}_j$ some label $\sigma(v) \in L_j$. We say that an edge $h = (v_1, \ldots, v_r)$, where $v_j \in \mathcal{V}_j$ for all $j \in [r]$, is strongly satisfied *under σ iff the labels of all its vertices are mapped to the same element in C, i.e. $\pi_h^j(\sigma(v_j)) = \pi_h^{j'}(\sigma(v_{j'}))$ for all $j, j' \in [r]$. In contrast, we say that the edge is* weakly satisfied *iff there exists some pair of vertices in h whose labels are mapped to the same element in C, i.e. $\pi_h^j(\sigma(v_j)) = \pi_h^{j'}(\sigma(v_{j'}))$ for some $j, j' \in [r]$, $j \neq j'$.*

For ease of exposition, we will often abuse the notation and denote by $j(v)$ the part of $\mathcal{V}$ to which a vertex v belongs, i.e. if $v \in \mathcal{V}_j$ for some $j \in [r]$, then we set $j(v) \leftarrow j$. The next theorem will be crucial in deriving our hardness result. The proof of this theorem follows from Feige's r-Prover system [8].

Theorem 1. *Let $r \in \mathbb{N}$ be a parameter. There is a polynomial time reduction from n-variable 3-SAT to r-HGLC with the following properties:*

- *(Yes-Instance) If the formula is satisfiable, there is a labeling that strongly satisfies every edge in G.*
- *(No-Instance) If the formula is not satisfiable, every labeling weakly satisfies at most a $2^{-\gamma r}$ fraction of the edges in G, for some universal constant γ.*
- *The number of vertices in the graph is $|\mathcal{V}| = n^{O(r)}$ and the number of edges is $|\mathcal{E}| = n^{O(r)}$. The sizes of the label sets are $|L_j| = 2^{O(r)}$ for all $j \in [r]$, and $|C| = 2^{O(r)}$. Further, we have $|\mathcal{V}_j| = |\mathcal{V}_{j'}|$ for all $j, j' \in [r]$, and each vertex $v \in \mathcal{V}$ has the same degree $r|\mathcal{E}|/|\mathcal{V}|$.*

We use a *partition system* that is motivated by the hardness proof of the Set Cover problem [8] but uses a different construction.

Definition 3 (Partition System). *Let $r \in \mathbb{N}$ and let C be any finite set. An (r, C)-partition system is a pair $(Z, \{p_c\}_{c \in C})$, where Z is an arbitrary (ground) set, such that the following properties hold.*

- *(Partition) For all $c \in C$, $p_c = \left(A_c^1, \ldots, A_c^r\right)$ is a partition of Z, that is $\bigcup_{j=1}^r A_c^j = Z$, and $A_c^{j'} \cap A_c^j = \emptyset$ for all $j, j' \in [r], j \neq j'$.*
- *(r-intersecting) For any r* distinct *indices $c_1, \ldots, c_r \in C$ and* not-necessarily distinct *indices $j_1, \ldots, j_r \in [r]$, we have that $\bigcap_{i=1}^r A_{c_i}^{j_i} \neq \emptyset$. In particular, $A_c^j \neq \emptyset$ for all c and j.*

In order to achieve a good lower bound on the approximation factor, we need partition systems with *small* ground sets. The most obvious way to build a partition system is to form an r-hypercube: Let $Z = [r]^{|C|}$, and for each $c \in C$ and $j \in [r]$, let A_c^j be the set of all elements in Z whose c-th component is j. It can easily be verified that this is an (r, C)-partition system with $|Z| = r^{|C|}$. With this construction, however, we would only get a hardness of $\Omega(\log \log m)$ for our problem. The following lemma shows that it is possible to construct an (r, C)-partition system probabilistically with $|Z| = r^{O(r)} \log |C|$.

Lemma 1. *There is an (r, C)-partition system with $|Z| = r^{O(r)} \log |C|$ elements. Further, such a partition system can be constructed efficiently with high probability.*

Proof. Let Z be any set of $r^{O(r)} \log |C|$ elements. We build a partition system $(Z, \{p_c\}_{c \in C})$ as described in Algorithm 1. By construction each p_c is a partition of Z, i.e. the first property stated in Def. 3 is satisfied. We bound the probability that the second property is violated.

Fix any choice of r *distinct* indices $c_1, \ldots, c_r \in C$ and *not necessarily distinct* indices $j_1, \ldots, j_r \in [r]$. We say that a *bad event* occurs when the intersection of the corresponding sets is empty, i.e. $\bigcap_{i=1}^r A_{c_i}^{j_i} = \emptyset$. To upper bound the probability of a bad event, we focus on events of the form $E_{e,i}$ – this occurs when an element $e \in Z$ is included in a set $A_{c_i}^{j_i}$. Since the indices $c_1 \ldots c_r$ are distinct, it follows that the events $\{E_{e,i}\}$ are mutually independent. Furthermore, note that we have $\Pr[E_{e,i}] = 1/r$ for all $e \in Z, i \in [r]$. Hence, the probability that an element $e \in Z$ does not belong to the intersection $\bigcap_{i=1}^r A_{c_i}^{j_i}$ is given by

Algorithm 1. A randomized construction of an (r, C)-partition system.

input : A ground set Z, a parameters $r \in \mathbb{N}$, and a set C.
foreach $c \in C$ **do**
  /* Construct the partition $p_c = (A_c^1, \ldots, A_c^r)$ */
  Initialize A_c^j to the empty set for all $j \in [r]$
  foreach *ground element* $e \in Z$ **do**
    Pick a $j \in [r]$ independently and uniformly at random and add e to A_c^j

$1 - \Pr[\bigcap_{i=1}^{r} E_{e,i}] = 1 - 1/r^r$. Accordingly, the probability that no element $e \in Z$ belongs to the intersection, which defines the bad event, is equal to $(1-1/r^r)^{|Z|}$.

Now, the number of choices for r distinct indices $c_1, \ldots, c_r$ and r not-necessarily distinct indices $j_1, \ldots, j_r$ is equal to $\binom{|C|}{r} \cdot r^r$. Hence, by a union-bound over all bad events, the second property stated in Def. 3 is violated with probability at most $\binom{|C|}{r} \cdot r^r \cdot (1-r^r)^{|Z|} \leq (|C|\, r)^r \cdot \exp(-|Z|/r^r)$. If we set $|Z| = d \cdot r^{d \cdot r} \log |C|$ with large enough constant d, the property is satisfied with high probability. □

3 Hardness of Robust k-Median on Uniform Metrics

First, we define *Minimum Congestion Set Packing* (MCSP), and then show a reduction from MCSP to Robust k-Median on uniform metrics. In Sect. 3.2, we will then show that MCSP is hard to approximate by reducing HGLC to MCSP.

Definition 4 (Minimum Congestion Set Packing (MCSP)). *An instance of this problem is a triple $(U, \mathcal{X}, t)$, where U is a universe of m elements, i.e. $|U| = m$, $\mathcal{X}$ is a collection of sets $\mathcal{X} = \{X \subseteq U\}$ such that $\bigcup_{X \in \mathcal{X}} X = U$, and $t \in \mathbb{N}$ and $t \leq |\mathcal{X}|$. The objective is to find a collection $\mathcal{X}' \subseteq \mathcal{X}$ of size t that minimizes $\textsc{Cong}(\mathcal{X}') = \max_{e \in U} \textsc{Cong}(e, \mathcal{X}')$. Here, $\textsc{Cong}(\mathcal{X}')$ refers to the* congestion *of the solution $\mathcal{X}'$, and $\textsc{Cong}(e, \mathcal{X}') = |\{X \in \mathcal{X}' : e \in X\}|$ is the congestion of the element $e \in U$ under the solution $\mathcal{X}'$.*

Lemma 2. *Given any MCSP instance $(U, \mathcal{X}, t)$, we can construct a Robust k-Median instance $(V, \mathcal{S}, d)$ with the same objective value in poly$(|U|, |\mathcal{X}|)$ time, such that $|U| = |\mathcal{S}|$, $|\mathcal{X}| = |V|$, d is a uniform metric, and $k = |V| - t$.*

Proof. We construct the Robust k-Median instance $(V, \mathcal{S}, d)$ as follows. For every $e \in U$ we create a set of clients $S(e)$, and for each $X \in \mathcal{X}$ we create a location $v(X)$. Thus, we get $V = \{v(X) : X \in \mathcal{X}\}$, and $\mathcal{S} = \{S(e) : e \in U\}$. We place the clients in $S(e)$ at the locations of the sets that contain e, i.e. $S(e) = \{v(X) : X \in \mathcal{X}, e \in X\}$ for all $e \in U$. The distance is defined as $d(u, v) = 1$ for all $u, v \in V, u \neq v$, and $d(v, v) = 0$. Finally, we set $k \leftarrow |V| - t$.

Now, it is easy to verify that the Robust k-Median instance $(V, \mathcal{S}, d)$ has a solution with objective ρ iff the corresponding MCSP instance $(U, \mathcal{X}, t)$ has a solution with objective ρ. The intuition is that a location $v(X) \in V$ is *not* included in the solution F to the Robust k-Median instance iff the corresponding set X is included in the solution $\mathcal{X}'$ to the MCSP instance. Indeed, let F be any

subset of $\mathcal{X}$ of size k (= the set of open facilities) and let $\mathcal{X}' = \mathcal{X} - F$. Further, let $[X \in \mathcal{X}']$ be an indicator variable that is set to 1 iff $X \in \mathcal{X}'$. Then

$$\begin{aligned}\textsc{Cong}(\mathcal{X}') &= \max_{e \in U} \textsc{Cong}(e, \mathcal{X}') = \max_{e \in U} \sum_{X; e \in X} [X \in \mathcal{X}'] \\ &= \max_{e \in U} \sum_{X; e \in X} \min_{Y \in F} d(X, Y) = \max_{S(e) \in \mathcal{S}} \sum_{v(X) \in S(e)} d(v(X), F).\end{aligned}$$

□

We devote the rest of Sect. 3 to MCSP and show that it is $\Omega(\log |U| / \log \log |U|)$ hard to approximate. This, in turn, will imply a $\Omega(\log |\mathcal{S}| / \log \log |\mathcal{S}|)$ hardness of approximation for Robust k-Median on uniform metrics. We will prove the hardness result via a reduction from HGLC.

3.1 Integrality Gap

Before proceeding to the hardness result, we show that a natural LP relaxation for the MCSP problem [1] has an integrality gap of $\Omega(\log m / \log \log m)$, where $m = |U|$ is the size of the universe of elements. In the LP, we have a variable $y(X)$ indicating that the set $X \in \mathcal{X}$ is chosen, and a variable z which represents the maximum congestion among the elements.

$$\begin{aligned}\min \quad & z \\ \text{s.t.} \quad & \sum_{X \in \mathcal{X}: e \in X} y(X) \leq z \text{ for all } e \in U \\ & \sum_{X \in \mathcal{X}} y(X) = t\end{aligned}$$

The Instance: Now, we construct a bad integrality gap instance $(U, \mathcal{X}, t)$. Let d be the intended integrality gap, let $\eta = d^2$, and let $U = \{I : I \subseteq [\eta], |I| = d\}$ be all subsets of $[\eta]$ of size d. The collection $\mathcal{X}$ consists of η sets $X_1, \ldots, X_\eta$, where $X_i = \{I : I \in U \text{ and } i \in I\}$. Note that the universe U consists of $|U| = m = \binom{\eta}{d}$ elements, and each element I is contained in exactly d sets, namely $I \in X_i$ if and only if $i \in I$. Finally, we set $t \leftarrow \eta/d$.

Analysis: The fractional solution simply assigns a value of $1/d$ to each variable $y(X_i)$; this ensures that the total (fractional) number of sets selected is $\eta/d = t$. Furthermore, each element is contained (fractionally) in exactly one set, so the fractional solution has cost one. Since $t = \eta/d = d$, any integral solution must choose d sets, say $X_{i_1}, \ldots, X_{i_d}$. Now consider $I = \{i_1, \ldots, i_d\}$ which belongs to set X_{i_λ} for all $\lambda \in [d]$ and hence the congestion of I is d. Finally, since $|U| = m \leq \eta^d \leq (d^2)^d$, we have $d = \Omega(\log m / \log \log m)$.

Tightness of the Result: The bound on the hardness and integrality gap is tight for the uniform metric case, as there is a simple $O(\log m / \log \log m)$-approximation algorithm. Pick each set X with probability equal to $\min(1, 2y(X))$.

The expected congestion is $2z$ for each element. By Chernoff's bound [9], an element is covered by no more than $z \cdot O(\log m / \log\log m)$ sets with high probability. A similar algorithm gives the same approximation guarantee for Robust k-Median on uniform metrics.

3.2 Reduction from r-Hypergraph Label Cover to Minimum Congestion Set Packing

The input is an instance (G, π, r) of r-HGLC (Def. 2). From this we construct the following instance $(U, \mathcal{X}, t)$ of MCSP (Def. 4).

- We define the universe U as a union of disjoint sets. For each edge $h \in \mathcal{E}$ in the hypergraph we have a set U_h. All these sets have the same size m^* and are pairwise disjoint, i.e. $U_h \cap U_{h'} = \emptyset$ for all $h, h' \in \mathcal{E}$, $h' \neq h$. The universe U is then the union of these sets $U = \bigcup_{h \in \mathcal{E}} U_h$. Since the U_h are mutually disjoint, we have $m = |U| = |\mathcal{E}| \cdot m^*$. Recall that C is the target set of π. Each set U_h is the ground set of an (r, C)-partition system (Def. 3) as given by Lemma 1. In particular we have $m^* = r^{O(r)} \log |C|$. We denote the r-partitions associated with U_h by $\{p_c(h)\}_{c \in C}$, where $p_c(h) = \big(A_c^1(h), \ldots, A_c^r(h)\big)$.
- We construct the collection of sets $\mathcal{X}$ as follows. For each $j \in [r]$, $v \in \mathcal{V}_j$ and $\ell \in L_j$, $\mathcal{X}$ contains the set $X(v, \ell)$, where $X(v, \ell) = \bigcup_{h : v \in h} A^j_{\pi^j_h(\ell)}(h)$. That is, $X(v, \ell) \cap U_h$ is empty if $v \notin h$ and is equal to $A^j_{\pi^j_h(\ell)}(h)$ if $v \in h$. Intuitively, choosing the set $X(v, \ell)$ corresponds to assigning label ℓ to the vertex v.
- We define $t \leftarrow |\mathcal{V}|$. Intuitively, this means each vertex in $\mathcal{V}$ gets one label.

We assume for the sequel that the r-HGLC instance is chosen according to Thm. 1. We assume that the parameter r satisfies $r^7 2^{-\gamma r} < 1$. In the proof of the main theorem, we will fix r to a specific value.

3.3 Analysis

We show that the reduction from HGLC to MCSP satisfies two properties. In Lemma 3, we show that for Yes-Instances (see Thm. 1) the corresponding MCSP instance admits a solution with congestion one. For No-Instances, Lemma 4 shows that any solution to the corresponding MCSP instance has congestion at least r.

Lemma 3 (Yes-Instance). *If the HGLC instance (G, π, r) admits a labeling that strongly satisfies every edge, then the MCSP instance $(U, \mathcal{X}, t)$ as in Sect. 3.2 admits a solution where the congestion of every element in U is exactly one.*

Proof. Suppose that there is a labeling σ that strongly satisfies every edge $h \in \mathcal{E}$. We will show how to pick $t = |\mathcal{V}|$ sets from $\mathcal{X}$ such that each element in U is contained in exactly one set. This implies that the maximum congestion is one. For each $j \in [r]$ and each vertex $v \in \mathcal{V}_j$, we choose the set $X(v, \sigma(v))$. Thus, the total number of sets chosen is exactly $|\mathcal{V}|$.

To see that the congestion is one, we concentrate on the elements in U_h, where $h = (v_1, \ldots, v_r)$, $v_j \in \mathcal{V}_j$ for all $j \in [r]$, is one of the edges in $\mathcal{E}$. The picked sets

that intersect U_h are $X(v_j, \sigma(v_j))$, where $j \in [r]$. Since h is strongly satisfied, π_h maps all vertex labels in h to a common $c \in C$, i.e. $\pi_h^j(\sigma(v_j)) = c$ for all $j \in [r]$. Thus $U_h \cap X(v_j, \sigma(v_j)) = A_c^j(h)$. By definition (Def. 3), the sets $A_c^1(h) \dots A_c^r(h)$ partition the elements in U_h. This completes the proof. □

Now, we turn to the proof of Lemma 4. Towards this end, we fix a collection $\mathcal{X}' \subseteq \mathcal{X}$ of size t and show that some element in U has congestion at least r under $\mathcal{X}'$. The intuition being that many edges in $G = (\mathcal{V}, \mathcal{E})$ are not even weakly satisfied, and the elements in U corresponding to those edges incur large congestion. Recall that for a $v \in \mathcal{V}$, we define $j(v) \in \mathbb{N}$ to be such that $v \in \mathcal{V}_{j(v)}$.

Claim 2. *For $v \in \mathcal{V}$, let $\mathcal{L}_v = \left\{\ell \in L_{j(v)} : X(v, \ell) \in \mathcal{X}'\right\}$. For $h \in \mathcal{E}$, let $\Lambda_h = \{X(v, \ell) \in \mathcal{X}' : v \in h\}$. If $\textsc{Cong}(\mathcal{X}') < r$, then $|\mathcal{L}_v| < r^2$ and $|\Lambda_h| < r^3$.*

Proof. Since $\Lambda_h = \bigcup_{v \in h} \mathcal{L}_v$, it suffices to prove $|\mathcal{L}_v| < r^2$ for all v. Assume otherwise, i.e., $|\mathcal{L}_v| \geq r^2$ for some $v \in \mathcal{V}_j$, $j \in [r]$. Let h be any hyper-edge with $v \in h$. Consider the images of the labels in $\mathcal{L}_v$ under π_h^j. Either there are at least r distinct images or at least r elements in L_v are mapped to the same $c \in C$.

In the former case, we have r pairwise distinct labels ℓ_1 to ℓ_r in $\mathcal{L}_v$ and r pairwise distinct labels c_1 to c_r in C such that $\pi_h^j(\ell_i) = c_i$ for $i \in [r]$. The set $X(v, \ell_i)$ contains $A_{c_i}^j(h)$ and $\bigcap_{i \in [r]} A_{c_i}^j(h) \neq \emptyset$ by property (2) of partition systems (Def. 3). Thus some element has congestion at least r.

In the latter case, we have r pairwise distinct labels ℓ_1 to ℓ_r in $\mathcal{L}_v$ and a label c in C such that $\pi_h^j(\ell_i) = c$ for $i \in [r]$. The set $X(v, \ell_i)$ contains $A_c^j(h)$ and hence every element in this non-empty set (property (2) of partition systems) has congestion at least r. □

Definition 5 (Colliding Edge). *We say that an edge $h \in \mathcal{E}$ is* colliding *iff there are sets $X(v, \ell), X(v', \ell') \in \mathcal{X}'$ with $v, v' \in h$, $v \neq v'$, and $\pi_h^{j(v)}(\ell) = \pi_h^{j(v')}(\ell')$.*

Claim 3. *Suppose that the solution $\mathcal{X}'$ has congestion less than r, and more than a $r^4 2^{-\gamma r}$ fraction of the edges in $\mathcal{E}$ are colliding. Then there is a labeling σ for G that weakly satisfies at least a $2^{-\gamma r}$ fraction of the edges in $\mathcal{E}$.*

Proof. For each $v \in \mathcal{V}$, we define $\mathcal{L}_v = \{\ell \in L_{j(v)} : X(v, \ell) \in \mathcal{X}'\}$. Then $|\mathcal{L}_v| < r^2$ by Claim 2. We construct a labeling function σ using Algorithm 2.

Now we bound the expected fraction of weakly satisfied edges under σ from below. Take any colliding edge $h \in \mathcal{E}$. Then there are vertices $v \in \mathcal{V}_j$, $v' \in \mathcal{V}_{j'}$ with

Algorithm 2. An algorithm for constructing a labeling function.

```
foreach vertex v ∈ V do
    if L_v ≠ ∅ then
        Pick a color σ(v) uniformly and independently at random from L_v
    else
        Pick an arbitrary color σ(v) from L_{j(v)}
```

$j \neq j'$, and colors $\ell \in \mathcal{L}_v$, $\ell' \in \mathcal{L}_{v'}$ such that $v, v' \in h$ and $\pi_h^j(\ell) = \pi_h^{j'}(\ell')$. By Claim 2, $|\mathcal{L}_v|$ and $|\mathcal{L}_{v'}|$ are both at most r^2. Since the colors $\sigma(v)$ and $\sigma(v')$ are chosen uniformly and independently at random from their respective palettes $\mathcal{L}_v$ and $\mathcal{L}_{v'}$, we have $\Pr[\sigma(v) = \ell \text{ and } \sigma(v') = \ell'] \geq 1/r^4$. In other words, every colliding edge is weakly satisfied with probability at least $1/r^4$. Since more than a $r^4 2^{-\gamma r}$ fraction of the edges in $\mathcal{E}$ are colliding, from linearity of expectation we infer that the expected fraction of edges weakly satisfied by σ is at least $2^{-\gamma r}$. □

Claim 4. *Let* $\Lambda_h = \{X(v,\ell) \in \mathcal{X}' : v \in h\}$ *and* $\lambda(h) = |\Lambda_h|$. $\sum_{h \in \mathcal{E}} \lambda(h) = r|\mathcal{E}|$.

Proof. This is a simple counting argument. Consider a bipartite graph H with vertex set $A \dot{\cup} B$, where each vertex in A represents a set $X(v,\ell)$, and each vertex in B represents an edge $h \in \mathcal{E}$. There is an edge between two vertices iff the set $X(v,\ell)$ contains some element in U_h. The quantity $\sum_{h \in \mathcal{E}} \lambda(h)$ counts the number of edges in H where one endpoint is included in the solution $\mathcal{X}'$. Since $\mathcal{X}'$ picks $t = |\mathcal{V}|$ sets and each set has degree $r|\mathcal{E}|/|\mathcal{V}|$ in H (Thm. 1), the total number of edges that are chosen is exactly $|\mathcal{V}| \times (r|\mathcal{E}|/|\mathcal{V}|) = r|\mathcal{E}|$. □

Let $\mathcal{E}' \subseteq \mathcal{E}$ denote the set of colliding edges, and define $\mathcal{E}'' = \mathcal{E} - \mathcal{E}'$. Suppose that we are dealing with a No-Instance (Thm. 1), i.e. the solution $\mathcal{X}'$ has congestion less than r and every labeling weakly satisfies at most a $2^{-\gamma r}$ fraction of the edges in $\mathcal{E}$. Then $\lambda(h) \leq r^3$ for all $h \in \mathcal{E}$ by Claim 2, and no more than $r^4 2^{-\gamma r}|\mathcal{E}|$ edges are colliding, i.e. $|\mathcal{E}'| \leq r^4 2^{-\gamma r}\,|\mathcal{E}|$, by Claim 3. Using these facts we conclude that $\sum_{h \in \mathcal{E}'} \lambda(h) \leq r^7 2^{-\gamma r}|\mathcal{E}| < |\mathcal{E}|$, as by assumption $r^7 2^{-\gamma r} < 1$. Now, applying Claim 4, we get $\sum_{h \in \mathcal{E}''} \lambda(h) = r|\mathcal{E}| - \sum_{h \in \mathcal{E}'} \lambda(h) > (r-1)|\mathcal{E}|$. In particular, there is an edge $h \in \mathcal{E}''$ with $\lambda(h) \geq r$.

Recall that $\Lambda_h = \{X(v,\ell) \in \mathcal{X}' : v \in h\}$ are the sets in $\mathcal{X}'$ that intersect U_h and note that $|\Lambda_h| = \lambda(h) \geq r$. Let $\mathcal{X}^* \subseteq \Lambda_h$ be a *maximal* collection of sets with the following property: For every two distinct sets $X(v,\ell), X(v',\ell') \in \mathcal{X}^*$ we have $\pi_h^{j(v)}(\ell) \neq \pi_h^{j(v')}(\ell')$. Hence, from the definition of a partition system (Def. 3), it follows that the intersection of the sets in $\mathcal{X}^*$ and the set U_h is nonempty.

Now, consider any set $X(v,\ell) \in \Lambda_h - \mathcal{X}^*$. Since the collection $\mathcal{X}^*$ is maximal, there must be at least one set $X(v',\ell')$ in $\mathcal{X}^*$ with $\pi_h^{j(v)}(\ell) = \pi_h^{j(v')}(\ell')$. Since h is not colliding, we must have $j(v) = j(v')$. Consequently we get $X(v,\ell) \cap U_h = X(v',\ell') \cap U_h$. In other words, for every set $X \in \Lambda_h - \mathcal{X}^*$, there is some set $X' \in \mathcal{X}^*$ where $X \cap U_h = X' \cap U_h$. Thus, $U_h \cap (\bigcap_{X \in \Lambda_h} X) = U_h \cap (\bigcap_{X \in \mathcal{X}^*} X) \neq \emptyset$. Every element in the intersection of the sets in Λ_h and U_h will have congestion $|\Lambda_h| \geq r$. This leads to the following lemma.

Lemma 4 (No-Instance). *If every labeling weakly satisfies at most a* $2^{-\gamma r}$ *fraction of the edges in the hypergragph Label Cover instance* (G, π, r)*, for some universal constant* γ *and* $r^7 2^{-\gamma r} < 1$ *then the congestion incurred by every solution to the MCSP instance* $(U, \mathcal{X}, t)$ *constructed in Sect. 3.2 is at least* r*.*

We are now ready to prove the main theorem of this section.

Theorem 5. *Robust k-Median* $(V, \mathcal{S}, d)$ *is* $\Omega(\log m / \log\log m)$ *hard to approximate on uniform metrics, where* $m = |\mathcal{S}|$*, unless* $\mathsf{NP} \subseteq \bigcap_{\delta > 0} \mathsf{DTIME}(2^{n^\delta})$.

Proof. Assume that there is a polynomial time algorithm for Robust k-Median that guarantees an approximation ratio in $o(\log |\mathcal{S}| \,/\, \log\log |\mathcal{S}|)$. Then, by Lemma 2, there is an approximation algorithm for the Minimum Congestion Set Packing problem with approximation guarantee $o(\log |U| \,/\, \log\log |U|)$.

Let $\delta > 0$ be arbitrary and set $r = \lfloor n^{\delta} \rfloor$, where n is the number of variables in the 3-SAT instance (Thm. 1). Then $r^7 2^{-\gamma r} < 1$ for all sufficiently large n. We first bound the size of the MCSP instance $(U, \mathcal{X}, t)$ constructed in Sect. 3.2. By Lemma 1, the size of an (r, C)-partition system is $|Z| = r^{O(r)} \log |C|$. By Thm. 1, we have $|C| = 2^{O(r)}$. So each set U_h has cardinality at most $r^{O(r)} \cdot r = r^{O(r)}$. Also recall that the number of sets in the MCSP instance is $|\mathcal{X}| = \sum_{j \in [r]} |\mathcal{V}_j| \cdot |L_j| = n^{O(r)}$, and that the number of elements is $|U| = m = |\mathcal{E}| \cdot r^{O(r)} \leq (nr)^{O(r)} = n^{O(r)} = n^{O(n^{\delta})} = 2^{O(r \log r)}$. Thus $r \geq \Omega(\log m / \log\log m)$.

The gap in the optimal congestion between the Yes-Instance and the No-Instance is at least r (Thm. 1 and Lemmas 3, 4). More precisely, for Yes-instances the congestion is at most one and for No-instances it is at least r. Since the approximation ratio of the alleged algorithm is $o(\log m / \log\log m)$, it is better than r for all sufficiently large n and hence it can be used to decide SAT.

The running time is polynomial in the size of the MCSP instance, i.e., is $\text{poly}(n^{O(n^{\delta})}) = n^{O(n^{\delta})} = 2^{O(n^{2\delta})}$. Since δ is arbitrary, the theorem follows. □

4 Hardness of Robust k-Median on Line Metrics

We modify the reduction from r-HGLC to Minimum Congestion Set Packing (MCSP) to give a $\Omega(\log m / \log\log m)$ hardness of approximation for Robust k-Median on line metrics as well, where $m = |\mathcal{S}|$ is the number of client-sets. For this section, it is convenient to assume that the label-sets are the initial segments of the natural numbers, i.e., $L_j = \{1, \ldots, |L_j|\}$ and $C = \{1, \ldots, |C|\}$.

Given a HGLC instance (G, π, r), we first construct a MCSP instance $(U, \mathcal{X}, t)$ in accordance with the procedure outlined in Sect. 3.2. Next, from this MCSP instance, we construct a Robust k-Median instance $(V, \mathcal{S}, d)$ as described below.

- We create a location in V for every set $X(v, \ell) \in \mathcal{X}$. To simplify the notation, the symbol $X(v, \ell)$ will represent both a set in the instance $(U, \mathcal{X}, t)$, and a location in the instance $(V, \mathcal{S}, d)$. Thus, we have $V = \{X(v, \ell) \in \mathcal{X}\}$. Furthermore, we create a set of clients $S(e)$ for every element $e \in U$, which consists of all the locations whose corresponding sets in the MCSP instance contain the element e. Thus, we have $\mathcal{S} = \{S(e) : e \in U\}$, where $S(e) = \{X(v, \ell) \in \mathcal{X} : e \in X(v, \ell)\}$ for all $e \in U$. This step is same as in Lemma 2.
- We now describe how to embed the locations in V on a given line. For every vertex $v \in \mathcal{V}_j, j \in [r]$, the locations $X(v, 1), \ldots, X(v, |L_j|)$ are placed next to one another in sequence, in such a way that the distance between any two consecutive locations is exactly one. Formally, this gives $d(X(v, \ell), X(v, \ell')) = |\ell' - \ell|$ for all $\ell, \ell' \in L_j$. Furthermore, we ensure that any two locations corresponding to two different vertices in $\mathcal{V}$ are *not close to each other*. To be more specific, we have the following guarantee: $d(X(v, \ell), X(v', \ell')) \geq 2$ whenever $v \neq v'$. It is easy to verify that d is a line metric.

– We define $k \leftarrow |\mathcal{X}| - t$.

Note that as $k = |\mathcal{X}| - t$, there is a one to one correspondence between the solutions to the MCSP instance and the solutions to the Robust k-Median instance. Specifically, a set in $\mathcal{X}$ is picked by a solution to the MCSP instance iff the corresponding location is *not* picked in the Robust k-Median instance.

Lemma 5 (Yes-Instance). *Suppose that there is a labeling strategy σ that strongly satisfies every edge in the HGLC instance (G, π, r). Then there is a solution to the Robust k-Median instance $(V, \mathcal{S}, d)$ with objective one.*

Proof. Recall the proof of Lemma 3. We construct a solution $\mathcal{X}' \subseteq \mathcal{X}$, $|\mathcal{X}'| = t$, to the MCSP instance $(U, \mathcal{X}, t)$ as follows. For every $v \in \mathcal{V}_j, j \in [r]$, the solution $\mathcal{X}'$ contains the set $X(v, \sigma(v))$. Now, focus on the corresponding solution $F_{\mathcal{X}'} \subseteq V$ to the Robust k-Median instance, which picks a location X iff $X \notin \mathcal{X}'$. Hence, for every vertex $v \in \mathcal{V}_j$, $j \in [r]$, all but one of the locations $X(v, 1), \ldots, X(v, |L_j|)$ are included in $F_{\mathcal{X}'}$. Since any two consecutive locations in such a sequence are unit distance away from each other, the cost of connecting any location in V to the set $F_{\mathcal{X}'}$ is either zero or one, i.e., $d(X, F_{\mathcal{X}'}) \in \{0, 1\}$ for all $X \in V = \mathcal{X}$.

For the rest of the proof, fix any set of clients $S(e) \in \mathcal{S}$, $e \in U$. The proof of Lemma 3 implies that the element e incurs congestion one under $\mathcal{X}'$. Hence, the element belongs to exactly one set in $\mathcal{X}'$, say X^*. Again, comparing the solution $\mathcal{X}'$ with the corresponding solution $F_{\mathcal{X}'}$, we infer that $S(e) - F_{\mathcal{X}'} = \{X^*\}$. In other words, every location in $S(e)$, except X^*, is present in the set $F_{\mathcal{X}'}$. The clients in such locations require zero cost for getting connected to $F_{\mathcal{X}'}$. Thus, the total cost of connecting the clients in $S(e)$ to the set $F_{\mathcal{X}'}$ is at most: $\sum_{X \in S(e)} d(X, F_{\mathcal{X}'}) = d(X^*, F_{\mathcal{X}'}) \leq 1$.

Thus, every set of clients in $\mathcal{S}$ requires unit cost for connecting to $F_{\mathcal{X}'}$. So the solution $F_{\mathcal{X}'}$ to the Robust k-Median instance indeed has objective one. □

Lemma 6 (No-Instance). *If every labeling weakly satisfies at most a $2^{-\gamma r}$ fraction of the edges in the HGLC instance (G, π, r), for some constant γ then every solution to the Robust k-Median instance $(V, \mathcal{S}, d)$ has objective at least r.*

Proof. Fix any solution $F \subseteq V$ to the Robust k-Median instance $(V, \mathcal{S}, d)$, and let $\mathcal{X}'_F \subseteq \mathcal{X}$ denote the corresponding solution to the MCSP instance $(U, \mathcal{X}, t)$. By Lemma 4 there is some element $e \in U$ with congestion at least r under $\mathcal{X}'_F$. In other words, there are at least r sets $X_1, \ldots, X_r \in \mathcal{X}'_F$ that contain the element e. The locations corresponding to these sets are not picked by the solution F. Furthermore, the way the locations have been embedded on a line ensures that the distance between any location and its nearest neighbor is at least one. Hence, we have $d(X_i, F) \geq 1$ for all $i \in [r]$. Summing over these distances, the total cost of connecting the clients in $S(e)$ to F is at least $\sum_{i \in [r]} d(X_i, F) \geq r$. Thus, the solution F to the Robust k-Median instance has objective at least r. □

Finally, applying Lemmas 5, 6, and an argument similar to the proof of Thm. 5, we get the following result.

Theorem 6. *The Robust k-Median problem $(V, \mathcal{S}, d)$ is $\Omega(\log m / \log\log m)$ hard to approximate even on line metrics, where $m = |\mathcal{S}|$, unless $\mathsf{NP} \subseteq \cap_{\delta > 0} \mathsf{DTIME}(2^{n^\delta})$.*

5 Conclusion and Future Work

We show a logarithmic lower bound for Robust k-median on the uniform and line metrics. However, the empirical results suggest that real-world instances are much easier, so it is interesting to see if realistic assumptions can be added to the problem in order to obtain constant approximation. For instance, one may assume that the diameter of each set S_i is small compared to the real diameter. This captures the "locality" of communities. Our hardness results do not apply in this case. Also, one may attack the problem from parameterized complexity's angle: Can we obtain an $O(1)$ approximation algorithm in time $g(k)\,\mathrm{poly}(n)$?

References

1. Anthony, B.M., Goyal, V., Gupta, A., Nagarajan, V.: A plant location guide for the unsure: Approximation algorithms for min-max location problems. Math. Oper. Res. 35(1), 79–101 (2010) (Also in SODA 2008)
2. Arora, S.: Polynomial time approximation schemes for euclidean traveling salesman and other geometric problems. J. ACM 45(5), 753–782 (1998)
3. Arora, S., Lund, C., Motwani, R., Sudan, M., Szegedy, M.: Proof verification and the hardness of approximation problems. J. ACM 45(3), 501–555 (1998)
4. Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., Pandit, V.: Local search heuristics for k-median and facility location problems. SIAM J. Comput. 33(3), 544–562 (2004)
5. Bansal, N., Khandekar, R., Könemann, J., Nagarajan, V., Peis, B.: On generalizations of network design problems with degree bounds. Math. Program. 141(1-2), 479–506 (2013)
6. Charikar, M., Guha, S.: Improved combinatorial algorithms for the facility location and k-median problems. In: FOCS, pp. 378–388. IEEE Computer Society (1999)
7. Charikar, M., Guha, S., Tardos, É., Shmoys, D.B.: A constant-factor approximation algorithm for the k-median problem. J. Comput. Syst. Sci. 65(1), 129–149 (2002)
8. Feige, U.: A threshold of ln n for approximating set cover. J. ACM 45(4), 634–652 (1998)
9. Hagerup, T., Rüb, C.: A guided tour of Chernoff bounds. Information Processing Letters 33(6), 305–308 (1990), `http://www.sciencedirect.com/science/article/pii/002001909090214I`
10. Kolliopoulos, S.G., Rao, S.: A nearly linear-time approximation scheme for the euclidean k-median problem. In: Nešetřil, J. (ed.) ESA 1999. LNCS, vol. 1643, pp. 378–389. Springer, Heidelberg (1999)
11. Li, S., Svensson, O.: Approximating k-median via pseudo-approximation. In: Boneh, D., Roughgarden, T., Feigenbaum, J. (eds.) STOC, pp. 901–910. ACM (2013)
12. Lin, J.H., Vitter, J.S.: Approximation algorithms for geometric median problems. Inf. Process. Lett. 44(5), 245–249 (1992)
13. Lund, C., Yannakakis, M.: On the hardness of approximating minimization problems. J. ACM 41(5), 960–981 (1994)
14. Raz, R.: A parallel repetition theorem. SIAM J. Comput. 27(3), 763–803 (1998)