Approximation Algorithms for Clustering Problems with Lower Bounds and Outliers*

Sara Ahmadian[1] and Chaitanya Swamy[†2]

1 **Combinatorics and Optimization, Univ. Waterloo, Waterloo, ON, Canada**
sahmadian@math.uwaterloo.ca
2 **Combinatorics and Optimization, Univ. Waterloo, Waterloo, ON, Canada**
sahmadian@math.uwaterloo.ca

Abstract

We consider clustering problems with *non-uniform lower bounds and outliers*, and obtain the *first approximation guarantees* for these problems. We have a set $\mathcal{F}$ of facilities with lower bounds $\{L_i\}_{i\in\mathcal{F}}$ and a set $\mathcal{D}$ of clients located in a common metric space $\{c(i,j)\}_{i,j\in\mathcal{F}\cup\mathcal{D}}$, and bounds k, m. A feasible solution is a pair $\big(S\subseteq\mathcal{F},\sigma:\mathcal{D}\mapsto S\cup\{\mathsf{out}\}\big)$, where σ specifies the client assignments, such that $|S|\leq k$, $|\sigma^{-1}(i)|\geq L_i$ for all $i\in S$, and $|\sigma^{-1}(\mathsf{out})|\leq m$. In the *lower-bounded min-sum-of-radii with outliers* (LBkSRO) problem, the objective is to minimize $\sum_{i\in S}\max_{j\in\sigma^{-1}(i)}c(i,j)$, and in the *lower-bounded k-supplier with outliers* (LBkSupO) problem, the objective is to minimize $\max_{i\in S}\max_{j\in\sigma^{-1}(i)}c(i,j)$.

We obtain an approximation factor of 12.365 for LBkSRO, which improves to 3.83 for the non-outlier version (i.e., $m=0$). These also constitute the *first* approximation bounds for the min-sum-of-radii objective when we consider lower bounds and outliers *separately*. We apply the primal-dual method to the relaxation where we Lagrangify the $|S|\leq k$ constraint. The chief technical contribution and novelty of our algorithm is that, departing from the standard paradigm used for such constrained problems, we obtain an $O(1)$-approximation *despite the fact that we do not obtain a Lagrangian-multiplier-preserving algorithm for the Lagrangian relaxation.* We believe that our ideas have broader applicability to other clustering problems with outliers as well.

We obtain approximation factors of 5 and 3 respectively for LBkSupO and its non-outlier version. These are the *first* approximation results for k-supplier with *non-uniform* lower bounds.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, G.1.6 Optimization, G.2 Discrete Mathematics

Keywords and phrases Approximation algorithms, facililty-location problems, primal-dual method, Lagrangian relaxation, k-center problems, minimizing sum of radii

Digital Object Identifier 10.4230/LIPIcs.ICALP.2016.69

1 Introduction

Clustering is an ubiquitous problem arising in applications in various fields such as data mining, machine learning, image processing, and bioinformatics. Many of these problems involve finding a set S of at most k "cluster centers", and an assignment σ mapping an underlying set $\mathcal{D}$ of data points located in some metric space $\{c(i,j)\}$ to S, to minimize

* A full version of the paper is available at http://arxiv.org/abs/1608.01700.

† This work was supported in part by the second author's NSERC grant 327620-09 and NSERC Discovery Accelerator Supplement Award.

43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016).
Editors: Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi;
Article No. 69; pp. 69:1–69:15
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

some objective function; examples include the *k-center* (minimize $\max_{j\in\mathcal{D}} c(\sigma(j), j)$) [20, 21], *k-median* (minimize $\sum_{j\in\mathcal{D}} c(\sigma(j), j)$) [9, 22, 25, 6], and *min-sum-of-radii* (minimize $\sum_{i\in S} \max_{j:\sigma(j)=i} c(i, j)$) [15, 11] problems. Viewed from this perspective, clustering problems can often be viewed as *facility-location* problems, wherein an underlying set of clients that require service need to be assigned to facilities that provide service in a cost-effective fashion. Both clustering and facility-location problems have been extensively studied in the Computer Science and Operations Research literature; see, e.g., [27, 29] in addition to the above references.

We consider clustering problems with (non-uniform) *lower-bound requirements* on the cluster sizes, and where a bounded number of points may be designated as *outliers* and left unclustered. One motivation for considering lower bounds comes from an *anonymity* consideration. In order to achieve data privacy, [28] proposed an anonymization problem where we seek to perturb (in a specific way) some of (the attributes of) the data points and then cluster them so that every cluster has at least L identical perturbed data points, thus making it difficult to identify the original data from the clustering. As noted in [2, 1], this anonymization problem can be abstracted as a lower-bounded clustering problem where the clustering objective captures the cost of perturbing data. Another motivation comes from a facility-location perspective, where (as in the case of *lower-bounded facility location*), the lower bounds model that it is infeasible or unprofitable to use services unless they satisfy a certain minimum demand (see, e.g., [26]). Allowing outliers enables one to handle a common woe in clustering problems, namely that data points that are quite dissimilar from any other data point can often disproportionately (and undesirably) degrade the quality of *any* clustering of the *entire* data set; instead, the outlier-version allows one to designate such data points as outliers and focus on the data points of interest.

Formally, adopting the facility-location terminology, our setup is as follows. We have a set $\mathcal{F}$ of facilities with lower bounds $\{L_i\}_{i\in\mathcal{F}}$ and a set $\mathcal{D}$ of clients located in a common metric space $\{c(i, j)\}_{i,j\in\mathcal{F}\cup\mathcal{D}}$, and bounds k, m. A feasible solution chooses a set $S \subseteq \mathcal{F}$ of at most k facilities, and assigns each client j to a facility $\sigma(j) \in S$, or designates j as an outlier by setting $\sigma(j) = \mathsf{out}$ so that $|\sigma^{-1}(i)| \geq L_i$ for all $i \in S$, and $|\sigma^{-1}(\mathsf{out})| \leq m$. We consider two clustering objectives: minimize $\sum_{i\in S} \max_{j:\sigma(j)=i} c(i, j)$, which yields the *lower-bounded min-sum-of-radii with outliers* (LBkSRO) problem, and minimize $\max_{i\in S} \max_{j:\sigma(j)=i} c(i, j)$, which yields the *lower-bounded k-supplier with outliers* (LBkSupO) problem. We refer to the non-outlier versions of the above problems (i.e., where $m = 0$) as LBkSR and LBkSup respectively.

Our contributions. We obtain the *first* results for clustering problems with *non-uniform lower bounds and outliers*. We develop various techniques for tackling these problems using which we obtain *constant-factor approximation guarantees* for LBkSRO and LBkSupO. Note that we need to ensure that none of the *hard* constraints involved here – at most k clusters, non-uniform lower bounds, and at most m outliers – are violated, which is somewhat challenging.

We obtain an approximation factor of 12.365 for LBkSRO (Theorem 7, Section 2.2), which improves to 3.83 for the non-outlier version LBkSR (Theorem 6, Section 2.1). These also constitute the *first* approximation results for the min-sum-of-radii objective when we consider: (a) lower bounds (even uniform bounds) but no outliers (LBkSR); and (b) outliers but no lower bounds. Previously, an $O(1)$-approximation was known only in the setting where there are *no lower bounds and no outliers* (i.e., $L_i = 0$ for all i, $m = 0$) [11].

For the k-supplier objective (Section 3), we obtain an approximation factor of 5 for LBkSupO (Theorem 16), and 3 for LBkSup (Theorem 15). These are the *first* approximation results for the k-supplier problem with non-uniform lower bounds. Previously, [1] obtained approximation factors of 4 and 2 respectively for LBkSupO and LBkSup for the special case of *uniform* lower bounds and when $\mathcal{F} = \mathcal{D}$ (often called the k-center version). Complementing our approximation bounds, we prove a factor-3 hardness of approximation for LBkSup (Theorem 17), which shows that our approximation factor of 3 is optimal for LBkSup.

Our techniques. Our main technical contribution is an $O(1)$-approximation algorithm for LBkSRO (Section 2.2). Whereas for the non-outlier version LBkSR (Section 2.1), one can follow an approach similar to that in [11] for the min-sum-of-radii problem without lower bounds or outliers, the presence of outliers creates substantial difficulties whose resolution requires various novel ingredients. As in [11], we view LBkSRO as a *k-ball-selection* (k-BS) problem of picking k suitable balls (see Section 2) and consider its LP-relaxation (P_2). Let OPT denote its optimal value. Following the Jain-Vazirani (JV) template for k-median [22], we move to the version where we may pick any number of balls but incur a fixed cost of z for each ball we pick. The dual LP (D_2) has α_j dual variables for the clients, which "pay" for (i, r) pairs (where (i, r) denotes the ball $\{j \in \mathcal{D} : c(i, j) \leq r\}$). For LB$k$SR (where $m = 0$), as observed in [11], it is easy to adapt the JV primal-dual algorithm for facility location to handle this fixed-cost version of k-BS: we raise the α_js of uncovered clients until all clients are covered by some fully-paid (i, r) pair (see PDAlg). This yields a so-called *Lagrangian-multiplier-preserving* (LMP) 3-approximation algorithm: if F is the primal solution constructed, then $3\sum_j \alpha_j$ can pay for $cost(F) + 3|F|z$; hence, by varying z, one can find two solutions F_1, F_2 for nearby values of z, and combine them to extract a low-cost k-BS-solution.

The presence of outliers in LBkSRO significantly complicates things. The natural adaptation of the primal-dual algorithm is to now stop when at least $|\mathcal{D}| - m$ clients are covered by fully-paid (i, r) pairs. But now, the dual objective involves a $-m \cdot \gamma$ term, where $\gamma = \max_j \alpha_j$, which potentially cancels the dual contribution of (some) clients that pay for the last fully-paid (i, r) pair, say f. Consequently, we *do not obtain an LMP-approximation*: if F is the primal solution we construct, we can only say that (roughly) $3(\sum_j \alpha_j - m \cdot \gamma)$ pays for $cost(F \setminus f) + 3|F \setminus f|z$ (see Theorem 8 (ii)). In particular, this means that *even if the primal-dual algorithm returns a solution with k pairs, its cost need not be bounded*, an artifact that never arises in LBkSR (or k-median). This in turn means that by combining the two solutions F_1, F_2 found for $z_1, z_2 \approx z_1$, we only obtain a solution of cost $O(OPT + z_1)$ (see Theorem 10).

Dealing with the case where $z_1 = \Omega(OPT)$ is technically the most involved portion of our algorithm (Section 2.2.2). We argue that in this case the solutions F_1, F_2 (may be assumed to) have a very specific structure: $|F_1| = k + 1$, and every F_2-ball intersects at most one F_1-ball, and vice versa. We utilize this structure to show that either we can find a good solution in a suitable neighborhood of F_1 and F_2, or F_2 itself must be a good solution.

We remark that the above difficulties (i.e., the inability to pay for the last "facility" and the ensuing complications) also arise in the *k-median problem with outliers*. We believe that our ideas also have implications for this problem and should yield a much-improved approximation ratio for this problem. (The current guarantee is a large (unspecified) constant [12].)

For the k-supplier problem, LBkSupO, we leverage the notion of skeletons and pre-skeletons defined by [14] in the context of *capacitated k-supplier with outliers*, wherein facilities have capacities instead of lower bounds limiting the number of clients that can be assigned to them. Roughly speaking, a skeleton $F \subseteq \mathcal{F}$ ensures there is a low-cost solution (F, σ). A pre-skeleton satisfies some of the properties of a skeleton. We show that if F is a pre-skeleton, then either F is a skeleton or $F \cup \{i\}$ is a pre-skeleton for some facility i. This allows one to find a sequence of facility-sets such that at least one of them is a skeleton. For a given set F, one can check if F admits a low-cost assignment σ, so this yields an $O(1)$-approximation algorithm.

Related work. There is a vast literature on clustering and facility-location (FL) problems (see, e.g., [27, 29]); we limit ourselves to work that is relevant to LBkSRO and LBkSupO.

The only prior work on clustering problems to incorporate both lower bounds *and* outliers is by Aggarwal et al. [1]. They obtain approximation ratios of 4 and 2 respectively for LBkSupO and LBkSup with *uniform* lower bounds, which they consider as a means of achieving anonymity. They also consider an alternate *cellular clustering* (CellC) objective and devise an $O(1)$-approximation algorithm for lower-bounded CellC with uniform lower bounds, and mention that this can be extended to an $O(1)$-approximation for lower-bounded CellC with outliers.

More work has been directed towards clustering problems involving outliers *or* lower bounds (but not both). Charikar et al. [10] consider (among other problems) the outlier-versions of the uncapacitated FL, k-supplier and k-median problems. They devise constant-factor approximations for the first two problems, and a bicriteria approximation for the k-median problem with outliers. They also proved a factor-3 approximation hardness result for k-supplier with outliers. This nicely complements our factor-3 hardness result for k-supplier with lower bounds but no outliers. Chen [12] obtained the first true approximation for k-median with outliers via a sophisticated combination of the primal-dual algorithm for k-median and local search that yields a large (unspecified) $O(1)$-approximation. Cygan and Kociumaka [14] consider the *capacitated k-supplier with outliers* problem, and devise a 25-approximation algorithm. We leverage some of their ideas in developing our algorithm for LBkSupO.

Lower-bounded clustering and FL problems remain largely unexplored and ill-understood. Besides LBkSup (which has also been studied in Euclidean spaces [16]) another such FL problem that has been studied is *lower-bounded facility location* (LBFL) [23, 19], wherein we seek to open facilities (which have lower bounds) and assign each client j to an open facility $\sigma(j)$ so as to minimize $\sum_{j \in \mathcal{D}} c(\sigma(j), j)$. Svitkina [30] obtained the first true approximation for LBFL, achieving an $O(1)$-approximation; the $O(1)$-factor was subsequently improved by [3]. Both results apply to LBFL with uniform lower bounds, and can be adapted to yield $O(1)$-approximations to the k-median variant (where we may open at most k facilities).

Doddi et al. [15] introduced the min-sum-of-diameters objective, which is closely related to the min-sum-of-radii objective (the former is at most twice the latter). Charikar and Panigrahi [11] devised the first (and current-best) $O(1)$-approximation algorithms for these problems, obtaining approximation ratios of 3.53 and 7.06 for the radii and diameter problems respectively. Various other results are known for specific metric spaces and when $\mathcal{F} = \mathcal{D}$, such as Euclidean spaces [18, 7] and metrics with bounded aspect ratios [17, 5].

The k-supplier and k-center (i.e., k-supplier with $\mathcal{F} = \mathcal{D}$) objectives have a rich history of study. Hochbaum and Shmoys [20, 21] obtained optimal approximation ratios of 3 and 2 for these problems respectively. Capacitated versions of k-center and k-supplier have also been studied: [24] devised a 6-approximation for uniform capacities, [13] obtained the first $O(1)$-approximation for non-uniform capacities, and this $O(1)$-factor was improved to 9 in [4].

Finally, our algorithm for LBkSRO leverages the template based on Lagrangian relaxation and the primal-dual method to emerge from the work of [22, 8] for the k-median problem.

2 Minimizing sum of radii with lower bounds and outliers

Recall that in the *lower-bounded min-sum-of-radii with outliers* (LBkSRO) problem, we have a facility-set $\mathcal{F}$ and client-set $\mathcal{D}$ located in a metric space $\{c(i,j)\}_{i,j\in\mathcal{F}\cup\mathcal{D}}$, lower bounds $\{L_i\}_{i\in\mathcal{F}}$, and bounds k and m. A feasible solution is a pair $\big(S\subseteq\mathcal{F},\sigma:\mathcal{D}\mapsto S\cup\{\mathsf{out}\}\big)$, where $\sigma(j)\in S$ indicates that j is assigned to facility $\sigma(j)$, and $\sigma(j)=\mathsf{out}$ designates j as an outlier, such that $|S|\le k$, $|\sigma^{-1}(i)|\ge L_i$ for all $i\in S$, and $|\sigma^{-1}(\mathsf{out})|\le m$. The cost $\mathit{cost}(S,\sigma)$ of such a solution is $\sum_{i\in S} r_i$, where $r_i:=\max_{j\in\sigma^{-1}(i)} c(i,j)$ denotes the *radius* of facility i; the goal is to find a solution of minimum cost. We use LBkSR to denote the non-outlier version where $m=0$.

It will be convenient to consider a relaxation of LBkSRO that we call the *k-ball-selection* (k-BS) problem, which focuses on selecting at most k balls centered at facilities of minimum total radius. More precisely, let $B(i,r):=\{j\in\mathcal{D}:c(i,j)\le r\}$ denote the ball of clients centered at i with radius r. Let $c_{\max}=\max_{i\in\mathcal{F},j\in\mathcal{D}} c(i,j)$. Let $\mathcal{L}_i:=\{(i,r):|B(i,r)|\ge L_i\}$, and $\mathcal{L}:=\bigcup_{i\in\mathcal{F}}\mathcal{L}_i$. The goal in k-BS is to find a set $F\subseteq\mathcal{L}$ with $|F|\le k$ and $\big|\mathcal{D}\setminus\bigcup_{(i,r)\in F}B(i,r)\big|\le m$ so that $\mathit{cost}(F):=\sum_{(i,r)\in F} r$ is minimized. (When formulating the LP-relaxation of the k-BS-problem, we equivalently view $\mathcal{L}$ as containing only pairs of the form $(i,c(i,j))$ for some client j, which makes $\mathcal{L}$ finite.) It is easy to see that any LBkSRO-solution yields a k-BS-solution of no greater cost. The key advantage of working with k-BS is that we do not explicitly consider the lower bounds (they are folded into the $\mathcal{L}_i$s) and we do not require the balls $B(i,r)$ for $(i,r)\in F$ to be disjoint. While a k-BS-solution F need not directly translate to a feasible LBkSRO-solution, one can show that it does yield a feasible LBkSRO-solution of cost at most $2\cdot\mathit{cost}(F)$. We prove a stronger version of this statement in Lemma 1. In the following two sections, we utilize this relaxation to devise the *first* constant-factor approximation algorithms for for LBkSR and LBkSRO. To our knowledge, our algorithm is also the first $O(1)$-approximation algorithm for the outlier version of the min-sum-of-radii problem *without* lower bounds.

We consider an LP-relaxation for the k-BS-problem, and to round a fractional k-BS-solution to a good integral solution, we need to preclude radii that are much larger than those used by an optimal solution. We therefore "guess" the t facilities in the optimal solution with the largest radii, and their radii, where $t\ge 1$ is some constant. That is, we enumerate over all $O\big((|\mathcal{F}|+|\mathcal{D}|)^{2t}\big)$ choices $F^O=\{(i_1,r_1),\dots,(i_t,r_t)\}$ of t (i,r) pairs from $\mathcal{L}$. For each such selection, we set $\mathcal{D}'=\mathcal{D}\setminus\bigcup_{(i,r)\in F^O}B(i,r)$, $\mathcal{L}'=\{(i,r)\in\mathcal{L}:r\le\min_{p=1,\dots,t} r_p\}$ and $k'=k-|F^O|$, and run our k-BS-algorithm on the modified k-BS-instance $(\mathcal{F},\mathcal{D}',c,\mathcal{L}',k',m)$ to obtain a k-BS-solution F. We translate $F\cup F^O$ to an LBkSRO-solution, and return the best of these solutions. The following lemma, and the procedure described therein, is repeatedly used to bound the cost of translating $F\cup F^O$ to a feasible LBkSRO-solution. We call pairs $(i,r),(i',r')\in\mathcal{F}\times\mathbb{R}_{\ge 0}$ *non-intersecting*, if $c(i,i')>r+r'$, and *intersecting* otherwise. Note that $B(i,r)\cap B(i',r')=\emptyset$ if (i,r) and (i',r') are non-intersecting. For a set $P\subseteq\mathcal{F}\times\mathbb{R}_{\ge 0}$ of pairs, define $\mu(P):=\{i\in\mathcal{F}:\exists r \text{ s.t. } (i,r)\in P\}$.

▶ **Lemma 1.** *Let $F^O\subseteq\mathcal{L}$, and $\mathcal{D}',\mathcal{L}',k'$ be as defined above. Let $F\subseteq\mathcal{L}$ be a k-BS-solution for the k-BS-instance $(\mathcal{F},\mathcal{D}',c,\mathcal{L}',k',m)$. Suppose for each $i\in\mu(F)$, we have a radius $r'_i\le\max_{r:(i,r)\in F} r$ such that the pairs in $U:=\bigcup_{i\in\mu(F)}(i,r'_i)$ are non-intersecting and $U\subseteq\mathcal{L}'$. Then there exists a feasible LBkSRO-solution (S,σ) with $\mathit{cost}(S,\sigma)\le\mathit{cost}(F)+\sum_{(i,r)\in F^O}2r$.*

2.1 Approximation algorithm for LBkSR

We now present our algorithm for the non-outlier version, LBkSR, which introduces many of the ideas underlying our algorithm for LBkSRO (Section 2.2). Let O^* be the cost of an optimal solution to the given LBkSR instance. For each selection $(i_1, r_1), \ldots, (i_t, r_t)$ of t pairs, we do the following. We set $\mathcal{D}' = \mathcal{D} \setminus \bigcup_{p=1}^{t} B(i_p, r_p)$, $\mathcal{L}' = \{(i,r) \in \mathcal{L} : r \leq R^* := \min_{p=1,\ldots,t} r_p\}$, $k' = k - t$, and consider the k-BS-problem of picking a min-cost set of at most k' pairs from $\mathcal{L}'$ whose corresponding balls cover $\mathcal{D}'$ (but our algorithm k-BSAlg will return pairs from $\mathcal{L}$). Consider the following natural LP-relaxation (P$_1$) of this problem, and its dual (D$_1$).

$$\begin{aligned} \min \quad & \sum_{(i,r)\in\mathcal{L}'} r \cdot y_{i,r} & \text{(P}_1\text{)} \\ \text{s.t.} \quad & \sum_{(i,r)\in\mathcal{L}':j\in B(i,r)} y_{i,r} \geq 1 \qquad \forall j \in \mathcal{D}' \\ & \sum_{(i,r)\in\mathcal{L}'} y_{i,r} \leq k' & (1) \\ & y \geq 0. \end{aligned}$$

$$\begin{aligned} \max \quad & \sum_{j\in\mathcal{D}'} \alpha_j - k' \cdot z & \text{(D}_1\text{)} \\ \text{s.t.} \quad & \sum_{j\in B(i,r)\cap\mathcal{D}'} \alpha_j - z \leq r \qquad \forall (i,r) \in \mathcal{L}' & (2) \\ & \alpha, z \geq 0. \end{aligned}$$

Let OPT denote the common optimal value of (P$_1$) and (D$_1$). As in the JV-algorithm for k-median, we Lagrangify constraint (1) and consider the unconstrained problem where we do not bound the number of pairs we may pick, but we incur a fixed cost z for each pair (i,r) that we pick (in addition to r). It is easy to adapt the JV primal-dual algorithm for facility location [22] to devise a simple *Lagrangian-multiplier-preserving* (LMP) 3-approximation algorithm for this problem (see PDAlg and Theorem 3). We use this LMP algorithm within a binary-search procedure for z to obtain two solutions F_1 and F_2 with $|F_2| \leq k' < |F_1|$, and show that these can be "combined" to extract a k-BS-solution F of cost at most $3.83 \cdot OPT + O(R^*)$. This combination step is more involved than in k-median. The main idea here is to use the F_2 solution as a guide to merge some F_1-pairs. We cluster the F_1 pairs around the F_2-pairs and setup a *covering-knapsack problem* whose solution determines for each F_2-pair (i,r), whether to "merge" the F_1-pairs clustered around (i,r) or select all these F_1-pairs (see step B2). Finally, we add back the pairs $(i_1, r_1), \ldots (i_t, r_t)$ selected earlier and apply Lemma 1 to obtain an LBkSR-solution. As required by Lemma 1, to aid in this translation, our k-BS-algorithm returns, along with F, a suitable radius $\mathsf{rad}(i)$ for every facility $i \in \mu(F)$. This yields a $(3.83 + \epsilon)$-approximation algorithm (Theorem 6).

While our approach is similar to the one in [11] for the min-sum-of-radii problem *without* lower bounds (although our combination step is notably simpler), an important distinction that arises is the following. In the absence of lower bounds, the ball-selection problem k-BS is *equivalent* to the min-sum-of-radii problem, but (as noted earlier) this is no longer the case when we have lower bounds since in k-BS we do not insist that the balls we pick be disjoint. Moving from overlapping balls in a k-BS-solution to an LBkSR-solution incurs, in general, a factor-2 blowup in the cost, but we avoid this blowup by exploiting the structure of the k-BS-solution obtained and carefully merging in the pairs $(i_1, r_1), \ldots, (i_t, r_t)$ (see Lemma 1). It is interesting that our approximation factor is quite close to the approximation factor (of 3.53) achieved in [11] for the min-sum-of-radii problem without lower bounds.

We now describe our algorithm in detail and analyze it. We describe a slightly simpler $(6.183 + \epsilon)$-approximation algorithm below (Theorem 2). We sketch the ideas behind the improved approximation ratio at the end of this section and defer the details to the full version.

▶ Algorithm 1.
Input: An LBkSR-instance $\mathcal{I} = \big(\mathcal{F}, \mathcal{D}, \{L_i\}, \{c(i,j)\}, k\big)$, parameter $\epsilon > 0$.
Output: A feasible solution (S, σ).

A1. Let $t = \min\big\{k, \lceil \frac{1}{\epsilon} \rceil\big\}$. For each set $F^O \subseteq \mathcal{L}$ with $|F^O| \leq t$, do the following.

A1.1. Set $\mathcal{D}' = \mathcal{D} \setminus \bigcup_{(i,r)\in F^O} B(i,r)$, $\mathcal{L}' = \{(i',r') \in \mathcal{L} : r \leq R^* = \min_{(i,r)\in F^O} r\}$, $k' = k - |F^O|$.

A1.2. If (P$_1$) is infeasible, then reject this guess and move to the next set F^O. If $\mathcal{D}' \neq \emptyset$, run k-BSAlg$(\mathcal{D}', \mathcal{L}', k', \epsilon)$ to obtain $\big(F, \{\mathsf{rad}(i)\}_{i\in F}\big)$; else set $(F, \mathsf{rad}) = (\emptyset, \emptyset)$.

A1.3. Apply the procedure in Lemma 1 taking $r'_i = \mathsf{rad}(i)$ for all $i \in \mu(F)$ to obtain (S, σ).

A1. Among all the solutions (S, σ) found in step A1, return the one with smallest cost.

▶ Algorithm **k-BSAlg$(\mathcal{D}', \mathcal{L}', k', \epsilon)$**.
Output: $F \subseteq \mathcal{L}$ with $|F| \leq k'$, a radius $\mathsf{rad}(i)$ for all $i \in \mu(F)$.

B1. Binary search for z.

B1.1. Set $z_1 = 0$ and $z_2 = 2k'c_{\max}$. For $p = 1, 2$, let $(F_p, \{\mathsf{rad}_p(i)\}, \alpha^p) \leftarrow$ PDAlg$(\mathcal{D}', \mathcal{L}', z_p)$, and let $k_p = |F_p|$. If $k_1 \leq k'$, stop and return $\big(F_1, \{\mathsf{rad}_1(i)\}\big)$. We prove in Theorem 3 that $k_2 \leq k'$; if $k_2 = k'$, stop and return $\big(F_2, \{\mathsf{rad}_2(i)\}\big)$.

B1.2. Repeat the following until $z_2 - z_1 \leq \delta_z = \frac{\epsilon OPT}{3n}$, where $n = |\mathcal{F}| + |\mathcal{D}|$. Set $z = \frac{z_1+z_2}{2}$. Let $(F, \{\mathsf{rad}(i)\}, \alpha) \leftarrow$ PDAlg$(\mathcal{D}', \mathcal{L}', z)$. If $|F| = k'$, stop and return $\big(F, \{\mathsf{rad}(i)\}\big)$; if $|F| > k'$, update $z_1 \leftarrow z$ and $(F_1, \mathsf{rad}_1, \alpha^1) \leftarrow (F, \mathsf{rad}, \alpha)$, else update $z_2 \leftarrow z$ and $(F_2, \mathsf{rad}_2, \alpha^2) \leftarrow (F, \mathsf{rad}, \alpha)$.

B2. Combining F_1 and F_2. Let $\pi : F_1 \mapsto F_2$ be any map such that (i', r') and $\pi(i', r')$ intersect $\forall (i', r') \in F_1$. (This exists since every $j \in \mathcal{D}'$ is covered by $B(i,r)$ for some $(i,r) \in F_2$.) Define star $\mathcal{S}_{i,r} = \pi^{-1}(i,r)$ for all $(i,r) \in F_2$ (see Fig. 1). Solve the following *covering-knapsack LP*.

$$\min \sum_{(i,r)\in F_2} \Big(x_{i,r}\big(2r + \textstyle\sum_{(i',r')\in \mathcal{S}_{i,r}} 2r'\big) + (1 - x_{i,r})\textstyle\sum_{(i',r')\in \mathcal{S}_{i,r}} r'\Big) \tag{C-P}$$

$$\text{s.t.} \quad \sum_{(i,r)\in F_2} \big(x_{i,r} + |\mathcal{S}_{i,r}|(1 - x_{i,r})\big) \leq k', \qquad 0 \leq x_{i,r} \leq 1 \quad \forall (i,r) \in F_2.$$

Let x^* be an extreme-point optimal solution to (C-P). The variable $x_{(i,r)}$ has the following interpretation. If $x^*_{i,r} = 0$, then we select all pairs in $\mathcal{S}_{i,r}$. Otherwise, if $\mathcal{S}_{i,r} \neq \emptyset$, we pick a pair in $(i', r') \in \mathcal{S}_{i,r}$, and include $(i', 2r + r' + \max_{(i'',r'')\in \mathcal{S}_{i,r}\setminus\{(i',r')\}} 2r'')$ in our solution. Notice that by expanding the radius of i' to $2r + r' + \max_{(i'',r'')\in \mathcal{S}_{i,r}\setminus\{(i',r')\}} 2r''$, we cover all the clients in $\bigcup_{(i'',r'')\in \mathcal{S}_{i,r}} B(i'', r'')$. Let F' be the resulting set of pairs.

B3. If $cost(F_2) \leq cost(F)$, return (F_2, rad_2), else return $\big(F', \{\mathsf{rad}_1(i)\}_{i\in\mu(F')}\big)$.

▶ Algorithm **PDAlg$(\mathcal{D}', \mathcal{L}', z)$**.
Output: $F \subseteq \mathcal{L}$, radius $\mathsf{rad}(i)$ for all $i \in \mu(F)$, dual solution α.

P1. Dual-ascent phase. Start with $\alpha_j = 0$ for all $j \in \mathcal{D}'$, $\mathcal{D}'$ as the set of *active clients*, and the set T of *tight pairs* initialized to $\emptyset$. We repeat the following until all clients become inactive: we raise the α_js of all active clients uniformly until constraint (2) becomes tight for some (i,r); we add (i,r) to T and mark all active clients in $B(i,r)$ as inactive.

P2. Pruning phase. Let T_I be a maximal subset of non-intersecting pairs in T picked by a greedy algorithm that scans pairs in T in non-increasing order of radius. Note that for each $i \in \mu(T_I)$, there is exactly one pair $(i,r) \in T_I$. We set $\mathsf{rad}(i) = r$, and $r_i = \max\ \{c(i,j) : j \in B(i', r'),\ (i', r') \in T,\ r' \leq r,\ (i', r') \text{ intersects } (i,r)\ ((i',r') \text{ could be } (i,r))\}$. Let $F = \{(i, r_i)\}_{i\in\mu(T_I)}$. Return F, $\{\mathsf{rad}(i)\}_{i\in\mu(T_I)}$, and α.

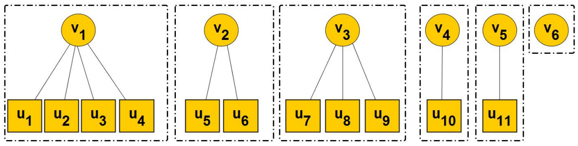

Figure 1 An example of stars formed by F_1 and F_2 where $F_1 = \{u_1, u_2, \dots, u_{11}\}$ and $F_2 = \{v_1, v_2, \dots, v_6\}$ depicted by squares and circles, respectively.

Analysis. We prove the following result.

▶ **Theorem 2.** *For any $\epsilon > 0$, Algorithm 1 returns a feasible* LBkSR*-solution of cost at most $\big(6.1821 + O(\epsilon)\big)O^*$ in time $n^{O(1/\epsilon)}$.*

We first prove that PDAlg is an LMP 3-approximation algorithm, i.e., its output (F, α) satisfies $cost(F) + 3|F|z \le 3\sum_{j\in\mathcal{D}'} \alpha_j$. (Theorem 3). Utilizing this, we analyze k-BSAlg, in particular, the output of the combination step B2, and argue that k-BSAlg returns a feasible solution of cost at most $\big(6.183 + O(\epsilon)\big) \cdot OPT + O(R^*)$ (Theorem 5). For the right choice of F^O, combining this with Lemma 1 yields Theorem 2.

▶ **Theorem 3.** *Suppose* PDAlg$(\mathcal{D}', \mathcal{L}', z)$ *returns* $(F, \{\mathsf{rad}(i)\}, \alpha)$. *Then*

(i) *the balls corresponding to F cover $\mathcal{D}'$;*

(ii) $cost(F) + 3|F|z \le 3\sum_{j\in\mathcal{D}'} \alpha_j \le 3(OPT + k'z)$*;*

(iii) $\big\{(i, \mathsf{rad}(i))\big\}_{i\in\mu(F)} \subseteq \mathcal{L}'$*, is a set of non-intersecting pairs,* $\mathsf{rad}(i) \le r_i \le 3R^*$ $\forall i \in \mu(F)$*;*

(iv) *if $|F| \ge k'$ then $cost(F) \le 3 \cdot OPT$; if $|F| > k'$, then $z \le OPT$. (Hence, $k_2 \le k'$ in step B1.1.)*

Let $\big(F, \{\mathsf{rad}(i)\}\big) = k$-BSAlg$(\mathcal{D}', \mathcal{L}', k', \epsilon)$. If k-BSAlg terminates in step B1, then $cost(F) \le 3 \cdot OPT$ due to part (ii) of Theorem 3, so assume otherwise. Let $a, b \ge 0$ be such that $ak_1 + bk_2 = k'$, $a + b = 1$. Let $C_1 = cost(F_1)$ and $C_2 = cost(F_2)$. Recall that $(F_1, \mathsf{rad}_1, \alpha^1)$ and $(F_2, \mathsf{rad}_2, \alpha^2)$ are the outputs of PDAlg for z_1 and z_2 respectively.

▶ **Claim 4.** *We have $aC_1 + bC_2 \le (3 + \epsilon)OPT$.*

▶ **Theorem 5.** k-BSAlg$(\mathcal{D}', \mathcal{L}', k', \epsilon)$ *returns a feasible solution* $\big(F, \{\mathsf{rad}(i)\}\big)$ *with* $cost(F) \le \big(6.183 + O(\epsilon)\big) \cdot OPT + O(R^*)$ *where* $\big\{(i, \mathsf{rad}(i))\big\}_{i\in\mu(F)} \subseteq \mathcal{L}'$ *is a set of non-intersecting pairs.*

Proof. The radii $\{\mathsf{rad}(i)\}_{i\in\mu(F)}$ are simply radii obtained from some execution of PDAlg, so $\big\{(i, \mathsf{rad}(i))\big\}_{i\in\mu(F)} \subseteq \mathcal{L}'$ and comprises non-intersecting pairs. If k-BSAlg terminates in step B1, we have a better bound on $cost(F)$. If not, and we return F_2, the cost incurred is C_2.

Otherwise, we return the solution F' found in step B2. Since (C-P) has only one constraint in addition to the bound constraints $0 \le x_{i,r} \le 1$, the extreme-point optimal solution x^* has at most one fractional component, and if it has a fractional component, then $\sum_{(i,r)\in F_2}\big(x^*_{i,r} + |\mathcal{S}_{i,r}|(1 - x^*_{i,r})\big) = k'$. For any $(i, r) \in F_2$ with $x^*_{i,r} \in \{0, 1\}$, the number of pairs we include is exactly $x^*_{i,r} + |\mathcal{S}_{i,r}|(1 - x^*_{i,r})$, and the total cost of these pairs is at most the contribution to the objective function of (C-P) from the $x^*_{i,r}$ and $(1 - x^*_{i,r})$ terms. If x^* has a fractional component $(i', r') \in F_2$, then $x^*_{i',r'} + |\mathcal{S}_{i',r'}|(1 - x^*_{i',r'})$ is a *positive* integer. Since we include at most one pair for (i', r'), this implies that $|F'| \le k'$. The cost of the pair we include is at most $15R^*$, since all $(i, r) \in F_1 \cup F_2$ satisfy $r \le 3R^*$. Therefore, $cost(F') \le OPT_{\text{C-P}} + 15R^*$. Also, $OPT_{\text{C-P}} \le 2bC_2 + (2b + a)C_1 = 2bC_2 + (1 + b)C_1$, since setting $x_{i,r} = b$ for all $(i, r) \in F_2$ yields a feasible solution to (C-P) of this cost.

So when we terminate in step B3, we return a solution F with $cost(F) \leq \min\{C_2, 2bC_2 + (1+b)C_1 + 15R^*\}$. We show that $\min\{C_2, 2bC_2 + (1+b)C_1\} \leq 2.0607(aC_1 + bC_2)$ for all $a, b \geq 0$ with $a + b = 1$. Combining this with Claim 4 yields the bound in the theorem. ◀

Proof. Proof of Theorem 2 It suffices to show that when the selection $F^O = \{(i_1, r_1), \ldots (i_t, r_t)\}$ in step A1 corresponds to the t facilities in an optimal solution with largest radii, we obtain the desired approximation bound. In this case, we have $R^* \leq \frac{O^*}{t} \leq \epsilon O^*$ and $OPT \leq O^* - \sum_{p=1}^{t} r_p$. Combining Theorem 5 and Lemma 1 then yields the theorem. ◀

Improved approximation ratio. The improved approximation ratio comes from a better way of combining F_1 and F_2 in step B2. We observe that the dual solutions α^1 and α^2 are component-wise close to each other (we can control the closeness by controlling δ_z). Thus, we may essentially assume that if $T_{1,I}$, $T_{2,I}$ denote the tight pairs yielding F_1, F_2 respectively, then every pair in $T_{1,I}$ intersects some pair in $T_{2,I}$, because we can augment $T_{2,I}$ to include non-intersecting pairs of $T_{1,I}$. This yields dividends when we combine solutions as in step B2, because we can now ensure that if $\pi(i', r') = (i, r)$, then the pairs of $T_{2,I}$ and $T_{1,I}$ yielding (i, r) and (i', r') respectively intersect, which yields an improved bound on $c_{i,i'}$. This yields an improved approximation of 3.83 for the combination step, and hence for the entire algorithm.

▶ **Theorem 6.** *For any $\epsilon > 0$, our algorithm returns a feasible* LBkSR*-solution of cost at most* $(3.83 + O(\epsilon))O^*$ *in time* $n^{O(1/\epsilon)}$.

2.2 Approximation algorithm for LBkSRO

We now build upon the ideas in Section 2.1 to devise an $O(1)$-approximation algorithm for the outlier version LBkSR. The high-level approach is similar to the one in Section 2.1. We again "guess" the t (i, r) pairs F^O corresponding to the facilities with largest radii in an optimal solution, and consider the modified k-BS-instance $(\mathcal{D}', \mathcal{L}', k', m)$ (where $\mathcal{D}', \mathcal{L}', k'$ are defined as before). If the LP-relaxation below, (P$_2$), for the k-BS-problem is infeasible, we move on to the next guess. Otherwise, we design a primal-dual algorithm for the Lagrangian relaxation of the k-BS-problem where we are allowed to pick any number of pairs from $\mathcal{L}'$ (leaving at most m uncovered clients) incurring a fixed cost of z for each pair picked, utilize this to obtain two solutions F_1 and F_2, and combine these to extract a low-cost solution. However, the presence of outliers introduces various difficulties both in the primal-dual algorithm and in the combination step. Consider the following LP-relaxation of the k-BS-problem and its dual.

$$\begin{aligned}
\min \quad & \sum_{(i,r)\in\mathcal{L}'} r \cdot y_{i,r} && (\text{P}_2) \\
\text{s.t.} \quad & \sum_{(i,r)\in\mathcal{L}': j\in B(i,r)} y_{i,r} + w_j \geq 1 && \forall j \in \mathcal{D}' \\
& \sum_{(i,r)\in\mathcal{L}'} y_{i,r} \leq k', \quad \sum_{j\in\mathcal{D}'} w_j \leq m \\
& y, w \geq 0.
\end{aligned}
\qquad
\begin{aligned}
\max \quad & \sum_{j\in\mathcal{D}'} \alpha_j - k' \cdot z - m \cdot \gamma && (\text{D}_2) \\
\text{s.t.} \quad & \sum_{j\in B(i,r)\cap\mathcal{D}'} \alpha_j - z \leq r && \forall (i,r) \in \mathcal{L}' \\
& \alpha_j \leq \gamma && \forall j \in \mathcal{D}' \\
& \alpha, z, \gamma \geq 0.
\end{aligned}
\qquad (3)$$

Let OPT denote the optimal value of (P$_2$). The natural modification of the earlier primal-dual algorithm PDAlg is to now stop the dual-ascent process when the number of active clients is at most m and set $\gamma = \max_{j\in\mathcal{D}'} \alpha_j$. This introduces the significant complication that one may not be able to pay for the $r+z$-cost of non-intersecting tight pairs selected in the pruning phase by the dual objective value $\sum_{j\in\mathcal{D}'}\alpha_j - m\cdot\gamma$, since clients with $\alpha_j=\gamma$ may be needed to pay for the $r+z$-cost of the last tight pair $f=(i_f,r_f)$ but their contribution gets canceled by the $-m\cdot\gamma$ term. This issue affects us in various guises. First, we no longer obtain an LMP-approximation for the unconstrained problem since we have to account for the $(r+z)$-cost of f separately. Second, unlike Claim 4, given solutions F_1 and F_2 obtained via binary search for $z_1, z_2 \approx z_1$ respectively with $|F_2|\le k'\le |F_1|$, we now only obtain a fractional k-BS-solution of cost $O(OPT+z_1)$. While one can modify the covering-knapsack-LP based procedure in step B2 of k-BSAlg to combine F_1, F_2, this only yields a good solution when $z_1=O(OPT)$. The chief technical difficulty is that z_1 may however be much larger than OPT. Overcoming this obstacle requires various novel ideas and is the key technical contribution of our algorithm. We design a second combination procedure that is guaranteed to return a good solution when $z_1=\Omega(OPT)$. This requires establishing certain structural properties for F_1 and F_2, using which we argue that one can find a good solution in the neighborhood of F_1 and F_2.

We now detail the changes to the primal-dual algorithm and k-BSAlg in Section 2.1, and analyze them to prove the following theorem.

▶ **Theorem 7.** *There exists a $\big(12.365+O(\epsilon)\big)$-approximation algorithm for* LBkSRO *that runs in time $n^{O(1/\epsilon)}$ for any $\epsilon>0$.*

Modified primal-dual algorithm PDAlg$^{\mathsf{o}}(\mathcal{D}',\mathcal{L}',z)$. This is quite similar to PDAlg (and we again return pairs from $\mathcal{L}$). We stop the dual-ascent process when there are at most m active clients. We set $\gamma=\max_{j\in\mathcal{D}'}\alpha_j$. Let $f=(i_f,r_f)$ be the last tight pair added to the tight-pair set T, and $B_f=B(i_f,r_f)$. We sometimes abuse notation and use (i,r) to also denote the singleton set $\{(i,r)\}$. For a set P of (i,r) pairs, define $\mathsf{uncov}(P):=\mathcal{D}'\setminus\bigcup_{(i,r)\in P}B(i,r)$. Note that $|\mathsf{uncov}(T\setminus f)|>m\ge|\mathsf{uncov}(T)|$. Let Out be a set of m clients such that $\mathsf{uncov}(T)\subseteq Out\subseteq\mathsf{uncov}(T\setminus f)$. Note that $\alpha_j=\gamma$ for all $j\in Out$.

The pruning phase is similar to before, but we only use f if necessary. Let T_I be a maximal subset of non-intersecting pairs picked by greedily scanning pairs in $T\setminus f$ in non-increasing order of radius. For $i\in\mu(T_I)$, set $\mathsf{rad}(i)$ to be the unique r such that $(i,r)\in T_I$, and let r_i be the smallest radius ρ such that $B(i,\rho)\supseteq B(i',r')$ for every $(i',r')\in T\setminus f$ such that $r'\le\mathsf{rad}(i)$ and (i',r') intersects $(i,\mathsf{rad}(i))$. Let $F'=\{(i,r_i)\}_{i\in\mu(T_I)}$. If $\mathsf{uncov}(F')\le m$, set $F=F'$. If $\mathsf{uncov}(F')>m$ and $\exists i\in\mu(F')$ such that $c(i,i_f)\le 2R^*$, then increase r_i so that $B(i,r_i)\supseteq B_f$ and let F be this updated F'. Otherwise, set $F=F\cup f$ and $r_{i_f}=\mathsf{rad}(i_f)=r_f$. We return $(F,f,Out,\{rad(i)\}_{i\in\mu(F)},\alpha,\gamma)$.

▶ **Theorem 8.** *Let $(F,f,Out,\{\mathsf{rad}(i)\},\alpha,\gamma)=\mathsf{PDAlg}^{\mathsf{o}}(\mathcal{D}',\mathcal{L}',z)$. Then:*
(i) $\mathsf{uncov}(F)\le m$*;*
(ii) $cost(F\setminus f)+3|F\setminus f|z-3R^*\le 3(\sum_{j\in\mathcal{D}'}\alpha_j-m\gamma)\le 3(OPT+k'z)$*;*
(iii) $\big\{(i,\mathsf{rad}(i))\big\}_{i\in\mu(F)}\subseteq\mathcal{L}'$*, is a set of non-intersecting pairs,* $\mathsf{rad}(i)\le r_i\le 3R^*\ \forall i\in\mu(F)$*;*
(iv) *if $|F\setminus f|\ge k'$ then $cost(F)\le 3\cdot OPT+4R^*$, and if $|F\setminus f|>k'$ then $z\le OPT$.*

Modified algorithm k-BSAlg$^{\mathsf{o}}(\mathcal{D}',\mathcal{L}',k',\epsilon)$. We again use binary search to find solutions F_1, F_2 and extract a low-cost solution from these. The only changes to step B1 are as follows. We start with $z_1=0$ and $z_2=2nk'c_{\max}$; for this z_2, one can argue PDAlg$^{\mathsf{o}}$ returns at most k' pairs. We stop when $z_2-z_1\le\delta_z:=\frac{\epsilon OPT}{3n2^n}$. We *do not stop* even if PDAlg$^{\mathsf{o}}$ returns a solution $(F,\dots)$ with $|F|=k'$ for some $z=\frac{z_1+z_2}{2}$, since *Theorem 8 is not strong enough to bound $cost(F)$ even when this happens!* If $|F|>k'$, we update $z_1\leftarrow z$ and the F_1-solution;

otherwise, we update $z_2 \leftarrow z$ and the F_2-solution. Thus, we maintain that $k_1 = |F_1| > k'$, and $k_2 = |F_2| \le k'$.

The main change is in the way solutions F_1, F_2 are combined. We adapt step B2 to handle outliers (procedure $\mathcal{A}$ in Section 2.2.1), but the key extra ingredient is that we devise an alternate combination procedure $\mathcal{B}$ (Section 2.2.2) that returns a low-cost solution when $z_1 = \Omega(OPT)$. We return the better of the solutions output by the two procedures. Combining Theorem 9 with Lemma 1 (for the right selection of t (i,r) pairs) yields Theorem 7.

▶ **Theorem 9.** *k-$\mathsf{BSAlg}^{\mathsf{o}}(\mathcal{D}', \mathcal{L}', k', \epsilon)$ returns a solution (F, rad) with $cost(F) \le \big(12.365 + O(\epsilon)\big) \cdot OPT + O(R^*)$ where $\big\{(i, \mathsf{rad}(i))\big\}_{i\in\mu(F)} \subseteq \mathcal{L}'$ comprises non-intersecting pairs.*

2.2.1 Combination subroutine $\mathcal{A}\Big((F_1, \mathsf{rad}_1), (F_2, \mathsf{rad}_2)\Big)$

As in step B2, we cluster the F_1-pairs around F_2-pairs in stars. However, unlike before, some $(i', r') \in F_1$ may remain *unclustered* and and we may not pick (i', r') or some pair close to it. Since we do not cover all clients covered by F_1, we need to cover a suitable number of clients from $\mathsf{uncov}(F_1)$. We again setup an LP to obtain a suitable collection of pairs, which is now a 2-dimensional covering knapsack LP, and use the structure of an extreme-point optimal solution to extract from it a good collection of pairs.

▶ **Theorem 10.** *We can obtain a solution $\big(F, \{\mathsf{rad}(i)\}_{i\in\mu(F)}\big)$ to the k-BS-problem with $cost(F) \le \big(6.1821 + O(\epsilon)\big)(OPT + z_1) + O(R^*)$ where $\big\{(i, \mathsf{rad}(i))\big\}_{i\in\mu(F)} \subseteq \mathcal{L}'$ is a set of non-intersecting pairs.*

2.2.2 Subroutine $\mathcal{B}\Big((F_1, f_1, Out_1, \mathsf{rad}_1, \alpha^1, \gamma^1), (F_2, f_2, Out_2, \mathsf{rad}_2, \alpha^2, \gamma^2)\Big)$

Subroutine $\mathcal{A}$ in the previous section yields a low-cost solution only if $z_1 = O(OPT)$. We complement subroutine $\mathcal{A}$ by now describing a procedure that returns a good solution when z_1 is large. We assume in this section that $z_1 > (1 + \epsilon)OPT$. Then $|F_1 \setminus f_1| \le k'$ (otherwise $z \le OPT$ by part (iv) of Theorem 8), so $|F_1 \setminus f_1| \le k' < |F_1|$, which means that $k_1 = k' + 1$ and $f_1 \in F_1$. Hence, $\alpha^1_j = \gamma^1$ for all $j \in B_{f_1} \cap \mathcal{D}'$. We utilize the following *continuity lemma*, which is essentially Lemma 6.6 in [11]; we include a proof in the full version of the paper.

▶ **Lemma 11.** *Let $(F_p, \ldots, \alpha^p, \gamma^p) = \mathsf{PDAlg}^{\mathsf{o}}(\mathcal{D}', \mathcal{L}', z_p)$ for $p = 1, 2$, where $0 \le z_2 - z_1 \le \delta_z$. Then, $\|\alpha^1_j - \alpha^2_j\|_\infty \le 2^n\delta_z$ and $|\gamma^1 - \gamma^2| \le 2^n\delta_z$. Thus, if (3) is tight for some $(i, r) \in \mathcal{L}'$ in one execution, then $\sum_{j\in B(i,r)\cap\mathcal{D}'} \alpha^p_j \ge r + z - 2^n\delta_z$ for $p = 1, 2$.*

First, we take care of some simple cases. If there exists $(i, r) \in F_1 \setminus f_1$ such that $\big|\mathsf{uncov}\big(F_1 \setminus \{f_1, (i, r)\} \cup (i, r + 12R^*)\big)\big| \le m$, then set $F = F_1 \setminus \{f_1, (i, r)\} \cup (i, r + 12R^*)$. We have $cost(F) = cost(F_1 \setminus f_1) + 12R^* \le 3 \cdot OPT + 15R^*$ (by part (ii) of Theorem 8). If there exist pairs $(i, r), (i', r') \in F_1$ such that $c(i, i') \le 12R^*$, take r'' to be the minimum $\rho \ge r$ such that $B(i', r') \subseteq B(i, \rho)$ and set $F = F_1 \setminus \{(i, r), (i', r')\} \cup (i, r'')$. We have $cost(F) \le cost(F_1 \setminus f_1) + 13R^* \le 3 \cdot OPT + 16R^*$. In both cases, we return $\big(F, \{\mathsf{rad}_1(i)\}_{i\in\mu(F)}\big)$.

So we assume in the sequel that neither of the above apply. In particular, all pairs in F_1 are well-separated. Let $AT = \{(i, r) \in \mathcal{L}' : \sum_{j\in B(i,r)\cap\mathcal{D}'} \alpha^1_j \ge r + z_1 - 2^n\delta_z\}$ and $AD = \{j \in \mathcal{D}' : \alpha^1_j \ge \gamma^1 - 2^n\delta_z\}$. By Lemma 11, AT includes the tight pairs of $\mathsf{PDAlg}^{\mathsf{o}}(\mathcal{D}', \mathcal{L}', z_p)$ for both $p = 1, 2$, and $Out_1 \cup Out_2 \subseteq AD$. Since the tight pairs T_2 used for building solution F_2 are almost tight in $(\alpha^1, \gamma^1, z_1)$, we swap them in and swap out pairs from F_1 one by one while maintaining a feasible solution. Either at some point, we will be able to remove f, which will give us a solution of size k', or we will obtain a bound on $cost(F_2)$. The following lemma is our main tool for bounding the cost of the solution returned.

▶ **Lemma 12.** *Let $F \subseteq \mathcal{L}'$, and $T_F = \big\{(i, r'_i)\big\}_{i\in\mu(F)}$ where $r'_i \le r$ for each $(i,r) \in F$. Suppose $T_F \subseteq AT$ and consists of non-intersecting pairs. If $|F| \ge k'$ and $|AD \setminus \bigcup_{(i,r)\in F} B(i,r))| \ge m$ then $cost(T_F) \le (1+\epsilon)OPT$. Moreover, if $|F| > k'$ then $z_1 \le (1+\epsilon)OPT$.*

Define a mapping $\psi : F_2 \to F_1 \setminus f_1$ as follows. Note that any $(i,r) \in F_2$ may intersect with at most one F_1-pair: if it intersects $(i',r'), (i'',r'') \in F_1$, then we have $c(i',i'') \le 12R^*$. First, for each $(i,r) \in F_2$ that intersects with some $(i',r') \in F_1$, we set $\psi(i,r) = (i',r')$. Let $M \subseteq F_2$ be the F_2-pairs mapped by ψ this way. For every $(i,r) \in F_2 \setminus M$, we arbitrarily match (i,r) with a *distinct* $(i',r') \in F_1 \setminus \psi(M)$. We claim that ψ is in fact a one-one function.

▶ **Lemma 13.** *Every $(i,r) \in F_1 \setminus f_1$ intersects with at most one F_2-pair.*

Let F'_2 be the pairs $(i,r) \in F_2$ such that if $(i',r') = \psi(i,r)$, then $r' < r$. Let $P = F'_2 \cap M$ and $Q = F'_2 \setminus M$. For every $(i',r') \in \psi(Q)$ and $j \in B(i',r')$, we have $j \in \mathsf{uncov}(F_2) \subseteq AD$ (else (i',r') would lie in $\psi(M)$). Starting with $F = F_1 \setminus f_1$, we iterate over $(i,r) \in F'_2$ and do the following. Let $(i',r') = \psi(i,r)$. If $(i,r) \in P$, we update $F \leftarrow F \setminus (i',r') \cup (i, r+2r')$ (so $B(i,r+2r') \supseteq B(i',r')$), else we update $F \leftarrow F \setminus (i',r') \cup (i,r)$. Let $T_F = \{(i, \mathsf{rad}_1(i))\}_{(i,r)\in F\cap F_1} \cup \{(i, \mathsf{rad}_2(i))\}_{(i,r)\in F\setminus F_1}$. Note that $|F| = k'$ and $\mathsf{uncov}(F) \subseteq AD$ at all times. Also, since (i,r) intersects only (i',r'), which we remove when (i,r) is added, we maintain that T_F is a collection of non-intersecting pairs and a subset of $AT \subseteq \mathcal{L}'$. This process continues until $|\mathsf{uncov}(F)| \le m$, or when all pairs of F'_2 are swapped in. In the former case, we argue that $cost(F)$ is small and return $\big(F, \{\mathsf{rad}_1(i)\}_{(i,r)\in F\cap F_1} \cup \{\mathsf{rad}_2(i)\}_{(i,r)\in F\setminus F_1}\big)$. In the latter case, we show that $cost(F'_2)$, and hence $cost(F_2)$ is small, and return (F_2, rad_2).

▶ **Lemma 14.**

(i) *If the algorithm stops with $|\mathsf{uncov}(F)| \le m$, $cost(F) \le (9+3\epsilon)OPT + 18R^*$.*
(ii) *If case (i) does not apply, then $cost(F_2) \le (3+3\epsilon)OPT + 9R^*$.*
(iii) *The pairs corresponding to the radii returned are non-intersecting, and a subset of $\mathcal{L}'$.*

3 Minimizing the maximum radius with lower bounds and outliers

The *lower-bounded k-supplier with outliers* (LBkSupO) problem is the min max-radius version of LBkSRO. The input and the set of feasible solutions are the same as in LBkSRO: the input is an instance $\mathcal{I} = \big(\mathcal{F}, \mathcal{D}, \{c(i,j)\}, \{L_i\}, k, m\big)$, and a feasible solution is $\big(S \subseteq \mathcal{F}, \sigma : \mathcal{D} \mapsto S \cup \{\mathsf{out}\}\big)$ with $|S| \le k$, $|\sigma^{-1}(i)| \ge L_i$ for all $i \in S$, and $|\sigma^{-1}(\mathsf{out})| \le m$. The cost of (S,σ) is now $\max_{i\in S} \max_{j\in\sigma^{-1}(i)} c(i,j)$. The case $m = 0$ is called the *lower-bounded k-supplier* (LBkSup) problem, and the setting where $\mathcal{D} = \mathcal{F}$ is often called the *k-center* version.

Let τ^* denote the optimal value; note that there are only polynomially many choices for τ^*. As is common in the study of min-max problems, we reduce the problem to a "graphical" instance, where given some value τ, we try to find a solution of cost $O(\tau)$ or deduce that $\tau^* > \tau$. We construct a bipartite unweighted graph $G_\tau = \big(V_\tau = \mathcal{D} \cup \mathcal{F}_\tau, E_\tau\big)$, where $\mathcal{F}_\tau = \{i \in \mathcal{F} : |B(i,\tau)| \ge L_i\}$, and $E_\tau = \{ij : c(i,j) \le \tau, i \in \mathcal{F}_\tau, j \in \mathcal{D}\}$. Let $dist_\tau(i,j)$ denote the shortest-path distance in G_τ between i and j, so $c(i,j) \le dist_\tau(i,j) \cdot \tau$. We say that an assignment $\sigma : \mathcal{D} \mapsto \mathcal{F}_\tau \cup \{\mathsf{out}\}$ is a *distance-α assignment* if $dist_\tau(j,\sigma(j)) \le \alpha$ for every client j with $\sigma(j) \ne \mathsf{out}$. We call such an assignment feasible, if it yields a feasible LBkSupO-solution, and we say that G_τ is feasible if it admits a feasible distance-1 assignment. It is not hard to see that given $F \subseteq \mathcal{F}_\tau$, the problem of finding a feasible distance-α-assignment $\sigma : \mathcal{D} \mapsto F \cup \{\mathsf{out}\}$ in G_τ (if one exists) can be solved by creating a network-flow instance with lower bounds and capacities.

Observe that an optimal solution yields a feasible distance-1 assignment in G_{τ^*}. We devise an algorithm that for every τ, either finds a feasible distance-α assignment in G_τ for some constant α, or detects that G_τ is not feasible. This yields an α-approximation algorithm since the smallest τ for which the algorithm returns a feasible LBkSupO-solution must be at most τ^*. We obtain Theorems 15 and 16 via this template, and complement these via a hardness result (Theorem 17) showing that our approximation factor for LBkSup is tight.

▶ **Theorem 15.** *There is a* 3*-approximation algorithm for* LBkSup*.*

▶ **Theorem 16.** *There is a* 5*-approximation algorithm for* LBkSupO*.*

▶ **Theorem 17.** *It is NP-hard to approximate* LBkSup *within a factor better than* 3*.*

Finding a distance-5 assignment for LBkSupO. The goal is to find a set $F \subseteq \mathcal{F}_\tau$ of at most k centers that are close to the centers in $F^* \subseteq \mathcal{F}_\tau$ for some feasible distance-1 assignment $\sigma^* : \mathcal{D} \mapsto F^* \cup \{\mathsf{out}\}$ in G_τ. If centers in F do not share a neighbor in G_τ, then clients in $N(i)$ can be assigned to i for each $i \in F$ to satisfy the lower bounds.

▶ **Definition 18** ([14]). Given the graph G_τ, a set $F \subseteq \mathcal{F}$ is called a *skeleton* if it satisfies the following properties.

(a) (*Separation property*) For $i, i' \in F$, $i \neq i'$, we have $dist_\tau(i, i') \geq 6$;

(b) There exists a feasible distance-1 assignment $\sigma^* : \mathcal{D} \mapsto F^* \cup \{\mathsf{out}\}$ in G_τ such that

- (*Covering property*) For all $i^* \in F^*$, $dist_\tau(i^*, F) \leq 4$, where $dist_\tau(i^*, F) = \min_{i \in F} dist_\tau(i^*, i)$.
- (*Injection property*) There exists $f : F \mapsto F^*$ such that $dist_\tau(i, f(i)) \leq 2$ for all $i \in F$.

If F satisfies the separation and injection properties, it is called a *pre-skeleton*.

▶ **Lemma 19.** *Let F be a pre-skeleton in G_τ. Define $U = \{i \in \mathcal{F}_\tau : dist_\tau(i, F) \geq 6\}$ and let $i = \arg\max_{i' \in U} |N(i')|$. Then, either F is a skeleton, or $F \cup \{i\}$ is a pre-skeleton.*

Suppose $F \subseteq \mathcal{F}_\tau$ is a skeleton and satisfies the properties with respect to a feasible distance-1 assignment (F^*, σ^*). The separation property ensures that the neighbor sets of any two locations $i, i' \in F$ are disjoint. The covering property ensures that F^* is at distance at most 4 from F, so there are at least $|\mathcal{D}| - m$ clients at distance at most 5 from F. Finally, the injection and separation properties together ensure that $|F| \leq k$. Thus, if F is a skeleton, then we can obtain a feasible distance-5 assignment $\sigma : \mathcal{D} \mapsto F \cup \{\mathsf{out}\}$.

If G_τ is feasible, then $\emptyset$ is a pre-skeleton. A skeleton can have size at most k. So using Lemma 19, we can find a sequence $\mathcal{F}'$ of at most $k+1$ subsets of $\mathcal{F}_\tau$ by starting with $\emptyset$ and repeatedly applying Lemma 19 until we either have a set of size k or the set U in Lemma 19 is empty. By Lemma 19, if G_τ is feasible then one of these sets must be a skeleton. So for each $F \in \mathcal{F}'$, we check if there exists a feasible distance-5 assignment $\sigma : \mathcal{D} \mapsto F \cup \{\mathsf{out}\}$, and if so, return (F, σ). Otherwise we return that G_τ is not feasible.

References

1 Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. Achieving anonymity via clustering. *ACM Transactions on Algorithms (TALG)*, 6(3):49, 2010.

2 Gagan Aggarwal, Tomás Feder, Krishnaram Kenthapadi, Rajeev Motwani, Rina Panigrahy, Dilys Thomas, and An Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

3 Sara Ahmadian and Chaitanya Swamy. Improved approximation guarantees for lower-bounded facility location. In *Proceedings of the 10th International Workshop on Approximation and Online Algorithms*, pages 257–271. Springer, 2012.

4 Hyung-Chan An, Aditya Bhaskara, Chandra Chekuri, Shalmoli Gupta, Vivek Madan, and Ola Svensson. Centrality of trees for capacitated k-center. *Mathematical Programming*, 154(1-2):29–53, 2015.

5 Babak Behsaz and Mohammad R Salavatipour. On minimum sum of radii and diameters clustering. *Algorithmica*, 73(1):143–165, 2015.

6 Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 737–756. SIAM, 2015.

7 Vasilis Capoyleas, Günter Rote, and Gerhard Woeginger. Geometric clusterings. *Journal of Algorithms*, 12(2):341–356, 1991.

8 Moses Charikar and Sudipto Guha. Improved combinatorial algorithms for facility location problems. *SIAM Journal on Computing*, 34(4):803–824, 2005.

9 Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the 31st annual ACM Symposium on Theory of Computing*, pages 1–10. ACM, 1999.

10 Moses Charikar, Samir Khuller, David Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 642–651. SIAM, 2001.

11 Moses Charikar and Rina Panigrahy. Clustering to minimize the sum of cluster diameters. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, pages 1–10. ACM, 2001.

12 Ke Chen. A constant factor approximation algorithm for k-median clustering with outliers. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 826–835. SIAM, 2008.

13 Marek Cygan, MohammadTaghi Hajiaghayi, and Samir Khuller. Lp rounding for k-centers with non-uniform hard capacities. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 273–282. IEEE, 2012.

14 Marek Cygan and Tomasz Kociumaka. Constant factor approximation for capacitated k-center with outliers. In *Proc. 31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014)*, volume 25 of *LIPIcs – Leibniz International Proceedings in Informatics*, pages 251–262. Schloss Dagstuhl –Leibniz-Zentrum fuer Informatik, 2014. doi:10.4230/LIPIcs.STACS.2014.251.

15 Srinivas R Doddi, Madhav V Marathe, SS Ravi, David Scot Taylor, and Peter Widmayer. Approximation algorithms for clustering to minimize the sum of diameters. In *Proceedings of the 11th Scandinavian Workshop on Algorithm Theory*, pages 237–250, 2000.

16 Alina Ene, Sariel Har-Peled, and Benjamin Raichel. Fast clustering with lower bounds: No customer too far, no shop too small. *arXiv preprint arXiv:1304.7318*, 2013.

17 Matt Gibson, Gaurav Kanade, Erik Krohn, Imran A Pirwani, and Kasturi Varadarajan. On metric clustering to minimize the sum of radii. *Algorithmica*, 57(3):484–498, 2010.

18 Matt Gibson, Gaurav Kanade, Erik Krohn, Imran A Pirwani, and Kasturi Varadarajan. On clustering to minimize the sum of radii. *SIAM Journal on Computing*, 41(1):47–60, 2012.

19 Sudipto Guha, Adam Meyerson, and Kamesh Munagala. Hierarchical placement and network design problems. In *Proceedings of 41st Annual IEEE Symposium on Foundations of Computer Science*, pages 603–612. IEEE, 2000.

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

21 Dorit S Hochbaum and David B Shmoys. A unified approach to approximation algorithms for bottleneck problems. *Journal of the ACM*, 33(3):533–550, 1986.

22 Kamal Jain and Vijay V Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *Journal of the ACM*, 48(2):274–296, 2001.

23 David R. Karger and Maria Minkoff. Building steiner trees with incomplete global knowledge. In *Proceedings of 41st Annual IEEE Symposium on Foundations of Computer Science*, pages 613–623. IEEE, 2000.

24 Samir Khuller and Yoram J Sussmann. The capacitated k-center problem. *SIAM Journal on Discrete Mathematics*, 13(3):403–418, 2000.

25 Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 901–910. ACM, 2013.

26 Andrew Lim, Fan Wang, and Zhou Xu. A transportation problem with minimum quantity commitment. *Transportation Science*, 40(1):117–129, 2006.

27 Pitu B Mirchandani and Richard L Francis. *Discrete location theory.* Wiley-Interscience, 1990.

28 Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

29 David B Shmoys. The design and analysis of approximation algorithms. *Trends in Optimization: American Mathematical Society Short Course, January 5-6, 2004, Phoeniz, Arizona*, 61:85, 2004.

30 Zoya Svitkina. Lower-bounded facility location. *ACM Transactions on Algorithms (TALG)*, 6(4):69, 2010.