Towards a Theoretical Understanding of Why Local Search Works for Clustering with Fair-Center Representation

Zhen Zhang[1,2], Junfeng Yang[1,2], Limei Liu[1,2], Xuesong Xu[1,2*], Guozhen Rong[3], Qilong Feng[4,2*]

[1]School of Advanced Interdisciplinary Studies, Hunan University of Technology and Business, China
[2]Xiangjiang Laboratory, China
[3]School of Computer and Communication Engineering, Changsha University of Science and Technology, China
[4]School of Computer Science and Engineering, Central South University, China
csuzz@foxmail.com, b12100031@hnu.edu.cn, seagullm@163.com, xuxs@hutb.edu.cn,
rongguozhen@csust.edu.cn, csufeng@mail.csu.edu.cn

Abstract

The representative k-median problem generalizes the classical clustering formulations in that it partitions the data points into several disjoint demographic groups and poses a lower-bound constraint on the number of opened facilities from each group, such that all the groups are fairly represented by the opened facilities. Due to its simplicity, the local-search heuristic that optimizes an initial solution by iteratively swapping at most a constant number of closed facilities for the same number of opened ones (denoted by the O(1)-swap heuristic) has been frequently used in the representative k-median problem. Unfortunately, despite its good performance exhibited in experiments, whether the O(1)-swap heuristic has provable approximation guarantees for the case where the number of groups is more than 2 remains an open question for a long time. As an answer to this question, we show that the O(1)-swap heuristic

(i) is guaranteed to yield a constant-factor approximation solution if the number of groups is a constant, and

(ii) has an unbounded approximation ratio otherwise.

Our main technical contribution is a new approach for theoretically analyzing local-search heuristics, which derives the approximation ratio of the O(1)-swap heuristic via linearly combining the increased clustering costs induced by a set of hierarchically organized swaps.

1 Introduction

Center-based clustering is one of the fundamental problems in the field of unsupervised learning, which aims to locate facilities (or called clustering centers) to serve a set of clients as cheaply as possible. This problem is useful in the task of data summarization, where the set of opened facilities is viewed as a summary of the data set. Despite its simplicity and popularity, algorithms for center-based clustering can yield unfair representations of the underlying groups of clients. One such example is in image searching for occupations (Kay, Matuszek, and Munson 2015). Here, the search result is a small but representative subset of the image database, in which fairly reflecting demographics (e.g., race and gender) is necessary, and classical clustering algorithms are not guaranteed to yield the desired results since they tend to minimize the clustering cost without concerning the attributes of the facilities. Motivated thus, lots of attention has been paid on *clustering with fair-center representation*, where the number of opened facilities from each demographic group is constrained to ensure fairness across demographics (Kleindessner, Awasthi, and Morgenstern 2019; Chiplunkar, Kale, and Ramamoorthy 2020; Thejaswi, Ordozgoiti, and Gionis 2021; Thejaswi et al. 2022; Angelidakis et al. 2022; Nguyen, Nguyen, and Jones 2022; Hotegni, Mahabadi, and Vakilian 2023).

Clustering with fair-center representation was recently formalized as the *representative k-median* (REP-k-MED) problem (Thejaswi, Ordozgoiti, and Gionis 2021; Thejaswi et al. 2022). As in the classical k-median problem, the goal of REP-k-MED is to open at most k facilities such that the sum of the distances from the clients to the nearest opened facilities is minimized. In REP-k-MED, however, the data points are partitioned into ℓ disjoint demographic groups and it is required that at least a given number of opened facilities are belong to each group. Formally, REP-k-MED is defined as follows.

Definition 1 (REP-k-MED) *An instance of* REP-k-MED *is specified by a metric space $(\mathcal{X}, d)$ with distance function d, a set $\mathcal{C} \subseteq \mathcal{X}$ of clients, a set $\mathcal{F} \subseteq \mathcal{X}$ of facilities, a collection $\mathbb{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_\ell\}$ of ℓ disjoint subsets of $\mathcal{F}$ satisfying $\bigcup_{t=1}^{\ell} \mathcal{G}_t = \mathcal{F}$, an integer $k \in [1, |\mathcal{F}|]$, and a vector $\vec{r} = (r[1], \ldots, r[\ell])$ of ℓ positive integers satisfying $\sum_{t=1}^{\ell} r[t] \leq k$. A feasible solution to the instance is a subset $\mathcal{S} \subseteq \mathcal{F}$ of facilities satisfying $|\mathcal{S}| \leq k$ and $|\mathcal{S} \cap \mathcal{G}_t| \geq r[t]\, \forall\, t \in \{1, \ldots, \ell\}$. The cost of such a solution is $\sum_{j \in \mathcal{C}} d(j, \mathcal{S})$, where $d(j, \mathcal{S})$ denotes the distance from j to its nearest facility in $\mathcal{S}$. The goal of* REP-k-MED *is to find a feasible solution with minimal cost.*

Thejaswi, Ordozgoiti, and Gionis (2021) experimentally showed that a multi-swap *local-search heuristic* yields high-quality solutions for REP-k-MED in real-world datasets, albeit how to analyze its approximation ratio was left as an open question. Starting with an arbitrary feasible solution, such a heuristic iteratively swaps a set of closed facilities

*Corresponding author

Algorithm 1: The p-swap heuristic for REP-k-MED

Input: An instance $\mathcal{I} = ((\mathcal{X}, d), \mathcal{C}, \mathcal{F}, \mathbb{G}, \ell, k, \vec{r})$ of REP-k-MED and an integer $p \geq 1$
Output: A locally-optimal solution $\mathcal{S}$ to $\mathcal{I}$

1: Let $\mathcal{S}$ be an arbitrary feasible solution to $\mathcal{I}$ satisfying $|\mathcal{S}| = k$.
2: **while** there exists a feasible solution $\mathcal{S}'$ to $\mathcal{I}$ satisfying $|\mathcal{S}' - \mathcal{S}| \leq p$ and $\sum_{j \in \mathcal{C}} d(j, \mathcal{S}') < \sum_{j \in \mathcal{C}} d(j, \mathcal{S})$ **do**
3: $\quad \mathcal{S} \Leftarrow \mathcal{S}'$.
4: **return** $\mathcal{S}$

for the same number of opened ones to reduce the clustering cost, until a *locally-optimal solution* that cannot be improved by any of such swaps is constructed.

As a simple and efficient way to obtain a locally-optimal solution, much work has been devoted on experimentally and theoretically understanding the effectiveness of local search for REP-k-MED (Thejaswi, Ordozgoiti, and Gionis 2021; Thejaswi et al. 2022), including studies on a special case of REP-k-MED referred to as *red-blue median* where the sum of the lower bounds on the number of opened facilities from the demographic groups equals k (Hajiaghayi, Khandekar, and Kortsarz 2010, 2012; Friggstad and Zhang 2016). It has been proved that the local-search heuristic trying to swap at most a constant number of facilities in each iteration yields constant-factor approximation for REP-k-MED if $\ell \leq 2$. However, for the harder case where $\ell > 2$, whether such a simple heuristic has provable approximation guarantees for REP-k-MED (including its special case called red-blue median) has remained an open question for a long time, see discussions in (Hajiaghayi, Khandekar, and Kortsarz 2012; Friggstad and Zhang 2016; Thejaswi, Ordozgoiti, and Gionis 2021).

The number of groups is more than 2 in most practical situations concerning REP-k-MED (e.g., where the groups encode age, race, or ethnicity), and requiring $\ell \leq 2$ means that the algorithm is quite limited in its applicability. Thus, it is necessary to consider a more general case where ℓ can be an arbitrary positive integer.

1.1 Our Results

In this paper we take a step further in analyzing the effectiveness of local search for REP-k-MED. Specifically, we analyze the p-swap heuristic described in Algorithm 1, and obtain the following guarantee.

Theorem 1 *Given an instance of* REP-k-MED *with ℓ demographic groups, the cost of each locally-optimal solution constructed by the $(\ell+1)^2$-swap heuristic is at most $4\ell + 5$ times the cost of an optimal solution to the instance.*

Theorem 1 says that the local-search heuristic using constant-size swaps has a constant-factor approximation guarantee, on the condition that ℓ is upper-bounded by a constant. For fixed ℓ, this affirmatively answers the open question about the effectiveness of local search proposed in (Hajiaghayi, Khandekar, and Kortsarz 2012; Friggstad and Zhang 2016; Thejaswi, Ordozgoiti, and Gionis 2021). We also give a lower bound on the swap size keeping the approximation ratio bounded, as described in Theorem 2.

Theorem 2 *There exists an instance of* REP-k-MED *with n clients and ℓ demographic groups satisfying $n \gg \ell$, such that each p-swap heuristic with $p < \ell$ has a locally-optimal solution whose cost is more than $n\ell^{-1} - 1$ times the cost of an optimal solution to the instance.*

Theorem 2 implies that bounding ℓ by a constant when using local-search heuristics is necessary from the theoretical point of view: It follows immediately that the local-search heuristic cannot approximate REP-k-MED within a factor less than $n\ell^{-1}-1$ using constant-size swaps when ℓ is super-constant. As a corollary of Theorem 1 and Theorem 2, we can answer the question of whether the local-search heuristic using constant-size swaps has provable approximation guarantees for REP-k-MED as follows.

(i) The local-search heuristic using constant-size swaps has the guarantee of yielding a constant-factor approximation solution when ℓ is a constant, and
(ii) has an unbounded approximation ratio when ℓ is super-constant.

1.2 Our Techniques

The analyses of local-search heuristics for clustering problems commonly construct a set of *test swaps* that close some facilities from the considered locally-optimal solution and open some facilities from an optimal one, so that the ratio between the costs of the locally-optimal and optimal solutions can be bounded using the fact that no such swap yields an improved solution. The approaches given in (Hajiaghayi, Khandekar, and Kortsarz 2010, 2012; Friggstad and Zhang 2016) partition the facilities from the locally-optimal and optimal solutions into a set of *blocks*, each of which consists of the facilities closed and opened in a test swap. These approaches associate the facilities from the locally-optimal solution with different labels depending on their demographic attributes and distances to the facilities from the optimal solution, and carefully select the members of each block according to the facility-labels, such that the changes in the cost induced by the corresponding swaps can be easily estimated and combined to yield an upper bound on the cost of the locally-optimal solution. This provides a clear way for deriving the approximation ratio of the locally-optimal solution. Unfortunately, the difficulty of constructing the blocks increases with the number of demographic groups (i.e., ℓ): Compared with the case where we only consider no more than two groups, balancing the numbers of opened and closed facilities from each group to construct valid and cost-bounded test swaps is much more challenging for the case where $\ell > 2$. This was verified in (Friggstad and Zhang 2016), where it was pointed out that getting the well structured blocks when ℓ is an arbitrary constant instead of upper-bounded by 2 is not possible.

In this paper we deal with the case where $\ell > 2$. As mentioned above, constructing the nicely structured test swaps according to the facility-labels, as done by the block-based approaches, seems unlikely in this harder case. Thus,

we no longer label the facilities and constrain the label-distributions of the test swaps. This weakens the properties guaranteed by the test swaps and makes them harder to analyze. As a remedy, we analyze the test swaps in a more refined hierarchical way, which we now briefly describe. For each to-be-clustered client j, denote by S_j and O_j the clustering costs of j induced by the locally-optimal and optimal solutions, respectively. It is shown that the changes in the cost of the locally-optimal solution induced by the test swaps can be upper-bounded by arithmetic expressions consisting of the terms of "$+S_j$", "$-S_j$", and "$+O_j$" for some clients. Our goal is to add these expressions together to get "$\alpha \sum_j O_j - \beta \sum_j S_j$" for two real numbers α and β satisfying $\alpha > \beta \geq 1$, which immediately indicates the approximation ratio of the heuristic due to the fact that the locally-optimal solution cannot be improved by the test swaps and the changes in the cost induced by these swaps are non-negative. To achieve this goal, we need to eliminate all the "$+S_j$" terms. We prove that our test swaps can be hierarchically organized into different levels such that the "$+S_j$" terms induced by each swap can be counteracted by repeatedly using the arithmetic expressions corresponding to the swaps with higher levels. These ideas lead to the proof of approximation guarantees of the local-search heuristic.

In our opinion, the method for hierarchically organizing and analyzing the test swaps and the construction of the swaps allowing the existence of such a hierarchical structure are our main technical contributions and the keys in obtaining the desired approximation guarantee.

1.3 Related Work

Clustering with fair-center representation was first studied by Kleindessner, Awasthi, and Morgenstern (2019), inspired by applications in data summarization for socioeconomic data (Moens, Uyttendaele, and Dumortier 1999; Girdhar and Dudek 2012; Kay, Matuszek, and Munson 2015). Since then this problem has been extensively studied under various objective functions, including k-center (Chiplunkar, Kale, and Ramamoorthy 2020; Angelidakis et al. 2022; Nguyen, Nguyen, and Jones 2022; Hotegni, Mahabadi, and Vakilian 2023), k-median (Thejaswi, Ordozgoiti, and Gionis 2021; Thejaswi et al. 2022; Hotegni, Mahabadi, and Vakilian 2023), and k-means (Thejaswi et al. 2022; Hotegni, Mahabadi, and Vakilian 2023). Thejaswi, Ordozgoiti, and Gionis (2021) showed that REP-k-MED can be reduced to the *matroid median* problem (Krishnaswamy et al. 2011) if the number of demographic groups is a constant, and can be approximated to a constant ratio based on the linear programming-based approaches for the latter (Li 2011; Swamy 2014; Krishnaswamy, Li, and Sandeep 2018). Hotegni, Mahabadi, and Vakilian (2023) gave a more scalable $O(1)$-approximation algorithm for the problem using small-size linear programming formulations, which can also work in a more general setting where the number of opened facilities from each demographic group is bounded by the given lower and upper bounds. Thejaswi et al. (2022) showed that combining a submodular optimization approach with the method for data reduction given in (Chen 2006) yields a $(1 + 2e^{-1} + \varepsilon)$-approximation algorithm for REP-k-MED with running time exponential in k, ℓ, and ε, where k is the upper bound on the number of opened facilities and ℓ is the number of demographic groups.

Another line of work on REP-k-MED is devoted on designing practical algorithms for the problem. Specifically, the technique of local search has been frequently used in REP-k-MED and exhibited good performance in plenty of experiments (Thejaswi, Ordozgoiti, and Gionis 2021; Thejaswi et al. 2022). Compared with the linear programming-based algorithms, the local-search heuristic is purely combinatorial and hence much easier to be implemented. Moreover, one can easily trade off the computational complexity against the solution quality when using the local-search heuristic, by changing the swap size, termination condition, and search range. Due to these reasons, much more attention has been paid on the local-search heuristics than the linear programming-based algorithms from the experimental point of view. This motivates our work in this paper where we theoretically analyze the effectiveness of the local-search heuristic.

The technique of local search plays an important role in many clustering problems (Cohen-Addad, Klein, and Mathieu 2019; Friggstad, Rezapour, and Salavatipour 2019; Cohen-Addad et al. 2022; Gupta et al. 2017; Bansal, Garg, and Gupta 2012; Zhang 2007). For example, the best approximation guarantee for the standard k-median problem was based on the local-search heuristic for almost 10 years (Arya et al. 2001). However, the analysis of the local-search heuristic for the standard k-median problem does not extend easily to REP-k-MED due to the increased difficulty in constructing feasible swap operations. Consider two disjoint demographic groups $\mathcal{G}_1$ and $\mathcal{G}_2$ as an example: After closing a facility from $\mathcal{G}_1$, to keep the cost bounded, we may need to open a facility from $\mathcal{G}_2$ to serve the clients previously served by the just-closed facility, in which case we are forced to simultaneously swap another pair of facilities to balance the number of opened facilities from each demographic group. This makes the cost of the solution after performing a test swap much more complex to analyze. In this paper we deal with this issue based on a new technique that hierarchically organizes the test swaps, which adds to the body of work exploring the power of local search.

2 The Approximation Guarantee of the $(\ell + 1)^2$-Swap Heuristic

In this section we prove Theorem 1. Generally speaking, we organize the proof as follows. In Section 2.1, we construct a set of test swaps between a locally-optimal solution given by Algorithm 1 and an optimal solution to the instance. After this, we estimate the changes in the cost induced by performing these swaps in Section 2.2. Finally, in Section 2.3, we hierarchically organize the test swaps, and show that summing the changes in the cost induced by them yields an expression linearly combining the costs of the locally-optimal and optimal solutions. Based on such an expression and the local optimality of the solution, we get the approximation ratio of the considered heuristic.

We now introduce some notations to be used through-

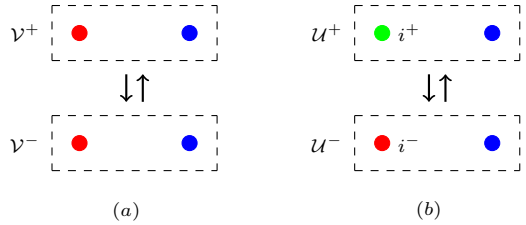

Figure 1: On the left is an example of the valid swaps, and on the right is an example of the almost-valid swaps, where facilities with the same color have the same demographic attribute. We have $g^-(\mathcal{U}) = g(i^-)$ and $g^+(\mathcal{U}) = g(i^+)$.

out this section. Let $\mathcal{I} = ((\mathcal{X}, d), \mathcal{C}, \mathcal{F}, \mathbb{G}, \ell, k, \vec{r})$ denote an instance of REP-k-MED, where $\mathbb{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_\ell\}$ and $\vec{r} = (r[1], \ldots, r[\ell])$. Define $[t] = \{1, \ldots, t\}$ for each integer $t \geq 1$. Let $\mathcal{O} \subseteq \mathcal{F}$ be an optimal solution to $\mathcal{I}$ with $|\mathcal{O}| = k$ and $\mathcal{S} \subseteq \mathcal{F}$ be a locally-optimal solution given by the $(\ell+1)^2$-swap heuristic described in Algorithm 1. We assume $\mathcal{S} \cap \mathcal{O} = \emptyset$, which is without loss of generality since one can duplicate each $i \in \mathcal{F}$ and assert that $\mathcal{S}$ contains only the copies and $\mathcal{O}$ contains only the original facilities. For each $j \in \mathcal{C}$, let o_j and s_j denote the facility nearest to j from $\mathcal{O}$ and $\mathcal{S}$ respectively, and define $S_j = d(j, s_j)$ and $O_j = d(j, o_j)$, where ties are broken arbitrarily. For each $i \in \mathcal{O}$ and $\mathcal{O}' \subseteq \mathcal{O}$, define $\mathcal{J}^*(i) = \{j \in \mathcal{C} : o_j = i\}$ and $\mathcal{J}^*(\mathcal{O}') = \bigcup_{i' \in \mathcal{O}'} \mathcal{J}^*(i')$. Similarly, define $\mathcal{J}(i) = \{j \in \mathcal{C} : s_j = i\}$ and $\mathcal{J}(\mathcal{S}') = \bigcup_{i \in \mathcal{S}'} \mathcal{J}(i)$ for each $i \in \mathcal{S}$ and $\mathcal{S}' \subseteq \mathcal{S}$. Given two real numbers t_1 and t_2, let $\Delta(t_1, t_2) = 1$ if $t_1 = t_2$, and let $\Delta(t_1, t_2) = 0$ otherwise.

For the case where $k > \sum_{t=1}^{\ell} r[t]$, the number of facilities from each demographic group opened by a feasible solution is indeterminate, and thus $\mathcal{O}$ and $\mathcal{S}$ may have different demographic distributions. This increases the difficulties in analyzing the approximation ratio of $\mathcal{S}$. As a remedy to this issue, we unify the demographic distributions of $\mathcal{O}$ and $\mathcal{S}$ by defining an additional demographic group for $2(k - \sum_{t=1}^{\ell} r[t])$ facilities from $\mathcal{O} \cup \mathcal{S}$. Specifically, let $\mathcal{S}_t$ be an arbitrary subset of $\mathcal{S} \cap \mathcal{G}_t$ satisfying $|\mathcal{S}_t| = r[t]$ and $\mathcal{O}_t$ be an arbitrary subset of $\mathcal{O} \cap \mathcal{G}_t$ with $|\mathcal{O}_t| = r[t]$ for each $t \in [\ell]$, and define $\mathcal{S}_{\ell+1} = \mathcal{S} \setminus \bigcup_{t=1}^{\ell} \mathcal{S}_t$ and $\mathcal{O}_{\ell+1} = \mathcal{O} \setminus \bigcup_{t=1}^{\ell} \mathcal{O}_t$. For each $t \in [\ell+1]$ and $i \in \mathcal{S}_t \cup \mathcal{O}_t$, we call $g(i) = t$ the demographic attribute of i.

2.1 Constructing the Test Swaps

A test swap considered in this section closes a set $\mathcal{V}^- \subseteq \mathcal{S}$ of facilities and opens the facilities from another set $\mathcal{V}^+ \subseteq \mathcal{O}$. Denote by $\mathcal{V} = (\mathcal{V}^- \mid \mathcal{V}^+)$ such a swap. We call $\mathcal{V}$ a *valid swap* if $|\mathcal{V}^-| = |\mathcal{V}^+| \neq 0$ and $|\mathcal{V}^- \cap \mathcal{S}_t| = |\mathcal{V}^+ \cap \mathcal{O}_t|$ for each $t \in [\ell+1]$, and a *non-valid swap* otherwise. It can be shown that $\mathcal{S}$ is still a feasible solution to $\mathcal{I}$ after performing a valid swap. Given a non-valid swap $\mathcal{U} = (\mathcal{U}^- \mid \mathcal{U}^+)$, we call $\mathcal{U}$ an *almost-valid swap* if there exist two facilities $i^- \in \mathcal{U}^-$ and $i^+ \in \mathcal{U}^+$ such that $(\mathcal{U}^- \setminus \{i^-\} \mid \mathcal{U}^+ \setminus \{i^+\})$ is a valid swap or $(\mathcal{U}^- \mid \mathcal{U}^+) = (\{i^-\} \mid \{i^+\})$, and let $g^-(\mathcal{U}) = g(i^-)$ and $g^+(\mathcal{U}) = g(i^+)$. Examples of the valid and almost-valid swaps are given in Figure 1.

Algorithm 2: Constructing the test swaps

$\mathbb{V} \Leftarrow \emptyset, \mathbb{U} \Leftarrow \emptyset$.
$\mathcal{S}^\dagger \Leftarrow \{i \in \mathcal{S} : \tau^{-1}(i) \neq \emptyset\}, \mathcal{O}^\dagger \Leftarrow \{\gamma(i) : i \in \mathcal{S}^\dagger\}, \mathcal{S}^\ddagger \Leftarrow \mathcal{S} \setminus \mathcal{S}^\dagger, \mathcal{O}^\ddagger \Leftarrow \mathcal{O} \setminus \mathcal{O}^\dagger$.
while $\exists \mathcal{S}' \subseteq \mathcal{S}^\dagger$ s.t. $|\mathcal{S}'| \leq (\ell+1)^2$ and $\mathcal{V} = (\mathcal{S}' \mid \bigcup_{i \in \mathcal{S}'} \{\gamma(i)\})$ is a valid swap **do** ▷ Loop-1
  $\mathbb{V} \Leftarrow \mathbb{V} \cup \{\mathcal{V}\}$.
  $\mathcal{S}^\dagger \Leftarrow \mathcal{S}^\dagger \setminus \mathcal{S}', \mathcal{O}^\dagger \Leftarrow \mathcal{O}^\dagger \setminus \bigcup_{i \in \mathcal{S}'} \{\gamma(i)\}$.
for each $i \in \mathcal{S}^\dagger$ **do**
  $\mathbb{U} \Leftarrow \mathbb{U} \cup \{(\{i\} \mid \{\gamma(i)\})\}$.
while $\exists \{\mathcal{U}_1, \mathcal{U}_2\} \subseteq \mathbb{U}$ s.t. $\mathcal{U} = (\mathcal{U}_1^- \cup \mathcal{U}_2^- \mid \mathcal{U}_1^+ \cup \mathcal{U}_2^+)$ is an almost-valid swap **do** ▷ Loop-2
  $\mathbb{U} \Leftarrow \mathbb{U} \cup \{\mathcal{U}\} \setminus \{\mathcal{U}_1, \mathcal{U}_2\}$.
while $\exists \mathbb{U}' \subseteq \mathbb{U}$, $\mathcal{Q}^- \subseteq \mathcal{S}^\ddagger$, and $\mathcal{Q}^+ \subseteq \tau^{-1}(\bigcup_{\mathcal{U} \in \mathbb{U}'} \mathcal{U}^-) \cap \mathcal{O}^\ddagger$ s.t. $|\mathbb{U}'| = |\mathcal{Q}^-| = |\mathcal{Q}^+| \leq \ell+1$ and $\mathcal{V} = (\bigcup_{\mathcal{U} \in \mathbb{U}'} \mathcal{U}^- \cup \mathcal{Q}^- \mid \bigcup_{\mathcal{U} \in \mathbb{U}'} \mathcal{U}^+ \cup \mathcal{Q}^+)$ is a valid swap **do** ▷ Loop-3
  $\mathbb{V} \Leftarrow \mathbb{V} \cup \{\mathcal{V}\}$.
  $\mathbb{U} \Leftarrow \mathbb{U} \setminus \mathbb{U}', \mathcal{S}^\ddagger \Leftarrow \mathcal{S}^\ddagger \setminus \mathcal{Q}^-, \mathcal{O}^\ddagger \Leftarrow \mathcal{O}^\ddagger \setminus \mathcal{Q}^+$.
while $\exists \mathcal{U} \in \mathbb{U}$, $i^- \in \mathcal{S}^\ddagger$, and $i^+ \in \mathcal{O}^\ddagger$ s.t. swap $\mathcal{V} = (\mathcal{U}^- \cup \{i^-\} \mid \mathcal{U}^+ \cup \{i^+\})$ is valid **do** ▷ Loop-4
  $\mathbb{V} \Leftarrow \mathbb{V} \cup \{\mathcal{V}\}$.
  $\mathbb{U} \Leftarrow \mathbb{U} \setminus \{\mathcal{U}\}, \mathcal{S}^\ddagger \Leftarrow \mathcal{S}^\ddagger \setminus \{i^-\}, \mathcal{O}^\ddagger \Leftarrow \mathcal{O}^\ddagger \setminus \{i^+\}$.
while $\exists i^- \in \mathcal{S}^\ddagger$ and $i^+ \in \mathcal{O}^\ddagger$ s.t. $\mathcal{V} = (\{i^-\} \mid \{i^+\})$ is a valid swap **do** ▷ Loop-5
  $\mathbb{V} \Leftarrow \mathbb{V} \cup \{\mathcal{V}\}$.
  $\mathcal{S}^\ddagger \Leftarrow \mathcal{S}^\ddagger \setminus \{i^-\}, \mathcal{O}^\ddagger \Leftarrow \mathcal{O}^\ddagger \setminus \{i^+\}$.
return $\mathbb{V}$

Our test swaps are constructed based on the following two functions that capture the neighbors of the facilities from $\mathcal{O} \cup \mathcal{S}$: Let $\tau(i)$ denote the facility from $\mathcal{S}$ nearest to i for each $i \in \mathcal{O}$, and let $\gamma(i')$ denote the facility from $\tau^{-1}(i')$ nearest to i' for each $i' \in \mathcal{S}$ satisfying $\tau^{-1}(i') \neq \emptyset$, where ties are broken arbitrarily. Let $\tau^{-1}(\mathcal{S}') = \bigcup_{i \in \mathcal{S}'} \tau^{-1}(i)$ for each $\mathcal{S}' \subseteq \mathcal{S}$. The procedure for constructing the test swaps is described in Algorithm 2. Note that this procedure is used only in our analysis.

To demonstrate Algorithm 2, we construct test swaps for the example given in Figure 2. We have $\mathcal{S}^\dagger = \{r_1, r_3, b_1\}$, $\mathcal{S}^\ddagger = \{r_2, b_2, b_3\}$, $\mathcal{O}^\dagger = \{r_2^*, r_3^*, b_2^*\}$, and $\mathcal{O}^\ddagger = \{r_1^*, b_1^*, b_3^*\}$ in the initialization step. In loop-1, $\mathcal{V}_1 = (\{r_1\} \mid \{\gamma(r_1)\})$ is added to $\mathbb{V}$ since it is a valid swap. In the for-loop and loop-2, a set $\mathbb{U}$ of almost-valid swaps satisfying $\bigcup_{i \in \mathcal{U}^-} \{\gamma(i)\} = \mathcal{U}^+$ for each $\mathcal{U} \in \mathbb{U}$ is constructed, and we have $\mathbb{U} = \{(\{b_1\} \mid \{\gamma(b_1)\}), (\{r_3\} \mid \{\gamma(r_3)\})\}$. Algorithm 2 combines the almost-valid swaps from $\mathbb{U}$ with some facilities from $\mathcal{S}^\ddagger \cup \mathcal{O}^\ddagger$ to construct a set of valid swaps in loop-3 and loop-4. Define $\mathcal{U}_1 = (\{b_1\} \mid \{\gamma(b_1)\})$ and $\mathcal{U}_2 = (\{r_3\} \mid \{\gamma(r_3)\})$. In loop-3, $\mathcal{U}_1$ and two facilities $r_2 \in \mathcal{S}^\ddagger$ and $b_3^* \in \tau^{-1}(\mathcal{U}_1^{-1})$ are combined into a valid swap $\mathcal{V}_2$. Similarly, in loop-4, $\mathcal{U}_2$ and two facilities $b_2 \in \mathcal{S}^\ddagger$ and $r_1^* \in \mathcal{O}^\ddagger$ are combined into a valid swap $\mathcal{V}_3$. Finally, the remained two facilities are combined into a valid swap $\mathcal{V}_4 = (\{b_3\} \mid \{b_1^*\})$

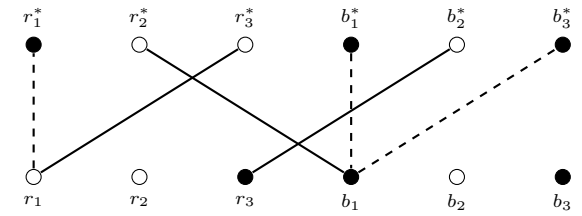

Figure 2: $\mathcal{S} = \{r_1, r_2, r_3, b_1, b_2, b_3\}$ and $\mathcal{O} = \{r_1^*, r_2^*, r_3^*, b_1^*, b_2^*, b_3^*\}$ are the locally-optimal and optimal solutions respectively, where facilities with the same color have the same demographic attribute. For each $i \in \mathcal{S}$ satisfying $\tau^{-1}(i) \neq \emptyset$ and $i^* \in \tau^{-1}(i) \backslash \{\gamma(i)\}$, we connect i and $\gamma(i)$ with a solid line, and connect i and i^* with a dashed line.

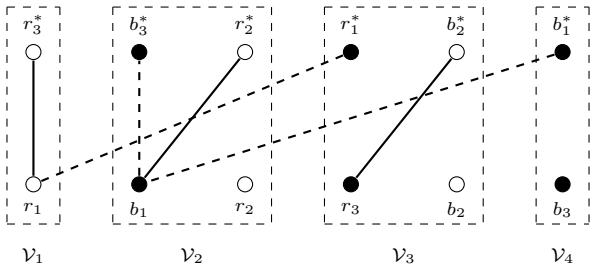

Figure 3: An example of the test swaps constructed by Algorithm 2.

in loop-5. These constructed swaps are shown in Figure 3. Intuitively, the idea of this procedure can be summarized as follows.

(i) For each $i \in \mathcal{S}$ satisfying $\tau^{-1}(i) \neq \emptyset$, Algorithm 2 assigns i and $\gamma(i)$ to the same swap, such that the increased cost induced by the swap closing i can be bounded, which we detail in Section 2.2.

(ii) Algorithm 2 constructs the swaps $\mathcal{V}$ satisfying $\mathcal{V}^+ \subseteq \tau^{-1}(\mathcal{V}^-)$ in priority, until the termination condition of loop-3 is reached and the remained facilities cannot form a swap satisfying the desired property. This provides a workable way for hierarchically organizing the swaps, as detailed in Section 2.3.

Let $\mathbb{V}$ denote the set of valid swaps constructed by Algorithm 2, and let $\mathbb{U}$ denote the set of almost-valid swaps constructed in loop-2. We define $\mathcal{S}^\dagger$, $\mathcal{O}^\dagger$, $\mathcal{S}^\ddagger$, and $\mathcal{O}^\ddagger$ as the same way as Algorithm 2, that is, let $\mathcal{S}^\dagger = \{i \in \mathcal{S} : \tau^{-1}(i) \neq \emptyset\}$, $\mathcal{O}^\dagger = \{\gamma(i) : i \in \mathcal{S}^\dagger\}$, $\mathcal{S}^\ddagger = \mathcal{S} \backslash \mathcal{S}^\dagger$, and $\mathcal{O}^\ddagger = \mathcal{O} \backslash \mathcal{O}^\dagger$. Define $\mathcal{T}^+ = \{t \in [\ell+1] : \sum_{\mathcal{U} \in \mathbb{U}} \Delta(g^+(\mathcal{U}), t) > 0\}$ and $\mathcal{T}^- = \{t \in [\ell+1] : \sum_{\mathcal{U} \in \mathbb{U}} \Delta(g^-(\mathcal{U}), t) > 0\}$ for brevity. In the following we give some useful properties of the swaps constructed by Algorithm 2.

The fact that Algorithm 2 iteratively combines two almost-valid swaps into a new one in loop-2 implies the following lemma.

Lemma 1 *$\mathcal{T}^- \cap \mathcal{T}^+ = \emptyset$.*

The following result says that each swap from $\mathbb{V}$ is of size no more than $(\ell+1)^2$, and thus performing it cannot reduce the cost of $\mathcal{S}$ due to the termination condition of the $(\ell+1)^2$-swap heuristic.

Lemma 2 *For each $\mathcal{V} \in \mathbb{V}$, we have $|\mathcal{V}^-| = |\mathcal{V}^+| \leq (\ell+1)^2$.*

The almost-valid swaps from $\mathbb{U}$ are combined with some facilities from $\mathcal{S}^\ddagger \cup \mathcal{O}^\ddagger$ into a set of valid swaps. The following is a useful structural property guaranteed by these swaps.

Lemma 3 *Considering two sets $\mathcal{Q}^- \subseteq \mathcal{S}^\ddagger$ and $\mathcal{Q}^+ \subseteq \mathcal{O}^\ddagger$ of facilities and a set $\mathbb{U}' \subseteq \mathbb{U}$ of almost-valid swaps satisfying $|\mathbb{U}'| = |\mathcal{Q}^-| = |\mathcal{Q}^+|$, if $(\bigcup_{\mathcal{U} \in \mathbb{U}'} \mathcal{U}^- \cup \mathcal{Q}^- \mid \bigcup_{\mathcal{U} \in \mathbb{U}'} \mathcal{U}^+ \cup \mathcal{Q}^+)$ is a valid swap, then it is the case that $\sum_{\mathcal{U} \in \mathbb{U}'} \Delta(g^-(\mathcal{U}), t) = \sum_{i \in \mathcal{Q}^+} \Delta(g(i), t)$ and $\sum_{\mathcal{U} \in \mathbb{U}'} \Delta(g^+(\mathcal{U}), t) = \sum_{i \in \mathcal{Q}^-} \Delta(g(i), t)$ for each $t \in [\ell+1]$.*

Finally, we are able to show that each facility from $\mathcal{S} \cup \mathcal{O}$ is involved in exactly one swap from $\mathbb{V}$. When we combine the increased costs induced by performing the swaps from $\mathbb{V}$ to derive the approximation ratio of $\mathcal{S}$, this result plays an important role.

Lemma 4 *We have $\mathcal{S} = \biguplus_{\mathcal{V} \in \mathbb{V}} \mathcal{V}^-$ and $\mathcal{O} = \biguplus_{\mathcal{V} \in \mathbb{V}} \mathcal{V}^+$.*

2.2 Estimating the Increased Costs

In this section we estimate the changes in the cost of $\mathcal{S}$ induced by performing the swaps from $\mathbb{V}$. Recall that i and $\gamma(i)$ are assigned to the same swap by Algorithm 2 for each $i \in \mathcal{S}^\dagger$. This immediately implies the following fact.

Fact 1 *$\gamma(i) \in \mathcal{V}^+$ for each $\mathcal{V} \in \mathbb{V}$ and $i \in \mathcal{V}^- \cap \mathcal{S}^\dagger$.*

Let $\mathcal{V}$ denote a test swap from $\mathbb{V}$. When performing $\mathcal{V}$ to adjust the locally-optimal solution $\mathcal{S}$, we open the facilities from $\mathcal{V}^+$ and close the facilities from $\mathcal{V}^-$, and the cost of the solution is changed to $\sum_{j \in \mathcal{C}} d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-)$. Considering a client $j \in \mathcal{C}$, Fact 1 implies that $\gamma(\tau(o_j)) \in \mathcal{V}^+$ for the case where $\tau(o_j) \in \mathcal{V}^-$. Consequently, we know that $d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-)$ can be upper-bounded by $d(j, \gamma(\tau(o_j)))$ if the swap closes $\tau(o_j)$, and $d(j, \tau(o_j))$ otherwise. The following lemma implies that such an upper bound on $d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-)$ is guaranteed to be a combination of S_j and O_j.

Lemma 5 *For each $j \in \mathcal{C}$, we have $d(j, \tau(o_j)) \leq S_j + 2O_j$ and $d(j, \gamma(\tau(o_j))) \leq 2S_j + 3O_j$.*

To estimate the cost of the solution $\mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-$, we partition $\mathcal{C}$ into several disjoint subsets and separately analyze the clustering costs of the clients from each subset, as detailed below.

(i) For each $j \in \mathcal{C} \backslash (\mathcal{J}(\mathcal{V}^-) \cup \mathcal{J}^*(\mathcal{V}^+))$, we have $s_j \in \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-$ and $d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-) \leq S_j$.

(ii) For each $j \in \mathcal{J}^*(\mathcal{V}^+)$, we have $o_j \in \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-$ and $d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-) \leq O_j$.

(iii) For each $j \in \mathcal{J}(\mathcal{V}^-) \backslash \mathcal{J}^*(\tau^{-1}(\mathcal{V}^-) \cup \mathcal{V}^+)$, the fact that $j \notin \mathcal{J}^*(\tau^{-1}(\mathcal{V}^-))$ implies that $\tau(o_j) \notin \mathcal{V}^-$, which in turn implies that $\tau(o_j) \in \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-$, and we have $d(j, \mathcal{S} \cup \mathcal{V}^+ \backslash \mathcal{V}^-) \leq d(j, \tau(o_j)) \leq S_j + 2O_j$ due to Lemma 5.

(iv) For each $j \in \mathcal{J}(\mathcal{V}^-) \cap \mathcal{J}^*(\tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+)$, the fact that $j \in \mathcal{J}^*(\tau^{-1}(\mathcal{V}^-))$ implies that $\tau(o_j) \in \mathcal{V}^-$, and we have $\gamma(\tau(o_j)) \in \mathcal{V}^+$ due to Fact 1. Consequently, it is the case that $\gamma(\tau(o_j)) \in \mathcal{S} \cup \mathcal{V}^+\backslash\mathcal{V}^-$, and Lemma 5 implies that $d(j, \mathcal{S} \cup \mathcal{V}^+\backslash\mathcal{V}^-) \leq d(j, \gamma(\tau(o_j))) \leq 2S_j + 3O_j$.

By the argument above, we know that each $\mathcal{V} \in \mathbb{V}$ satisfies

$$\begin{aligned} 0 &\leq \sum_{j\in\mathcal{C}} d(j, \mathcal{S} \cup \mathcal{V}^+\backslash\mathcal{V}^-) - \sum_{j\in\mathcal{C}} S_j \\ &\leq \sum_{j\in\mathcal{J}^*(\mathcal{V}^+)} (O_j - S_j) + \sum_{j\in\mathcal{J}(\mathcal{V}^-)\backslash\mathcal{J}^*(\tau^{-1}(\mathcal{V}^-)\cup\mathcal{V}^+)} 2O_j \\ &\quad + \sum_{j\in\mathcal{J}(\mathcal{V}^-)\cap\mathcal{J}^*(\tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+)} (3O_j + S_j), \end{aligned} \tag{1}$$

where the first step follows from the termination condition of the $(\ell+1)^2$-swap heuristic described in Algorithm 1 and Lemma 2.

2.3 Hierarchically Organizing the Test Swaps

It can be seen that inequality (1) involves "$+O_j$" and "$-S_j$" terms for the clients from $\mathcal{C}$. We want to add both sides of this inequality over $\mathcal{V} \in \mathbb{V}$ to get $O(1)\sum_{j\in\mathcal{C}} O_j - \sum_{j\in\mathcal{C}} S_j \geq 0$, which immediately indicates the desired approximation guarantee for the local-search heuristic. However, inequality (1) also involves a "$+S_j$" term that needs to be counteracted. We show that the swaps from $\mathbb{V}$ allow the existence of a hierarchical structure, which yields a feasible way to deal with this issue, as detailed in the following lemma.

Lemma 6 *We can partition $\mathbb{V}$ into h disjoint subsets $\mathbb{V}_1, \ldots, \mathbb{V}_h$, such that*

(i) $h \in [3, \ell+2]$,
(ii) $\bigcup_{\mathcal{V}\in\mathbb{V}_h} \tau^{-1}(\mathcal{V}^-) = \emptyset$, *and*
(iii) *each $t \in [h-1]$ satisfies $\bigcup_{\mathcal{V}\in\mathbb{V}_t} \tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+ \subseteq \bigcup_{\mathcal{V}\in\mathbb{V}_t^+} \mathcal{V}^+$, where $\mathbb{V}_t^+ = \bigcup_{t'=t+1}^{h} \mathbb{V}_{t'}$.*

Instead of immediately proving Lemma 6, we first show the implication of the lemma. Let $\mathbb{V}_1, \ldots, \mathbb{V}_h$ denote the h subsets of $\mathbb{V}$ constructed by Lemma 6. We have $\bigcup_{\mathcal{V}\in\mathbb{V}_t} \tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+ \subseteq \bigcup_{\mathcal{V}\in\mathbb{V}_t^+} \mathcal{V}^+$ for each $t \in [h-1]$. Combining this with the fact that given a swap $\mathcal{V} \in \mathbb{V}$, inequality (1) contains a "$-S_j$" term for each $j \in \mathcal{J}^*(\mathcal{V}^+)$ and a "$+S_j$" term for each $j \in \mathcal{J}(\mathcal{V}^-)\cap\mathcal{J}^*(\tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+)$, we know that the "$+S_j$" terms induced by the swaps from $\mathbb{V}_1$ can be counteracted via multiplying inequality (1) by a factor of 2 for each $\mathcal{V} \in \mathbb{V}_1^+$. After this, the swaps from $\mathbb{V}_2$ induce some "$+2S_j$" terms, which can be canceled via multiplying inequality (1) by factor 3 for each $\mathcal{V} \in \mathbb{V}_2^+$. By the same argument, we can multiply inequality (1) by factor t for each $t \in [h]$ and each swap from $\mathbb{V}_t$ to cancel all the "$+S_j$" terms. It is shown in Lemma 7 that this yields the desired approximation ratio of $\mathcal{S}$.

Proof (of Lemma 6) Denote by $\mathbb{V}^\dagger$ the set of swaps constructed in loop-4 of Algorithm 2. For each swap $\mathcal{V} \in \mathbb{V}^\dagger$, $\mathcal{V}$ is a combination of two facilities from $\mathcal{S}^\ddagger \cup \mathcal{O}^\ddagger$ and an almost-valid swap $\mathcal{U} \in \mathbb{U}$, and we define $g(\mathcal{V}) = g^-(\mathcal{U})$. Define $\mathbb{V}_t^\dagger = \{\mathcal{V} \in \mathbb{V}^\dagger : g(\mathcal{V}) = t\}$ for each $t \in [\ell+1]$. We construct a graph G according to the members of $\mathbb{V}^\dagger$ as follows: We construct a vertex v_t for each $t \in [\ell+1]$ satisfying $\mathbb{V}_t^\dagger \neq \emptyset$, and denote by $\mathcal{P}(G)$ the vertex set of G; for each $\{t_1, t_2\} \subseteq \mathcal{P}(G)$, if there exists a swap $\mathcal{V} \in \mathbb{V}_{t_1}^\dagger$ and a facility $i \in \tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+$ satisfying $g(i) = t_2$, then we construct an arc from v_{t_1} to v_{t_2}. The following claim gives a useful property of G.

Claim 1 *G is a directed acyclic graph with $|\mathcal{P}(G)| \leq \ell$.*

For each $\{v_1, v_2\} \subseteq \mathcal{P}(G)$, let $f(v_1, v_2)$ be the number of vertices lying in a longest path from v_1 to v_2 if G has at least one path from v_1 to v_2, and let $f(v_1, v_2) = 2$ otherwise. Denote by $\mathcal{P}_0(G) \subseteq \mathcal{P}(G)$ the set of vertices with an in-degree of 0. Let $f(v) = 2$ for each $v \in \mathcal{P}_0(G)$ and $f(v) = \max_{v'\in\mathcal{P}_0(G)} f(v', v) + 1$ for each $v \in \mathcal{P}(G)\backslash\mathcal{P}_0(G)$. We have $f(v) \in [2, |\mathcal{P}(G)| + 1]$ for each $v \in \mathcal{P}(G)$. Define $h = \max_{v\in\mathcal{P}(G)} f(v) + 1$. Claim 1 implies that

$$3 \leq h \leq |\mathcal{P}(G)| + 2 \leq \ell + 2. \tag{2}$$

Based on the graph G, we partition $\mathbb{V}$ as follows.

(i) Let $\mathbb{V}_1$ be the set of swaps constructed in loop-1 and loop-3 of Algorithm 2,
(ii) let $\mathbb{V}_h$ be the set of swaps constructed in loop-5 of Algorithm 2, and
(iii) let $\mathbb{V}_q = \bigcup_{f(v_t)=q} \mathbb{V}_t^\dagger$ for each $q \in \{2, \ldots, h-1\}$.

Using Claim 1 and the structural properties of the swaps from $\mathbb{V}$ given in Section 2.1, we obtain that such a partition of $\mathbb{V}$ has the following guarantee.

Claim 2 *Each $t \in [h-1]$ satisfies $\bigcup_{\mathcal{V}\in\mathbb{V}_t} \tau^{-1}(\mathcal{V}^-)\backslash\mathcal{V}^+ \subseteq \bigcup_{\mathcal{V}\in\mathbb{V}_t^+} \mathcal{V}^+$.*

Considering a swap $\mathcal{V} \in \mathbb{V}_h$, we know that $\mathcal{V}$ is constructed in loop-5 of Algorithm 2 and hence $\mathcal{V}^- \subseteq \mathcal{S}^\ddagger$. Moreover, it is the case that $\tau^{-1}(\mathcal{V}^-) = \emptyset$ due to the definition of $\mathcal{S}^\ddagger$. Combining this with inequality (2) and Claim 2, we complete the proof of Lemma 6. □

Denote by $\mathbb{V}_1, \ldots, \mathbb{V}_h$ the h subsets of $\mathbb{V}$ constructed by Lemma 6. Multiplying both sides of inequality (1) by a factor of t for each $t \in [h]$ and $\mathcal{V} \in \mathbb{V}_t$, and summing both sides of the inequality over $\mathcal{V} \in \mathbb{V}$, we get the following result, which says that $\mathcal{S}$ is a $(4\ell+5)$-approximation solution to $\mathcal{I}$ and hence Theorem 1 is true.

Lemma 7 $\sum_{j\in\mathcal{C}} S_j \leq (4\ell+5)\sum_{j\in\mathcal{C}} O_j$.

3 The Locality Gap of the $(\ell-1)$-Swap Heuristic

In this section we prove Theorem 2. Define $[t] = \{1, \ldots, t\}$ for each positive integer t. Motivated by a lower bound for the red-blue median problem given in (Krishnaswamy et al. 2011), we construct a bad instance of REP-k-MED, which is illustrated in Figure 4. In this instance, we are given ℓ demographic groups $\{i_1^*, i_1\}, \ldots, \{i_\ell^*, i_\ell\}$ of facilities and a

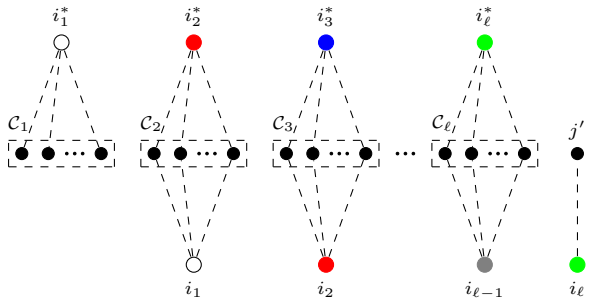

Figure 4: Illustration of a bad instance of REP-k-MED. For each $j \in \bigcup_{t\in[\ell]} \mathcal{C}_t \cup \{j'\}$ and $i \in \bigcup_{t\in[\ell]}\{i_t^*, i_t\}$, we have $d(i,j) = 0$ if i and j are connected with a dashed line and $d(i,j) = 1$ otherwise.

set $\bigcup_{t\in[\ell]} \mathcal{C}_t \cup \{j'\}$ of clients satisfying $|\mathcal{C}_t| = m$ for each $t \in [\ell]$. The constraint posed on the instance is that the number of the opened facilities is upper-bounded by ℓ, and at least one facility from each demographic group needs to be opened. As shown in Figure 4, the instance involves $\ell + 1$ locations, the distance between each pair of which is 1. Here, i_1^* and the clients from $\mathcal{C}_1$ are in the first location, j' is in the $(\ell+1)$-th location, and for each $t \in \{2, \ldots, \ell\}$, the facilities from $\{i_t^*, i_{t-1}\}$ and the clients from $\mathcal{C}_t$ are in the t-th location.

Define $\mathcal{C} = \bigcup_{t\in[\ell]} \mathcal{C}_t \cup \{j'\}$, $\mathcal{O} = \{i_t^* : t \in [\ell]\}$, and $\mathcal{S} = \{i_t : t \in [\ell]\}$. Let $n = |\mathcal{C}| = m\ell + 1$ denote the number of clients involved in the instance. It can be seen that $\mathcal{O}$ is an optimal solution to the instance, whose cost is $\sum_{j\in\mathcal{C}} d(j,O) = d(j',O) = 1$. Moreover, we have $\sum_{j\in\mathcal{C}} d(j,S) = \sum_{j\in\mathcal{C}_1} d(j,S) = m$, and thus

$$\frac{\sum_{j\in\mathcal{C}} d(j,S)}{\sum_{j\in\mathcal{C}} d(j,O)} = m > n\ell^{-1} - 1. \tag{3}$$

If we can show that swapping less than ℓ facilities between $\mathcal{S}$ and $\mathcal{O}$ cannot reduce the cost of $\mathcal{S}$, then the termination condition of Algorithm 1 implies that $\mathcal{S}$ is a locally-optimal solution for each p-swap heuristic satisfying $p < \ell$. Combining this with inequality (3), we can complete the proof of Theorem 2.

It remains to consider the changes in the cost of $\mathcal{S}$ induced by the swaps of size less than ℓ. For the sake of contradiction, assume that there exists a set $\mathcal{V}^- \subset \mathcal{S}$ and a set $\mathcal{V}^+ \subset \mathcal{O}$, such that $|\mathcal{V}^-| = |\mathcal{V}^+| < \ell$ and $\mathcal{S}\backslash\mathcal{V}^- \cup \mathcal{V}^+$ is a feasible solution satisfying $\sum_{j\in\mathcal{C}} d(j, \mathcal{S}\backslash\mathcal{V}^- \cup \mathcal{V}^+) < \sum_{j\in\mathcal{C}} d(j,\mathcal{S})$. Define $\mathcal{S}' = \mathcal{S}\backslash\mathcal{V}^- \cup \mathcal{V}^+$ for brevity. We separately consider the following two cases: (1) $i_1^* \notin \mathcal{S}'$, and (2) $i_1^* \in \mathcal{S}'$.

For case (1), we have

$$\sum_{j\in\mathcal{C}} d(j,\mathcal{S}') \geq \sum_{j\in\mathcal{C}_1} d(j,\mathcal{S}') = m = \sum_{j\in\mathcal{C}} d(j,\mathcal{S}),$$

which is a contradiction.

For case (2), the assumption that $\mathcal{S}'$ has lower cost than $\mathcal{S}$ implies that

$$\sum_{t=2}^{\ell} \sum_{j\in\mathcal{C}_t} d(j,\mathcal{S}') < \sum_{j\in\mathcal{C}} d(j,\mathcal{S}) = m,$$

which in turn implies that

$$\mathcal{S}' \cap \{i_t^*, i_{t-1}\} \neq \emptyset \tag{4}$$

for each $t \in \{2, \ldots, \ell\}$. Moreover, the fact that a feasible solution to the considered instance opens at least one facility from each demographic group implies that

$$\mathcal{S}' \cap \{i_t^*, i_t\} \neq \emptyset \tag{5}$$

for each $t \in [\ell]$, and the fact that the number of opened facilities is upper-bounded by ℓ implies that

$$|\mathcal{S}'| \leq \ell. \tag{6}$$

Combining inequality (6) with inequality (4) and the assumption that $i_1^* \in \mathcal{S}'$ yields

$$|\mathcal{S}' \cap \{i_t^*, i_{t-1}\}| = 1 \tag{7}$$

for each $t \in \{2, \ldots, \ell\}$, and combining inequality (6) with inequality (5) yields

$$|\mathcal{S}' \cap \{i_t^*, i_t\}| = 1 \tag{8}$$

for each $t \in [\ell]$. Using inequality (7), inequality (8), and the assumption that $i_1^* \in \mathcal{S}'$, we have $i_t \notin \mathcal{S}'$ for each $t \in [\ell]$, which implies that $\mathcal{S}' \cap \mathcal{S} = \emptyset$ and thus $|\mathcal{V}^-| = |\mathcal{S}| = \ell$. This contradicts the assumption that $|\mathcal{V}^-| < \ell$.

By the argument above, we know that $\mathcal{S}$ is a locally-optimal solution with approximation ratio larger than $n\ell^{-1} - 1$ for each p-swap heuristic with $p < \ell$. This implies that Theorem 2 is true.

4 Conclusions

In this paper we study the effectiveness of the local-search heuristic with constant-size swaps for the representative k-median problem. It is shown that such a heuristic yields a constant-factor approximation if the number of demographic groups, denoted by ℓ, is a constant, and has an unbounded approximation ratio otherwise. This answers the open question that whether the local-search heuristic has provable approximation guarantees when there are more than two demographic groups, which has existed for a long time. We give a lower bound of ℓ on the swap size keeping the approximation ratio bounded, while the swap size selected in this paper is $(\ell+1)^2$. How to decrease this gap seems to be an interesting question.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62202161, 62172446), Open Project of Xiangjiang Laboratory (22XJ02002, 22XJ03013), Natural Science Foundation of Hunan Province (2023JJ40240), Scientific Research Fund of Hunan Provincial Education Department (23B0597, 23A0462), and National Key Research and Development Program of China (2022YFC3302302).

References

Angelidakis, H.; Kurpisz, A.; Sering, L.; and Zenklusen, R. 2022. Fair and Fast k-Center Clustering for Data Summarization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 669–702.

Arya, V.; Garg, N.; Khandekar, R.; Meyerson, A.; Munagala, K.; and Pandit, V. 2001. Local Search Heuristic for k-Median and Facility Location Problems. In *Proceedings of the 33rd Annual ACM Symposium on Theory of Computing*, 21–29.

Bansal, M.; Garg, N.; and Gupta, N. 2012. A 5-Approximation for Capacitated Facility Location. In *Proceedings of the 20th Annual European Symposium on Algorithms*, volume 7501, 133–144.

Chen, K. 2006. On k-Median Clustering in High Dimensions. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms6*, 1177–1185.

Chiplunkar, A.; Kale, S. S.; and Ramamoorthy, S. N. 2020. How to Solve Fair k-Center in Massive Data Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 1877–1886.

Cohen-Addad, V.; Gupta, A.; Hu, L.; Oh, H.; and Saulpic, D. 2022. An Improved Local Search Algorithm for k-Median. In *Proceedings of the 33rd ACM-SIAM Symposium on Discrete Algorithms*, 1556–1612.

Cohen-Addad, V.; Klein, P. N.; and Mathieu, C. 2019. Local Search Yields Approximation Schemes for k-Means and k-Median in Euclidean and Minor-Free Metrics. *SIAM J. Comput.*, 48(2): 644–667.

Friggstad, Z.; Rezapour, M.; and Salavatipour, M. R. 2019. Local Search Yields a PTAS for k-Means in Doubling Metrics. *SIAM J. Comput.*, 48(2): 452–480.

Friggstad, Z.; and Zhang, Y. 2016. Tight Analysis of a Multiple-Swap Heurstic for Budgeted Red-Blue Median. In *Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming*, volume 55, 75:1–75:13.

Girdhar, Y. A.; and Dudek, G. 2012. Efficient On-Line Data Summarization Using Extremum Summaries. In *Proceeding of the 29th IEEE International Conference on Robotics and Automation*, 3490–3496.

Gupta, S.; Kumar, R.; Lu, K.; Moseley, B.; and Vassilvitskii, S. 2017. Local Search Methods for k-Means with Outliers. *Proc. VLDB Endow.*, 10(7): 757–768.

Hajiaghayi, M.; Khandekar, R.; and Kortsarz, G. 2010. Budgeted Red-Blue Median and Its Generalizations. In *Proceedings of the 18th Annual European Symposium on Algorithms*, volume 6346, 314–325.

Hajiaghayi, M.; Khandekar, R.; and Kortsarz, G. 2012. Local Search Algorithms for the Red-Blue Median Problem. *Algorithmica*, 63(4): 795–814.

Hotegni, S. S.; Mahabadi, S.; and Vakilian, A. 2023. Approximation Algorithms for Fair Range Clustering. In *Proceedings of the 40th International Conference on Machine Learning*.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828.

Kleindessner, M.; Awasthi, P.; and Morgenstern, J. 2019. Fair k-Center Clustering for Data Summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 3448–3457.

Krishnaswamy, R.; Kumar, A.; Nagarajan, V.; Sabharwal, Y.; and Saha, B. 2011. The Matroid Median Problem. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, 1117–1130.

Krishnaswamy, R.; Li, S.; and Sandeep, S. 2018. Constant Approximation for k-Median and k-Means with Outliers via Iterative Rounding. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 646–659.

Li, S. 2011. A 1.488 Approximation Algorithm for the Uncapacitated Facility Location Problem. In *The Proceedings of 38th International Colloquium on Automata, Languages and Programming*, volume 6756, 77–88.

Moens, M.; Uyttendaele, C.; and Dumortier, J. 1999. Abstracting of Legal Cases: The Potential of Clustering Based on the Selection of Representative Objects. *J. Am. Soc. Inf. Sci.*, 50(2): 151–161.

Nguyen, H. L.; Nguyen, T. D.; and Jones, M. 2022. Fair Range k-center. *CoRR*, abs/2207.11337.

Swamy, C. 2014. Improved Approximation Algorithms for Matroid and Knapsack Median Problems and Applications. In *Proceedings of the 17th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems and 18th International Workshop on Randomization and Computation*, volume 28, 403–418.

Thejaswi, S.; Gadekar, A.; Ordozgoiti, B.; and Osadnik, M. 2022. Clustering with Fair-Center Representation: Parameterized Approximation Algorithms and Heuristics. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1749–1759.

Thejaswi, S.; Ordozgoiti, B.; and Gionis, A. 2021. Diversity-Aware k-Median: Clustering with Fair Center Representation. In *Proceedings of the 32nd European Conference on Machine Learning and the 25th European Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 12976, 765–780.

Zhang, P. 2007. A New Approximation Algorithm for the k-Facility Location Problem. *Theor. Comput. Sci.*, 384(1): 126–135.