



# Exact Exponential Algorithms for Clustering Problems

Fedor V. Fomin  

Department of Informatics, University of Bergen, Norway

Petr A. Golovach  

Department of Informatics, University of Bergen, Norway

Tanmay Inamdar<sup>1</sup>  

Department of Informatics, University of Bergen, Norway

Nidhi Purohit<sup>1</sup> 

Department of Informatics, University of Bergen, Norway

Saket Saurabh 

The Institute of Mathematical Sciences, HBNI, Chennai, India

Department of Informatics, University of Bergen, Norway

---

## Abstract

In this paper we initiate a systematic study of exact algorithms for some of the well known clustering problems, namely  $k$ -MEDIAN and  $k$ -MEANS. In  $k$ -MEDIAN, the input consists of a set  $X$  of  $n$  points belonging to a metric space, and the task is to select a subset  $C \subseteq X$  of  $k$  points as *centers*, such that the sum of the distances of every point to its nearest center is minimized. In  $k$ -MEANS, the objective is to minimize the sum of *squares* of the distances instead. It is easy to design an algorithm running in time  $\max_{k \leq n} \binom{n}{k} n^{\mathcal{O}(1)} = \mathcal{O}^*(2^n)$  (here,  $\mathcal{O}^*(\cdot)$  notation hides polynomial factors in  $n$ ). In this paper we design first non-trivial exact algorithms for these problems. In particular, we obtain an  $\mathcal{O}^*((1.89)^n)$  time *exact* algorithm for  $k$ -MEDIAN that works for any value of  $k$ . Our algorithm is quite general in that it does not use any properties of the underlying (metric) space – it does not even require the distances to satisfy the triangle inequality. In particular, the same algorithm also works for  $k$ -MEANS. We complement this result by showing that the running time of our algorithm is asymptotically optimal, up to the base of the exponent. That is, unless the Exponential Time Hypothesis fails, there is no algorithm for these problems running in time  $2^{\epsilon n} \cdot n^{\mathcal{O}(1)}$ .

Finally, we consider the “facility location” or “supplier” versions of these clustering problems, where, in addition to the set  $X$  we are additionally given a set of  $m$  candidate centers (or facilities)  $F$ , and objective is to find a subset of  $k$  centers from  $F$ . The goal is still to minimize the  $k$ -MEDIAN/ $k$ -MEANS/ $k$ -CENTER objective. For these versions we give a  $\mathcal{O}(2^n (mn)^{\mathcal{O}(1)})$  time algorithms using subset convolution. We complement this result by showing that, under the Set Cover Conjecture, the “supplier” versions of these problems do not admit an exact algorithm running in time  $2^{(1-\epsilon)n} (mn)^{\mathcal{O}(1)}$ .

**2012 ACM Subject Classification** Theory of computation → Facility location and clustering; Theory of computation → Parameterized complexity and exact algorithms

**Keywords and phrases** clustering,  $k$ -median,  $k$ -means, exact algorithms

**Digital Object Identifier** 10.4230/LIPIcs.IPEC.2022.13

**Funding** The research leading to these results has received funding from the Research Council of Norway via the project BWCA (grant no. 314528) and the European Research Council (ERC) via grant LOPPRE, reference 819416.

---

<sup>1</sup> Part of this work was done when the two authors were visiting IMSc, Chennai.



## 1 Introduction

Clustering is a fundamental area in the domain of optimization problems with numerous applications. In this paper, we focus on some of the most fundamental problems in the clustering literature, namely  $k$ -MEDIAN,  $k$ -MEANS, and  $k$ -CENTER. We formally define the optimization version  $k$ -MEDIAN.

$k$ -MEDIAN

*Input:* Given a metric space  $(X, d)$ , where  $X = \{x_1, \dots, x_n\}$  is a collection of  $n$  points, with distance function  $d$  on  $X$  and a positive integer  $k$ .

*Task:* Find a pair  $(C, P)$ , where  $C = \{c_1, \dots, c_k\} \subseteq X$  is a set of centers and  $P$  is a partition of  $X$  into  $k$  subsets  $\{X_1, \dots, X_k\}$  (clusters). Here,  $X_i$  is the cluster corresponding to the center  $c_i \in C$ . The goal is to minimize the following cost, over all pairs  $(C, P)$ .

$$\text{cost}(C, P) = \sum_{i=1}^k \sum_{x \in X_i} d(c_i, x)$$

$k$ -MEANS is a variant of  $k$ -MEDIAN, where the only difference is that we want to minimize the sum of squares of the distances, i.e.,  $\sum_{i=1}^k \sum_{x \in X_i} (d(c_i, x))^2$ . In  $k$ -CENTER, the objective is to minimize the maximum distance of a point and its nearest center, i.e.,  $\max_{i=1}^k \max_{x \in X_i} d(c_i, x)$ .

The special cases of  $k$ -MEDIAN have a long history, and they are known in the literature as Fermat-Weber problem [23, 24]. A recent formulation of  $k$ -MEANS can be traced back to Steinhaus [20] and MacQueen [19]. Lloyd proposed a heuristic algorithm [17] for  $k$ -MEANS that is extremely simple to implement for euclidean spaces, and it remains popular even today.  $k$ -CENTER was proved to be NP-complete by Hsu and Nemhauser [10]. All three problems have been studied from the perspective of approximation algorithms for last several decades. These three problems – as well as several of their generalizations – are known to admit constant factor approximations in polynomial time. More recently, these problems have also been studied from the perspective of Fixed-Parameter Tractable (FPT) algorithms, where one allows the running times of the form  $f(k) \cdot n^{\mathcal{O}(1)}$  for some computable function  $f$ .  $k$ -MEDIAN and  $k$ -MEANS are known to admit improved approximation guarantees using FPT algorithms [4], and these approximation guarantees are tight up to certain complexity-theoretic assumptions.

A result that initiated this study is an *exact* algorithm for  $k$ -CENTER<sup>1</sup> by Agarwal and Procopiuc [1], who give an  $n^{\mathcal{O}(k^{1-\frac{1}{d}})}$  time algorithm in  $\mathbb{R}^d$ . In particular, in two dimensional space, their algorithm runs in  $2^{\mathcal{O}(\sqrt{n} \log n)}$  time for any value of  $k$ , i.e., in *sub-exponential* time. This led us towards a natural question, namely, studying the complexity of  $k$ -MEDIAN,  $k$ -MEANS,  $k$ -CENTER in general metrics.

Note that it is easy to design an exact algorithm that runs in time  $\binom{n}{k} \cdot n^{\mathcal{O}(1)}$  – it simply enumerates all sets of centers of size  $k$ , and the corresponding partition of  $X$  into clusters is obtained by assigning each point to its nearest center. Then, we simply return the solution with the minimum cost. However, note that when  $k$  belongs to the range  $n/2 \pm o(n)$ ,  $\binom{n}{k} \simeq 2^n$ . Thus, the naïve algorithm has running time  $\mathcal{O}^*(2^n)$  in the worst case.

<sup>1</sup> We note that the result of [1] holds for a slightly different variant, where the centers can be placed anywhere in  $\mathbb{R}^d$ . This formulation is more natural and standard in euclidean spaces.

For many problems, the running time of  $\mathcal{O}^*(2^n)$  is often achievable by a brute-force enumeration of all the solutions. However, for many NP-hard problems, it is often possible to obtain improved running times. The field of exact algorithms for NP-hard problems is several decades old. In 2003, Woeginger wrote a survey [25] on this topic, which revived the field. This eventually led to a plethora of new results and techniques, such as subset convolution [2], measure and conquer [8], and monotone local search [7]. A detailed survey on this topic can be found in a textbook by Kratsch and Fomin [16]. We study the aforementioned classical clustering problems from this perspective. In other words, we ask whether the classical clustering problems such as  $k$ -MEDIAN and  $k$ -MEANS admit moderately exponential-time algorithms, i.e., algorithms with running time  $c^n \cdot n^{\mathcal{O}(1)}$  for a constant  $c < 2$  that is as small as possible. We indeed answer this question in the affirmative, leading to the following theorem.

► **Theorem 1.** *There is an exact algorithm for  $k$ -MEDIAN ( $k$ -MEANS) in time  $(1.89)^n n^{\mathcal{O}(1)}$ , where  $n$  is the number of points in  $X$ .*

To explain the idea behind this result, consider the following fortuitous scenario. Suppose that the optimal solution only contains clusters of size exactly 2. In this case, it is easy to solve the problem optimally by reducing the problem to finding a minimum-weight matching in the complete graph defining the metric<sup>2</sup>. Note that the problem of finding Minimum-Weight Perfect Matching is known to be polynomial-time solvable by the classical result of Edmonds [13]. This idea can also be extended if the optimal solution only contains clusters of size 1 and 2, by finding matching in an auxiliary graph. However, the idea does not generalize to clusters of size 3 and more, since we need to solve a problem that has a flavor similar to the 3-dimensional matching problem or the “star partition” problem, which are known to be NP-hard [9, 3, 15]. Nevertheless, if the number of points belonging to the clusters of size at least 3 is *small*, one can “guess” these points, and solve the remaining points using matching. However, the number of points belonging to the clusters of size at least 3 can be quite large – it can be as high as  $n$ . But note that the number of *centers* corresponding to clusters of size at least 3 can be at most  $n/3$ . We show that “guessing” the subset of centers of such clusters is sufficient (as opposed to guessing *all* the points in such clusters), in the sense that an optimal clustering of the “residual” instance can be found – again – by finding a minimum-weight matching in an appropriately constructed auxiliary graph.

We briefly explain the idea behind the construction of this auxiliary graph. Note that in order to find an optimal clustering in the “residual” instance, we need to figure out the following things: (1) the set of points that are involved in clusters of size 1, i.e., *singleton* clusters, (2) the pairs of points that become clusters of size 2, and (3) for each center  $c_i$  of a cluster of size at least 3, the set of at least two additional points that are connected to  $c_i$ . We find the set of points of type (1) by matching them to a set of *dummy* points with zero-weight edges. The pairs of points involved in clusters of size 2 naturally correspond to a matching, such that the weight of each edge corresponds to the distance between the corresponding pair of points. Finally, to find points of type (3), we make an appropriate number of *copies* of each guessed center  $c_i$  that will be matched to the corresponding points. Although the high-level idea behind the construction of the graph is very natural, it is non-trivial to construct the graph such that a minimum-weight perfect matching in the auxiliary graph exactly corresponds to an optimal clustering (assuming we guess the centers

<sup>2</sup> Note that the cluster-center always belongs to its own cluster, which implies that a cluster of size 2 contains one *additional* point. This immediately suggests the connection to minimum-weight matching.

correctly). Thus, this construction pushes the boundary of applicability of matching in order to find an optimal clustering. Since the minimum-weight perfect matching problem can be solved in polynomial time, the running time of our algorithm is dominated by guessing the set of centers of clusters of size at least 3. As mentioned previously, the number of such centers is at most  $n/3$ , which implies that the number of guesses is at most  $\binom{n}{n/3} \leq (1.89)^n$ , which dominates the running time of our algorithm. We describe this result in Section 3. We complement these moderately exponential algorithms by showing that these running times are asymptotically optimal. Formally, assuming the Exponential Time Hypothesis (ETH), as formulated by Impagliazzo and Paturi [11], we show that these problems do not admit an algorithms running in time  $2^{o(n)} \cdot n^{\mathcal{O}(1)}$ . A formal definition of ETH is given in Section 2, and we prove the ETH-hardness result in Section 4.

We note that our algorithm as well as the hardness result also holds for  $k$ -CENTER. However, it is folklore that the *exact* versions of  $k$ -CENTER and DOMINATING SET are equivalent. Thus, using the currently best known algorithm for DOMINATING SET by Iwata [12], it is possible to obtain an  $\mathcal{O}^*((1.4689)^n)$  time algorithm for  $k$ -CENTER.

We also consider a “facility location” or “supplier” version, which is a generalization of the clustering problems defined above. In this setting, we are given a set of clients (or points)  $X$ , and a set of facilities (or centers)  $F$ . In general the sets  $X$  and  $F$  may be different, or even disjoint. In these versions, the set of  $k$  centers  $C$  must be chosen from  $F$ , i.e.,  $C \subseteq F$ . We formally state the “supplier” version of  $k$ -MEDIAN, which we call  $k$ -MEDIAN FACILITY LOCATION<sup>3</sup>.

*k*-MEDIAN FACILITY LOCATION

*Input:* Given a metric space  $(X \cup F, d)$ , where  $X = \{x_1, \dots, x_n\}$  of  $n$  points, called clients,  $F$  is a set of  $m$  centers, and a positive integer  $k$ .

*Task:* Find a pair  $(C, P)$ , where  $C = \{c_1, \dots, c_k\} \subseteq F$  of size at most  $k$  and  $P$  is a partition of  $X$  into  $k$  subsets  $\{X_1, \dots, X_k\}$  (clusters) such that each client in cluster  $X_i$  is assigned to center  $c_i$  so as to minimize the  $k$ -median cost of clustering, defined as follows:

$$\text{cost}(C, P) = \sum_{i=1}^k \sum_{x \in X_i} d(c_i, x)$$

It is also possible to define the analogous versions of  $k$ -MEANS and  $k$ -CENTER— the latter has been studied in the approximation literature under the name of  $k$ -SUPPLIER. In this paper, we show that these “facility location” versions of  $k$ -MEDIAN/ $k$ -MEANS/ $k$ -CENTER are computationally harder, as compared to the normal versions, in the following sense. Consider the concrete example of  $k$ -MEDIAN and  $k$ -MEDIAN FACILITY LOCATION. As mentioned earlier, we beat the “trivial” bound of  $\mathcal{O}(2^n)$ , by giving a  $\mathcal{O}((1.89)^n)$  time algorithm for  $k$ -MEDIAN. On the other hand, we show that for  $k$ -MEDIAN FACILITY LOCATION, it is not possible to obtain a  $2^{(1-\epsilon)n} \cdot (mn)^{\mathcal{O}(1)}$  time algorithm for any fixed  $\epsilon > 0$  (note that  $m = |F|$  is the number of facilities and  $n = |X|$  is the number of clients). For showing this result, we use the SET COVER CONJECTURE, which is a complexity theoretic hypothesis proposed by Cygan et al. [5]. We match this lower bound by designing an algorithm with running time  $2^n \cdot (mn)^{\mathcal{O}(1)}$

<sup>3</sup> We note that a slight generalization of this problem has been considered by Jain and Vazirani [14], who called it “a common generalization of  $k$ -median and Facility Location”, and gave a constant approximation in polynomial time.

under some mild assumptions. The details are in Section 6. While this algorithm is not obvious, it is a relatively straightforward application of the subset convolution technique. This algorithm also works for the supplier versions of  $k$ -MEANS and  $k$ -CENTER; however, again there is a much simpler algorithm for  $k$ -SUPPLIER with a similar running time.

Finally, note that designing an algorithm for the supplier versions with running time  $2^m \cdot (mn)^{O(1)}$  is trivial by simple enumeration. It is not known whether the base of the exponent can be improved by showing an algorithm with running time  $(2 - \epsilon)^m (mn)^{O(1)}$  for some fixed  $\epsilon > 0$ , or whether this is not possible assuming a similar complexity-theoretic hypothesis, such as SET COVER CONJECTURE, or Strong Exponential Time Hypothesis (SETH). We leave this open for a future work.

## 2 Preliminaries

We denote by  $G = (V(G), E(G))$  a graph with vertex set  $V(G)$  and edge set  $E(G)$ . Cardinality of a set  $S$  denoted by  $|S|$  is the number of elements of the set. We denote an (undirected) edge between vertices  $u$  and  $v$  as  $uv$ . We denote by  $N(v) = \{u \in V(G) \mid (u, v) \in E(G)\}$  be the *open neighbourhood* (or simply neighbourhood) of  $v$ , and let  $N[v] = N(v) \cup \{v\}$  be the *closed neighbourhood* of  $v$ .

A *matching*  $M$  of a graph  $G$  is a set of edges such that no two edges have common vertices. A vertex  $v \in V(G)$  is said to be saturated by  $M$  if there is an edge in  $M$  incident to  $v$ , otherwise it is said to be unsaturated. We also say that  $M$  saturates  $v$ . We say that a vertex  $u$  is matched to a vertex  $v$  in  $M$  if there is an edge  $e \in M$  such that  $e = (u, v)$ . A perfect matching in a graph  $G$  is a matching which saturates every vertex in  $G$ . Given a weight function  $w: E(G) \rightarrow \mathbb{R}_{\geq 0}$ , the minimum weight perfect matching problem is to find a perfect matching  $M$  (if it exists) of minimum weight  $w(M) = \sum_{e \in M} w(e)$ . It is well known to be solvable in polynomial time by the Blossom algorithm of Edmonds [13].

A  $q$ -CNF formula  $\phi = C_1 \wedge \dots \wedge C_m$  is a boolean formula over  $n$  variables  $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ , such that each clause  $C_i$  is a disjunction of at most  $q$  literals of the form  $x_i$  or  $\neg x_i$ , for some  $1 \leq i \leq n$ . In a  $q$ -SAT instance we are given a  $q$ -CNF formula  $\phi$ , and the question is to decide whether  $\phi$  is satisfiable. Impagliazzo and Paturi [11] formulated the following hypothesis, called Exponential Time Hypothesis. Note that this ETH is a stronger assumption than  $P \neq NP$ .

*Exponential Time Hypothesis (ETH)* states that  $q$ -SAT,  $q \geq 3$  cannot be solved within a running time of  $2^{o(n)}$  or  $2^{o(m)}$ , where  $n$  is the number of variables and  $m$  is the number of clauses in the input  $q$ -CNF formula.

## 3 Proof of Theorem 1

Before delving into the proof of Theorem 1, we discuss the approach at a high level. We begin by “guessing” a subset of centers from an (unknown) optimal solution. For each guess, the problem of finding the best (i.e., minimum-cost) clustering that is “compatible” with the guess is reduced to finding a minimum weight perfect matching in an auxiliary graph  $G$ . The graph  $G$  is constructed in such a way that this clustering can be extracted by essentially looking at the minimum-weight perfect matching. Note that MINIMUM WEIGHT PERFECT MATCHING problem is well known to be solvable in polynomial time by the Blossom algorithm of Edmonds [13]. Finally, we simply return a minimum-cost clustering found over all guesses.

Let us fix some optimal  $k$ -median solution and let  $k_1^*$ ,  $k_2^*$  and  $k_3^*$  be a partition of  $k$ , where  $k_1^*$ : the number of clusters of size exactly 1, call *Type1*;  $k_2^*$ : the number of clusters of size exactly 2, call *Type2*; and  $k_3^*$ : the number of clusters of size at least 3, call *Type3*. Let

## 13:6 Exact Exponential Algorithms for Clustering Problems

$C_3^* \subseteq X$  be *Type3* centers, and say  $C_3^* = \{c_1, \dots, c_{k_3^*}\}$ . Observe that number of clusters with *Type3* centers is at most  $\frac{n}{3}$ . Suppose not, then the number of clusters with *Type3* centers is greater than  $\frac{n}{3}$ . Each *Type3* cluster contains at least three points. This contradicts that the number of input points is  $n$ .

**Algorithm.** First, we guess the partition of  $k$  into  $k_1, k_2, k_3$  as well as a subset  $C_3 \subseteq X$  of size at most  $n/3$ . For each such guess  $(k_1, k_2, k_3, C_3)$ , we construct the auxiliary graph  $G$  (as defined subsequently) corresponding to this guess, and compute a minimum weight perfect matching  $M$  in  $G$ . Let  $M^*$  be a minimum weight perfect matching over *all* the guesses. We extract the corresponding clustering  $(C^*, P^*)$  from  $M^*$  (also explained subsequently), and return as an optimal solution of the given instance.

**Running time.** Note that there are at most  $\mathcal{O}(k^2)$  tuples  $(k_1, k_2, k_3)$  such that  $k_1 + k_2 + k_3 \leq k$  (note that  $k_i$ 's are non-negative integers). Furthermore, there are at most  $\sum_{i=0}^{n/3} \binom{n}{i} \leq (1.89)^n$  subsets of  $X$  of size at most  $n/3$ . Finally, constructing the auxiliary graph, and finding a minimum-weight perfect matching takes polynomial time. Thus, the running time is dominated by the number of guesses for  $C_3$ , which implies that we can bound the running time of our algorithm by  $\mathcal{O}^*((1.89)^n)$ .

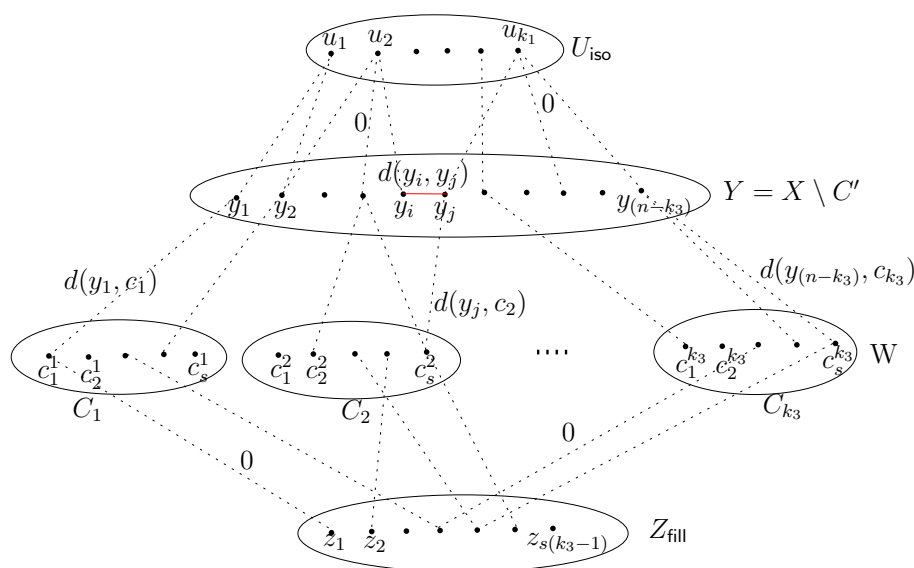
**Construction of Auxiliary Graph.** From now on assume that our algorithm made the right guesses, i.e., suppose that  $(k_1, k_2, k_3) = (k_1^*, k_2^*, k_3^*)$  and  $C_3^* = C'$ . Then, we initialize the *Type3* centers by placing each center from  $C'$  into a separate cluster. At this point, to achieve this, we reduce the problem to the classical MINIMUM WEIGHT PERFECT MATCHING on an auxiliary graph  $G$ , which we define as follows. (See Figure 1 for an illustration of the construction).

- For each  $i \in \{1, \dots, k_3\}$ , construct a set of  $s = n - k_3 - 2k_2 - k_1$  vertices  $C_i = \{c_1^i, \dots, c_s^i\}$ . Denote  $W = \cup_{i=1}^{k_3} C_i$ ; the block of vertices  $C_i$  corresponds to center  $c_i$ .
- Let  $Y = X \setminus C'$ , that is, a set consisting of unclustered points in  $X$ . Observe  $|Y| = n - k_3$ . Denote  $Y = \{y_1, \dots, y_{(n-k_3)}\}$ . For simplicity, we slightly abuse the notation by keeping the vertices in  $G$  same as points in  $Y$ . That is, for each  $i \in \{1, \dots, (n - k_3)\}$ , place a vertex  $y_i$  in the set  $Y$ . Make each  $y_i$  adjacent to all vertices of  $W$ .
- For each  $i \in \{1, \dots, k_1\}$ , construct an auxiliary vertex  $u_i$ . Denote  $U_{\text{iso}} = \{u_1, \dots, u_{k_1}\}$ . Make each  $u_i$  adjacent to every vertex of  $Y$ .
- Construct a set of  $s(k_3 - 1)$  vertices,  $Z_{\text{fill}} = \{z_1, \dots, z_{s(k_3-1)}\}$ , that we call fillers and make vertices of  $Z_{\text{fill}}$  adjacent to the vertices of  $W$ .

We define edge weights. For an edge  $(u, v) \in E(G)$ , we will use  $w(u, v)$  to denote  $w((u, v))$  to avoid clutter.

- For every  $i \in \{1, \dots, (n - k_3)\}$  and every  $j \in \{1, \dots, k_3\}$  set  $w(y_i, c_h^j) = d(y_i, c_j)$  for  $h \in \{1, \dots, s\}$ , i.e, weight of all edges joining  $y_i$  in  $Y$  with the vertices of  $C_i$  corresponding to center  $c_j$ .
- For every  $i, j \in \{1, \dots, n - k_3\}$ ,  $i \neq j$ , set  $w(y_i, y_j) = d(y_i, y_j)$ , i.e, the weight of edges between vertices of  $Y$ .
- For every  $i \in \{1, \dots, k_1\}$  and  $j \in \{1, \dots, (n - k_3)\}$ , set  $w(u_i, y_j) = 0$ , i.e., the edges incident to the vertices of  $U_{\text{iso}}$  have zero weights.
- For every  $i \in \{1, \dots, s(k_3 - 1)\}$  and  $j \in \{1, \dots, k_3\}$ ,  $w(z_i c_h^j) = 0$ , for  $h \in \{1, \dots, s\}$ , i.e., the edges incident to the fillers have zero weights.





■ **Figure 1** Illustration of the graph  $G$  produced in the reduction from  $k$ -MEDIAN to MINIMUM WEIGHT PERFECT MATCHING. To avoid clutter, we only show some representative edges. Recall that we guess the set of  $k_3$  centers of type 3, and corresponding to each such center  $c_i$ , we add a set  $C_i$  consisting of  $s$  copies corresponding to that center. Next, we have the set  $Y$  corresponding to  $n - k_3$  unclustered points. Finally,  $U_{\text{iso}}$  and  $Z_{\text{fill}}$  consist of auxiliary vertices in order to ensure a perfect matching. The weights of vertices among  $Y$  correspond to the corresponding original distance; whereas the weight of an edge between  $y_\ell \in Y$ , and a copy  $c_i^j$  corresponding to a type 3 center  $c_i$  is defined to be  $d(y_\ell, c_i)$ . The weights of all other edges are equal to zero.

► **Lemma 2.** *The graph  $G$  has a perfect matching.*

**Proof.** We construct a set  $M \subseteq E(G)$  that saturates every vertex in  $G$ .

Note that  $|U_{\text{iso}}| < |Y|$  and every vertex of  $U_{\text{iso}}$  is adjacent to every vertex of  $Y$ . Therefore, we can construct  $M_1 \subseteq E(G)$  by arbitrarily mapping each vertex of  $U_{\text{iso}}$  to a distinct vertex of  $Y$ . Clearly,  $M_1$  is matching saturating vertices of  $U_{\text{iso}}$ . Since  $|U_{\text{iso}}| = k_1$ ,  $M_1$  saturates  $k_1$  vertices of  $Y$ . Denote by  $Y'$  the set of vertices of  $Y$  that are not saturated by  $M_1$ . Observe  $|Y'| = s + 2k_2$ .

Every vertex of  $Z_{\text{fill}}$  is adjacent to every vertex of  $W$  and  $|Z_{\text{fill}}| < |W|$ . Construct  $M_2 \subseteq E(G)$  by arbitrarily mapping each vertex of  $Z_{\text{fill}}$  to a distinct vertex of  $W$ . Thus,  $M_2$  is a matching which saturates every vertex of  $Z_{\text{fill}}$  and since  $|Z_{\text{fill}}| = s(k_3 - 1)$ , it also saturates  $s(k_3 - 1)$  vertices of  $W$ . Denote by  $W'$  the set of vertices of  $W$  that is not saturated by  $M_2$ . Observe  $|W'| = s$ . Recall, every vertex of  $W'$  is adjacent to every vertex of  $Y'$  and note that  $|W'| < |Y'|$ . Therefore, construct  $M_3 \subseteq E(G)$  by arbitrarily matching each vertex of  $W'$  with a distinct vertex of  $Y'$ .

Thus, the matching  $M_3$  saturates  $s$  vertices in both the sets  $W'$  and  $Y'$ . Denote  $M' = M_1 \cup M_2 \cup M_3$ . Clearly, the vertices of  $U_{\text{iso}}$ ,  $W$  and  $Z_{\text{fill}}$  are saturated by  $M'$ .

Denote by  $Y'' = Y \setminus Y'$  the set of vertices of  $Y$  that are not saturated by  $M'$ . Note that  $|Y''| = 2k_2$ . Consider  $M_4 \subseteq E(G)$  which maps these  $2k_2$  vertices to each other. We set  $M = M' \cup M_4$ . It is easy to see that  $M$  is a perfect matching. ◀

We next show one-to-one correspondence between perfect matchings of  $G$  and  $k$ -median clusterings of  $X$ .

► **Lemma 3.** *Let  $\text{OPT}_{\text{mm}}(G)$  = weight of minimum weight perfect matching, and  $\text{OPT}_{k\text{med}}(X)$  = optimal clustering cost of  $k$ -median clustering of  $X$ . Then,  $\text{OPT}_{\text{mm}}(G) = \text{OPT}_{k\text{med}}(X)$ .*

**Proof.** In the forward direction, let  $M$  denote a minimum weight perfect matching  $M \subseteq E(G)$ . We construct a  $k$ -median clustering of  $X$  of same cost.

Observe that each vertex of  $Z_{\text{fill}}$  is only adjacent to the vertices of  $W$  and  $|Z_{\text{fill}}| < |W|$ . Let  $W_1 \subseteq W$  be a set of vertices matched to vertices of  $Z_{\text{fill}}$ . Since  $G$  has a perfect matching, it saturates  $Z_{\text{fill}}$ , where  $|Z_{\text{fill}}| = s(k_3 - 1)$ . Then,  $|W_1| = s(k_3 - 1)$ . Let  $W_2 = W \setminus W_1$  be set of vertices matched to vertices of  $Y$ . Clearly,  $|W_2| = s$ .

For every  $i \in \{1, \dots, (n - k_3)\}$ , vertex  $y_i \in Y$  is saturated by  $M$ . Therefore, we construct the  $k$ -median clustering  $\{X_1, \dots, X_k\}$  of  $X$ , where each  $X_i \in \{\text{Type1}, \text{Type2}, \text{Type3}\}$ , for  $i \in \{1, \dots, k\}$  as follows.

Let  $Y' \subseteq Y$  be the set of vertices that are matched to vertices of  $U_{\text{iso}}$  in  $M$ , where  $|U_{\text{iso}}| = k_1 < |Y|$ . Corresponding to each such vertex in  $Y'$ , select a center in the solution  $C$ , call  $C_{\text{Type1}} = \{c_{\text{Type1}}^1, \dots, c_{\text{Type1}}^{k_1}\}$ . Correspondingly, also construct a singleton cluster  $X_i = \{c_{\text{Type1}}^i\}$ , for  $i \in \{1, \dots, k_1\}$ . Let  $X_{\text{Type1}}$  denote set of all *Type1* clusters.

We now construct *Type3* clusters: Let  $Y'_i \subseteq Y$  be the set of vertices matched to set  $C_i$ , for  $i \in \{1, \dots, k_3\}$  in  $M$ . Consider  $X_i = Y'_i \cup \{c_i\}$ . Clearly,  $X_i$ , for  $i \in \{1, \dots, k_3\}$  corresponds to *Type3* clusters in  $X$ . Let  $X_{\text{Type3}}$  denote set of all *Type3* clusters. Recall, we already guess set  $C' = \{c_1, \dots, c_{k_3}\}$ , that is, *Type3* centers correctly.

Lastly, we construct clusters of *Type2*. Denote by  $Y''$  set of unclustered points in  $Y$ . Observe these points form a set of  $k_2$  disjoint edges in  $M$ . Arbitrarily, select one of the endpoint of each edge as a center in the solution  $C$ , call  $C_{\text{Type2}} = \{c_{\text{Type2}}^1, \dots, c_{\text{Type2}}^{k_2}\}$ . That is, for an edge  $(y_1, y_2) \in M$ , where  $y_1, y_2 \in Y''$ , select center as  $y_1$  or  $y_2$ . Then construct a cluster  $X_i$ , for  $i \in \{1, \dots, k_2\}$  by placing both the endpoints of the edge in the same cluster. Denote by  $X_{\text{Type2}}$  the set of all *Type2* clusters.

Clearly,  $X_i \in \{\text{Type1}, \text{Type2}, \text{Type3}\}$ , for  $i \in \{1, \dots, k\}$  is a partition of  $X$ . Note, since *Type1* clusters are isolated points, therefore, they contribute zero to the total cost of clustering. Now we upper bound the cost of the obtained  $k$ -median clustering:

$$\sum_{i=1}^k \sum_{x \in X_i} d(c_i, x) = \sum_{i=1}^{k_2} \sum_{y \in X_{\text{Type2}}} d(c_{\text{Type2}}^i, y) + \sum_{i=1}^{k_3} \sum_{y \in X_{\text{Type3}}} d(c_i, y) = \text{OPT}_{\text{mm}}(G).$$

For the reverse direction, consider a  $k$ -median clustering  $\{X_1, \dots, X_k\}$  of  $X$  into  $\{\text{Type1}, \text{Type2}, \text{Type3}\}$  clusters of  $X$  such that  $|\text{Type1}| = k_1$ ,  $|\text{Type2}| = k_2$  and  $|\text{Type3}| = k_3$  and  $C' = \{c_1, \dots, c_{k_3}\}$ , that is, centers of *Type3* clusters with  $\text{OPT}_{k\text{med}}(X)$ . We construct a perfect matching  $M \subseteq E(G)$  of  $G$  as follows.

Observe that each *Type1* cluster is a singleton cluster. Construct  $M_1 \subseteq E(G)$  by iterating over each singleton vertex in  $Y$  correspond to each cluster and matched it to a distinct vertex in  $U_{\text{iso}}$ . Since  $|\text{Type1}| = |U_{\text{iso}}| = k_1$ ,  $M_1$  is a matching saturating set  $U_{\text{iso}}$ . Also,  $M_1$  saturates  $k_1$  vertices in  $Y$ .

Corresponding to each *Type2* cluster, construct  $M_2 \subseteq E(G)$  by adding an edge between both the end vertices in  $Y$ . Clearly,  $M_2$  is a disjoint set of  $k_2$  edges in  $G$  and saturates  $2k_2$  vertices in  $Y$ .

Denote  $Y' \subseteq Y$  be the set of vertices matched by  $M_1 \cup M_2$ . Clearly,  $|Y'| = k_1 + 2k_2$ . Let  $Y'' = Y \setminus Y'$  be the set of remaining unmatched vertices in  $Y$ . Then,  $|Y''| = |Y| - |Y'| = n - k_3 - 2k_2 - k_1 = s$ .



Note, we already guessed  $C' = \{c_1, \dots, c_{k_3}\}$  and we have a cluster  $X_i$  corresponding to each  $C_i$ , for  $i \in \{1, \dots, k_3\}$ . Construct  $M_3 \subseteq E(G)$  by matching each vertex of  $X_i \setminus \{c_i\}$  in  $Y''$  to a distinct copy of  $c_i$  in  $W$ . Since  $|Y''| < |W|$ ,  $M_3$  saturates  $Y''$ . Let  $W_1 \subseteq W$  be the set of vertices saturated by  $M_3$ . Note that  $|Y''| = s$ , then  $|W_1| = s$ . Let  $W_2 = W \setminus W_1$  be the set of vertices not saturated by  $M_3$ , where  $|W| = sk_3$ . Then,  $|W_2| = s(k_3 - 1)$ . Every vertex of  $Z_{\text{fill}}$  is only adjacent to every vertex of  $W$  (in particular of  $W_2$ ). We construct  $M_4 \subseteq E(G)$  by matching each vertex of  $Z_{\text{fill}}$  to a distinct vertex of  $W_2$ . Since  $|Z_{\text{fill}}| = |W_2| = s(k_3 - 1)$ ,  $M_4$  saturates  $Z_{\text{fill}}$  and  $W_2$ .

To evaluate the weight of  $M$ , recall that the edges of  $G$  incident to set  $U_{\text{iso}}$  and filler vertices  $Z_{\text{fill}}$  have zero weights, that is,  $w(M_1) = w(M_4) = 0$ . Then

$$\begin{aligned} w(M) &= w(M_2) + w(M_3) = \sum_{e \in M_2} w(e) + \sum_{e \in M_3} w(e) \\ &= \sum_{c_i: X_i \in X_{\text{Type2}}} \sum_{y \in C_i} d(y, c_i) + \sum_{c_i: X_i \in X_{\text{Type3}}} \sum_{y \in C_i} d(y, c_i) \\ &= \text{OPT}_{\text{kmed}}(X). \end{aligned}$$

It is straightforward to see that the construction of the graph  $G$  from an instance  $(X, d)$  of  $k$ -MEDIAN can be done in polynomial time. Then, because a perfect matching of minimum weight of the graph  $G$  can be found in polynomial time [13] and the total number of guesses is at most  $(1.89)^n n^{\mathcal{O}(1)}$ ,  $k$ -MEDIAN can be solved exactly in  $(1.89)^n n^{\mathcal{O}(1)}$  time. This completes the proof of the theorem.  $\blacktriangleleft$

► **Remark 4.** Note that even if the distances satisfy the triangle inequality, the sum of *squares* of distances do not. Nevertheless, our algorithm also works for  $k$ -MEANS, where we want to minimize the sum of squares of distances; or even more generally, if we want to minimize the sum of  $z$ -th powers of distances, for some fixed  $z \geq 1$ . In fact, our algorithm works for non-metric distance functions – it is easy to modify construction of graph  $G$  so that it works with asymmetric distance functions, which are quite popular in the context of asymmetric traveling salesman problem [21, 22]. Finally, we note that it may be possible to improve the running time (i.e., the base of the exponent) using the metric properties of distances, and we leave this open for a future work. However, in the next section, we show the running time of an exact algorithm cannot be substantially improved, i.e., to  $\mathcal{O}^*(2^{o(n)})$ .

## 4 ETH Hardness

In this section, we establish result around the (im)possibility of solving  $k$ -MEDIAN problem in subexponential time in the number of points. For this, we use the result of Lokshtanov et al. [18] which states that, assuming ETH, DOMINATING SET problem cannot be solved in time  $2^{o(n)}$  time, where  $n$  is the number of vertices of graph.

Given an unweighted, undirected graph  $G = (V, E)$ , a dominating set  $S$  is a subset of  $V$  such that each  $v \in V$  is dominated by  $S$ , that is, we either have  $v \in S$  or there exists an edge  $(uv) \in E(G)$  such that  $u \in S$ . The decision version of DOMINATING SET is defined as follows.

DOMINATING SET

*Input:* Given an unweighted, undirected graph  $G(V, E)$ , positive integer  $k$ .  
*Task:* Determine whether  $G$  has a dominating set of size at most  $k$ .

Lokshtanov et. al [18] proved the following result.

## 13:10 Exact Exponential Algorithms for Clustering Problems

► **Proposition 5** ([18]). *Assuming ETH, there is no  $2^{o(n)}$  time algorithm for DOMINATING SET problem, where  $n$  is the number of vertices of  $G$ .*

We use this known fact about DOMINATING SET to prove the following.

► **Theorem 6.**  *$k$ -MEDIAN cannot be solved in time  $2^{o(n)}$  time unless the exponential-time hypothesis fails, where  $n$  is the number of points in  $X$ .*

**Proof.** We give a reduction from DOMINATING SET problem to  $k$ -MEDIAN problem. Let  $(G = (V, E), k)$  be the given instance of DOMINATING SET. We assume that there is no dominating set in  $G$  of size at most  $k - 1$ . This assumption is without loss of generality, since we can use the following reduction iteratively for  $k' = 1, 2, \dots, k$ , which only incurs a polynomial overhead.

Now we construct an instance  $(X, d)$  of  $k$ -MEDIAN as follows. First, let  $X = V(G)$ , i.e., we treat each vertex of the graph as a point in the metric space, and we use the terms vertex and point interchangeably. Recall that the graph  $G = (V, E)$  is unweighted, but we suppose that the weight of every edge in  $E(G)$  is 1. Then, we let  $d$  be the shortest path metric in  $G$ . The following observations are immediate.

► **Observation 7.**

- For all  $u \in V(G)$ ,  $d(u, u) = 0$ .
- For all distinct  $u, v \in V(G)$ ,  $d(u, v) = 1 \iff (u, v) \in E(G)$ , and  $d(u, v) \geq 2 \iff (u, v) \notin E(G)$ .

We now show that there is a dominating set of size  $k$  iff there is a  $k$ -median clustering of cost exactly  $n - k$ .

In the forward direction, let  $S \subseteq V(G)$  be a dominating set of size  $k$ . We obtain the corresponding  $k$ -median clustering as follows. We let  $S = \{c_1, c_2, \dots, c_k\}$  to be the set of centers. For a center  $c_i \in S$ , we define  $X'_i = N[c_i]$ . Since  $S$  is a dominating set, every vertex in  $V(G) \setminus S$  has a neighbor in  $S$ . Therefore,  $\bigcup_{1 \leq i \leq k} X'_i = V(G)$ . Now, we remove all *other* centers except for  $c_i$  from the set  $X'_i$ . Furthermore, if a vertex belongs to multiple  $X'_i$ 's, we arbitrarily keep it only a single  $X'_i$ . Let  $\{X_1, X_2, \dots, X_k\}$  be the resulting partition of  $V(G)$ . Observe that in the resulting clustering, centers pay a cost of zero, whereas every other vertex has a center at distance 1. Therefore, the cost of the clustering is exactly  $n - k$ .

In the other direction, let  $(S, \{X_1, X_2, \dots, X_k\})$  be a given  $k$ -median clustering of cost  $n - k$ . We claim that  $S$  is a dominating set of size  $k$ . Consider any vertex  $u \in V(G) \setminus S$ , and suppose  $u \in X_i$  corresponding to the center  $c_i$ . Since  $u \notin S$ ,  $d(u, S) \geq d(u, c_i) \geq 1$ . This holds for all  $n - k$  points of  $V(G) \setminus S$ . Now, if  $u \in X_i$ , and  $d(u, c_i) > 1$  for some vertex  $u \in V(G) \setminus S$ , then this contradicts the assumption that the given clustering has cost  $n - k$ . This implies that every  $u \in V(G) \setminus S$  has a center in  $S$  at distance exactly 1, i.e.,  $u$  has a neighbor in  $S$ . This concludes the proof.

This reduction takes polynomial time. Observe that the number of points in the resulting instance is equal to  $n$ , the number of vertices in  $G$ . Therefore, if there is an algorithm for  $k$ -MEDIAN with running time subexponential in the number of points  $n$  then it would give a  $2^{o(n)}$  time algorithm for DOMINATING SET, which would refute ETH, via Proposition 5. ◀

## 5 SeCoCo Hardness

In this section, we consider the variant of  $k$ -MEDIAN, which we call  $k$ -MEDIAN FACILITY LOCATION. Recall that in this problem, we are given a metric space  $(X \cup F, d)$ , where  $X$  is a set of  $n$  clients,  $F$  is a set of  $m$  centers and integer  $k > 0$ . The goal is to select a set  $C \subseteq F$  of  $k$  centers and assign each client in  $X$  to a center in  $C$ , such that the  $k$ -median cost of clustering is minimized.

We show that there is no algorithm solves  $k$ -MEDIAN FACILITY LOCATION problem in time  $\mathcal{O}(2^{(1-\epsilon)n} \text{poly}(m))$ , for every fixed  $\epsilon > 0$ . For this, we use the SET COVER CONJECTURE by Cygan et al. [5].

The decision version of SET COVER problem is defined as follows.

SET COVER

*Input:* Given a universe  $\mathcal{U} = \{u_1, \dots, u_n\}$  of  $n$  elements and a family  $\mathcal{S} = \{S_1, \dots, S_m\}$  of  $m$  subsets of  $\mathcal{U}$  and an integer  $k$

*Task:* Determine whether there is a set cover of size at most  $k$ .

To state SET COVER CONJECTURE [5] more formally, let  $\Delta$ -SET COVER denote the SET COVER problem where all the sets have size at most  $\Delta > 0$ .

► **Conjecture 8.** SET COVER CONJECTURE (SeCoCo)[5]. For every fixed  $\epsilon > 0$  there is  $\Delta(\epsilon) > 0$ , such that no algorithm (even randomized) solves  $\Delta$ -SET COVER in time  $\mathcal{O}(2^{(1-\epsilon)n} \cdot \text{poly}(m))$ .

Using this result, we show the following.

► **Theorem 9.** Assuming SET COVER CONJECTURE, for any fixed  $\epsilon > 0$ , there is no  $\mathcal{O}(2^{(1-\epsilon)n} \cdot \text{poly}(m))$  time algorithm for  $k$ -MEDIAN FACILITY LOCATION, where  $n$  is the number of clients.

**Proof.** We give a reduction from SET COVER to  $k$ -MEDIAN FACILITY LOCATION problem.

Given an instance  $(\mathcal{U}, \mathcal{S})$  of SET COVER problem, where  $\mathcal{U} = \{u_1, \dots, u_n\}$  and  $\mathcal{S} = \{S_1, \dots, S_m\}$ , such that  $S_i \subseteq \mathcal{U}$ , we create an instance of  $k$ -MEDIAN FACILITY LOCATION by building a bipartite graph  $G = ((X \cup F), E)$  as follows.

- For each element  $u_i \in \mathcal{U}$ , we create a client, say  $x_i$ , for  $i \in \{1, \dots, n\}$ . Denote  $X = \{x_1, \dots, x_n\}$ .
  - For each set  $S_i \in \mathcal{S}$ , we create a center, say  $c_i$ , for  $i \in \{1, \dots, m\}$ . Denote  $F = \{c_1, \dots, c_m\}$ .
  - For every  $i \in \{1, \dots, n\}$  and every  $j \in \{1, \dots, m\}$ , if  $u_i \in S_j$ , then connect corresponding  $x_i$  and  $c_j$  with an edge of weight 1, i.e., client  $x_i$  pays cost 1 when assigned to facility  $c_j$ .
- This finishes the construction of  $G$ . Now, let  $d$  be the shortest path metric in graph  $G$ .

We show that there is set cover of size at most  $k$  if and only if there is  $k$ -median clustering of cost  $n$ .

In the forward direction, assume there is a set cover  $\mathcal{S}' \subseteq \mathcal{S}$  of size at most  $k$ . Assume  $\mathcal{S}' = \{S_1, \dots, S_k\}$ . For a set  $S_i$ , we make the corresponding vertex  $c_i \in F$  a center. Then, we create its corresponding cluster  $X_i$  as follows. We add all the points  $x_j$  such that  $(c_i x_j) \in E$ . Finally, we make the clusters  $X_i$  pairwise disjoint, by arbitrarily choosing exactly one cluster for every client, if the client is present in multiple clusters. Clearly,  $\{X_1, \dots, X_k\}$  is a partition of  $X$ . We now calculate the cost of the obtained  $k$ -median clustering.

$$\sum_{i=1}^k \sum_{x \in X_i} d(c_i, x) = \sum_{i=1}^k |X_i| = n.$$

## 13:12 Exact Exponential Algorithms for Clustering Problems

In the reverse direction, suppose there is a  $k$ -median clustering  $\{X_1, \dots, X_k\}$  of  $X$  of cost  $n$ . Let  $C = \{c_1, \dots, c_k\} \subseteq F$  be a set of centers. Every client must be at distance at least 1 from its corresponding center. We claim that each client in a cluster is at distance exactly 1 from its corresponding center. Suppose not, then there exists a client with distance strictly greater than 1 from its center. The total number of clients is  $n$ . This contradicts that the cost of  $k$ -median clustering is  $n$ . Thus, every element is chosen in some set corresponding to set  $C$ . Therefore, a subfamily  $\mathcal{S}' \subseteq \mathcal{S}$  corresponding to set  $C$  forms a cover of  $\mathcal{U}$ . Since  $|C| = k$ ,  $\mathcal{S}'$  is a cover of  $\mathcal{U}$  of size at most  $k$ .

Clearly, this reduction takes polynomial time. Furthermore, observe that the number of clients in the resulting instance is same as the number of elements in  $\mathcal{U}$ . Therefore, if there is an  $\mathcal{O}(2^{(1-\epsilon)n} \cdot \text{poly}(m))$  time algorithm for  $k$ -MEDIAN FACILITY LOCATION then it would give a  $\mathcal{O}(2^{(1-\epsilon)n} \cdot \text{poly}(m))$  time algorithm for SET COVER, which, in turn, refutes SET COVER CONJECTURE.  $\blacktriangleleft$

We briefly note that the same hardness construction also shows a similar hardness result for the “supplier” versions of  $k$ -MEANS and  $k$ -CENTER.

### 6 A $2^n \cdot \text{poly}(m, n)$ time Algorithm for $k$ -Median Facility Location

Let  $(X \cup F, d)$  be a given instance of  $k$ -MEDIAN FACILITY LOCATION, where  $n = |X|$  denotes the number of clients, and  $m = |F|$  denotes the number of centers. In this section, we give a  $2^n \cdot \text{poly}(m, n)$ -time exact algorithm, under a mild assumption that any distance in the input is a non-negative integer that is bounded by a polynomial in the input size.<sup>4</sup> Let  $M := n \cdot D$ , where  $D$  denotes the maximum inter-point distance in the input. Note that  $M = \text{poly}(m, n)$ .

We define  $k$  functions  $\text{cost}_1, \text{cost}_2, \dots, \text{cost}_k : 2^X \rightarrow M$ , where  $\text{cost}_i(Y)$  denotes the minimum cost of clustering the clients of  $Y$  into at most  $i$  clusters. In other words,  $\text{cost}_i(Y)$  is the optimal  $i$ -MEDIAN FACILITY LOCATION cost, restricted to the instance  $(Y \cup F, d)$ . First, notice that  $\text{cost}_1(Y)$  is simply the minimum cost of clustering all points of  $Y$  into a single cluster. This value can be computed in  $\mathcal{O}(mn)$  time by iterating over all centers in  $F$ , and selecting the center  $c$  that minimizes the cost  $\sum_{p \in Y} d(p, c)$ . Thus, the values  $\text{cost}_1(Y)$  for all subsets  $Y \subseteq X$  can be computed in  $\mathcal{O}(2^n mn)$  time. Next, we have the following observation.

► **Observation 10.** For any  $Y \subseteq X$  and for any  $1 \leq i \leq k$ ,

$$\text{cost}_i(Y) = \min_{\substack{A \cup B = Y \\ A \cap B = \emptyset}} \text{cost}_{i-1}(A) + \text{cost}_1(B).$$

Note that since we are interested in clustering of  $Y$  into *at most*  $i$  clusters, we do not need to “remember” the set of facilities realizing  $\text{cost}_{i-1}(A)$  and  $\text{cost}_1(B)$  in Observation 10. Next, we discuss the notion of subset convolution that will be used to compute  $\text{cost}_i(\cdot)$  values that is faster than the naïve computation.

**Subset Convolutions.** Given two functions  $f, g : 2^X \rightarrow \mathbb{Z}$ , the *subset convolution* of  $f$  and  $g$  is the function  $(f * g) : 2^X \rightarrow \mathbb{Z}$ , defined as follows.

$$\forall Y \subseteq X : \quad (f * g)(Y) = \sum_{\substack{A \cup B = Y \\ A \cap B = \emptyset}} f(A) \cdot g(B) \tag{1}$$

<sup>4</sup> Since the integers are encoded in binary, this implies that the length of the encoding of any distance is  $\mathcal{O}(\log(m) + \log(n))$ .

It is known that, given all the  $2^n$  values of  $f$  and  $g$  in the input, all the  $2^n$  values of  $f * g$  can be computed in  $\mathcal{O}(2^n \cdot n^3)$  arithmetic operations, see e.g., Theorem 10.15 in the Parameterized Algorithms book [6]. This is known as *fast subset convolution*. Now, let  $(f \oplus g)(Y) = \min_{\substack{A \cup B = Y \\ A \cap B = \emptyset}} f(A) + g(B)$ . We observe that  $f \oplus g$  is equal to the subset convolution  $f * g$  in the integer min-sum semiring  $(\mathbb{Z} \cup \{\infty\}, \min, +)$ , i.e., in Equation (1), we use the mapping  $+$   $\mapsto$   $\min$ , and  $\cdot$   $\mapsto$   $+$ . This, combined with a simple “embedding trick” enables one to compute all values of  $f \oplus g : 2^X \rightarrow \{-N, \dots, N\}$  in time  $2^n n^{\mathcal{O}(1)} \cdot \mathcal{O}(N \log N \log \log N)$  using fast subset convolution – see Theorem 10.17 of [6]. Finally, Observation 10 implies that  $\text{cost}_i$  is exactly  $\text{cost}_{i-1} \oplus \text{cost}_1$ , and we observe that the function values are upper bounded by  $n \cdot D = M$ . We summarize this discussion in the following proposition.

► **Proposition 11.** *Given all the  $2^n$  values of  $\text{cost}_{i-1}$  and  $\text{cost}_1$  in the input, all the  $2^n$  values of  $\text{cost}_i$  can be computed in time  $2^n n^{\mathcal{O}(1)} \cdot \mathcal{O}(M \log M \log \log M)$ .*

Using Proposition 11, we can compute all the  $2^n$  values of  $\text{cost}_2(\cdot)$ , using the pre-computed values  $\text{cost}_1(\cdot)$ . Then, we can use the values of  $\text{cost}_2(\cdot)$  and  $\text{cost}_1(\cdot)$  to compute the values of  $\text{cost}_3(\cdot)$ . By iterating in this manner  $k - 1 \leq n$  times, we compute the values of  $\text{cost}_k(\cdot)$  for all  $2^n$  subsets of  $k$ , and the overall time is upper bounded by  $2^n mn^{\mathcal{O}(1)} \cdot \mathcal{O}(M \log M \log \log M)$ , which is  $2^n \cdot \text{poly}(m, n)$ , if  $M = \text{poly}(m, n)$ . Note that  $\text{cost}_k(X)$  corresponds to the optimal cost of  $k$ -MEDIAN FACILITY LOCATION. Finally, the computed values of the functions  $\text{cost}_i(\cdot)$  can be used to also compute a clustering  $\{X_1, X_2, \dots, X_k\}$  of  $X$ , and the corresponding centers  $\{c_1, c_2, \dots, c_k\}$ . We omit the straightforward details.

► **Theorem 12.**  *$k$ -MEDIAN FACILITY LOCATION can be solved optimally in  $2^n \cdot \text{poly}(m, n)$  time, assuming the distances are integers that are bounded by polynomial in the input size.*

Note that the algorithm does not require the underlying distance function to satisfy the triangle inequality. In particular, we obtain an analogous result the “facility location” version of the  $k$ -MEANS objective. Finally, the algorithm works for  $k$ -SUPPLIER, which is a similar variant of  $k$ -CENTER. However, in this case there is a much simpler reduction to SET COVER which gives an  $2^n \cdot \text{poly}(m, n)$  time algorithm. For this, we first “guess” the optimal radius  $r$ , and define a set system that consists of balls of radius  $r$  around the given centers. We omit the details.

---

## References

- 1 Pankaj K Agarwal and Cecilia Magdalena Procopiu. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.
- 2 Andreas Björklund, Thore Husfeldt, and Mikko Koivisto. Set partitioning via inclusion-exclusion. *SIAM Journal on Computing*, 39(2):546–563, 2009.
- 3 Jérémie Chalopin and Daniël Paulusma. Packing bipartite graphs with covers of complete bipartite graphs. *Discret. Appl. Math.*, 168:40–50, 2014. doi:10.1016/j.dam.2012.08.026.
- 4 Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight fpt approximations for  $k$ -median and  $k$ -means. *arXiv preprint*, 2019. arXiv:1904.12334.
- 5 Marek Cygan, Holger Dell, Daniel Lokshtanov, Dániel Marx, Jesper Nederlof, Yoshio Okamoto, Ramamohan Paturi, Saket Saurabh, and Magnus Wahlström. On problems as hard as cnf-sat. *ACM Trans. Algorithms*, 12, 2016.
- 6 Marek Cygan, Fedor V Fomin, Łukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk, and Saket Saurabh. *Parameterized algorithms*, volume 5(4). Springer, 2015.
- 7 Fedor V Fomin, Serge Gaspers, Daniel Lokshtanov, and Saket Saurabh. Exact algorithms via monotone local search. *Journal of the ACM (JACM)*, 66(2):1–23, 2019.

## 13:14 Exact Exponential Algorithms for Clustering Problems

- 8 Fedor V Fomin, Fabrizio Grandoni, and Dieter Kratsch. A measure & conquer approach for the analysis of exact algorithms. *Journal of the ACM (JACM)*, 56(5):1–32, 2009.
- 9 M. R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.
- 10 Wen-Lian Hsu and George L Nemhauser. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1(3):209–215, 1979.
- 11 Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *J. Comput. Syst. Sci.*, 9:367–375, 2001.
- 12 Yoichi Iwata. A faster algorithm for dominating set analyzed by the potential method. In *International Symposium on Parameterized and Exact Computation*, pages 41–54. Springer, 2011.
- 13 Edmonds Jack. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.
- 14 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 (JACM)*, 48(2):274–296, 2001.
- 15 David G. Kirkpatrick and Pavol Hell. On the complexity of general graph factor problems. *SIAM J. Comput.*, 12(3):601–609, 1983. doi:10.1137/0212040.
- 16 Dieter Kratsch and FV Fomin. *Exact exponential algorithms*. Springer-Verlag Berlin Heidelberg, 2010.
- 17 Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.
- 18 Daniel Lokshtanov, Dániel Marx, and Saket Saurabh. Lower bounds based on the exponential time hypothesis. *Bull. EATCS*, 105:41–72, 2011.
- 19 James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1(14), pages 281–297. Oakland, CA, USA, 1967.
- 20 Hugo Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.*, 1(804):801, 1956.
- 21 Ola Svensson, Jakub Tarnawski, and László A Végh. A constant-factor approximation algorithm for the asymmetric traveling salesman problem. *Journal of the ACM (JACM)*, 67(6):1–53, 2020.
- 22 Vera Traub and Jens Vygen. An improved approximation algorithm for atsp. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pages 1–13, 2020.
- 23 Wikipedia. Geometric median – Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Geometric%20median&oldid=1061179886>, 2022. [Online; accessed 21-April-2022].
- 24 Wikipedia. Weber problem – Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Weber%20problem&oldid=916663348>, 2022. [Online; accessed 21-April-2022].
- 25 Gerhard J Woeginger. Exact algorithms for np-hard problems: A survey. In *Combinatorial optimization – eureka, you shrink!*, pages 185–207. Springer, 2003.