Journal of Computer and System Sciences 136 (2023) 171–194

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

journal homepage: www.elsevier.com/locate/jcss

Parameterized complexity of categorical clustering with size constraints ☆

Fedor V. Fomin, Petr A. Golovach, Nidhi Purohit *

Department of Informatics, University of Bergen, PB 7803, 5020 Bergen, Norway

ARTICLE INFO

Article history:
Received 4 October 2021
Received in revised form 19 February 2023
Accepted 24 March 2023
Available online 13 April 2023

Keywords:
Categorical clustering
k-median
Fixed-parameterized algorithm
Kernelization

ABSTRACT

In the Categorical Clustering problem, we are given a set of vectors (matrix) $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ over Σ^m, where Σ is a finite alphabet, and integers k and B. The task is to partition $\mathbf{A}$ into k clusters such that the median objective of the clustering in the Hamming norm is at most B. Fomin, Golovach, and Panolan [ICALP 2018] proved that the problem is fixed-parameter tractable for the binary case $\Sigma = \{0, 1\}$. We extend this algorithmic result to a popular capacitated clustering model, where in addition the sizes of the clusters are lower and upper bounded by integer parameters p and q, respectively. Our main theorem is that the problem is solvable in time $2^{\mathcal{O}(B \log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$.

1. Introduction

While many problems in machine learning concern numerical data, there is a large class of problems about learning from categorical data. The term categorical data refers to the type of data whose values are discrete and belong to a specific finite set of categories. It could be text, some numeric values, or even unstructured data like images. The most popular clustering objectives for numerical data are k-means and k-median, which are based on distances in the ℓ_1 and ℓ_2-norm. For categorical data, other metrics, like Hamming distance, could be much more useful.

We study the parameterized complexity of clustering problems with constraints on the sizes of the clusters. The need for clustering with constraints comes from various applications. The survey of Banerjee and Ghosh [5] contains various examples of clustering with balancing constraints in Direct Marketing [39], Category Management [33], Clustering of Documents [3,28], and Energy Aware Sensor Networks [22,23] among others. However, introducing constraints on the sizes of clusters usually makes clustering tasks much more computationally challenging.

In this paper, we focus on categorical data clustering, where data features admit a fixed number of possible values. We work with vectors from Σ^m, where Σ is a finite alphabet. The most commonly used similarity measure for categorical data is the Hamming distance. For two vectors $\mathbf{a}, \mathbf{b} \in \Sigma^m$ or, equivalently, for two strings of length m over Σ, we use $d_H(\mathbf{a}, \mathbf{b})$, to denote the *Hamming distance* between $\mathbf{a}$ and $\mathbf{b}$, that is, the number of indices $i \in \{1, \ldots, m\}$ where the i-th elements of $\mathbf{a}$ and $\mathbf{b}$ differ. The task of the vanilla Categorical Clustering problem is, given an $m \times n$ matrix $\mathbf{A}$ with columns $(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ over a finite alphabet Σ, a positive integer k, and a nonnegative integer B, decide whether there is a partition $\{I_1, \ldots, I_k\}$ of $\{1, \ldots, n\}$ and vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$ such that

☆ A preliminary version of the paper appeared in the proceedings of WADS 2021. The research leading to these results has been supported by the Research Council of Norway via the project BWCA (grant no. 314528) and the European Research Council (ERC) via grant LOPPRE, reference 819416.

* Corresponding author.
E-mail addresses: fedor.fomin@uib.no (F.V. Fomin), petr.golovach@uib.no (P.A. Golovach), nidhi.purohit@uib.no (N. Purohit).

https://doi.org/10.1016/j.jcss.2023.03.006

$$\sum_{i=1}^{k}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq B.$$

The sets $I_1,\ldots,I_k$ are called *clusters* and the vectors $\mathbf{c}_1,\ldots,\mathbf{c}_k$ are *medians* (or *centers*).[1] We consider the generalization of the problem, where the size of each cluster should be within a given interval:

CAPACITATED CLUSTERING

Input: An $m\times n$ matrix $\mathbf{A}$ with columns $(\mathbf{a}_1,\ldots,\mathbf{a}_n)$ over a finite alphabet Σ, a positive integer k, a nonnegative integer B, and positive integers p and q such that $p\leq q$.

Task: Decide whether there is a partition $\{I_1,\ldots,I_k\}$ of $\{1,\ldots,n\}$, where $p\leq |I_i|\leq q$, and vectors $\mathbf{c}_1,\ldots,\mathbf{c}_k\in\Sigma^m$ such that

$$\sum_{i=1}^{k}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq B.$$

Parameterized algorithms for the vanilla variant of CAPACITATED CLUSTERING (without constraints on the sizes of clusters) were given by Fomin, Golovach, and Panolan in [19]. One of the main results of their paper is the theorem providing an algorithm of running time $2^{\mathcal{O}(B\log B)}\cdot(nm)^{\mathcal{O}(1)}$ for vanilla clustering over the binary field. In other words, the problem is fixed-parameter tractable (FPT) parameterized by B. The main question that we address in this paper is whether clustering constraints impact the problem's parameterized complexity.

Our results Our main result is that CAPACITATED CLUSTERING is fixed-parameter tractable when parameterized by the budget B and the alphabet size. More precisely, we show the following:

Theorem 1. CAPACITATED CLUSTERING *can be solved in* $2^{\mathcal{O}(B\log B)}|\Sigma|^B\cdot(mn)^{\mathcal{O}(1)}$ *time.*

Fomin, Golovach, and Panolan [19, Theorem 1] proved that CATEGORICAL CLUSTERING for binary matrices is FPT when parameterized by the budget B. Theorem 1 generalizes this result. Interestingly, for approximation algorithms, introducing clustering constraints makes the problem much more computationally challenging. However, from the parameterized complexity perspective, adding constraints on the sizes of clusters does not change the complexity of the problem. We note that Theorem 1 is tight in the sense that it is unlikely that the dependence on the alphabet size could be made polynomial because the results of Fomin, Golovach, and Simonov [20] imply that CAPACITATED CLUSTERING is W[1]-hard when parameterized by B and m.

We also observe that CAPACITATED CLUSTERING is NP-complete even for binary matrices, $k=2$ and $p=q=\frac{n}{k}$. Theorem 1 can be used to establish fixed-parameter tractability of several other variants of constrained clustering discussed in the literature. In some applications, it is natural to require that the sizes of clusters be approximately equal, see e.g. [37]. We consider variants of CATEGORICAL CLUSTERING, where the input contains additional parameters besides a matrix $\mathbf{A}=(a_1,\ldots,a_n)$ and integers k and B, and the task is to find clusters $I_1,\ldots,I_k$ and medians $\mathbf{c}_1,\ldots,\mathbf{c}_r\in\Sigma^m$ such that $\sum_{i=1}^{k}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq B$ and the sizes of the clusters satisfy special balance properties.

- In BALANCED CLUSTERING, we are additionally given a nonnegative integer δ and it should hold that $||I_i|-|I_j||\leq\delta$ for all $i,j\in\{1,\ldots,k\}$, that is, the sizes of clusters can differ by at most δ.
- In FACTOR-BALANCED CLUSTERING, we are given a real $\alpha\geq 1$ and it is required that $|I_i|\leq\alpha|I_j|$ for all $i,j\in\{1,\ldots,k\}$, that is, the ratio of the sizes of the clusters is upper bounded by α.

By making use of Theorem 1, we prove that BALANCED CLUSTERING and FACTOR-BALANCED CLUSTERING are solvable in time $2^{\mathcal{O}(B\log B)}|\Sigma|^B\cdot(mn)^{\mathcal{O}(1)}$.

Finally, we discuss kernelization for these problems. In particular, we show that BALANCED CLUSTERING admits a polynomial kernel under the combined parameterization by k, B, and δ. We also observe that neither of the considered problems has a polynomial kernel when parameterized by B only, unless coNP $\subseteq$ NP/poly, even for the binary case.

High-level overview of the proof of Theorem 1 The algorithm for the vanilla problem of Fomin et al. [19], as well as the algorithm of Fomin, Golovach and Simonov for clustering in ℓ_p-norm [20], uses the result of Marx [30] about the enumeration of subhypergraphs with certain properties of a given hypergraph of a special type. Basically, these algorithms can be seen as an intricate reduction of a clustering instance to a hypergraph of a special type and then calling the result of Marx as a

[1] Some authors call $\mathbf{c}_1,\ldots,\mathbf{c}_k$ *means* in the case of Hamming distances.

black box. In the context of the categorical clustering problems, a similar reduction implies that all potential medians can be listed in $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time (see Lemma 4).

However, this strategy does not work to prove Theorem 1. Here the difficulties are due to the constraints on the sizes of clusters. The algorithm for Categorical Clustering in [19] uses an observation that identical columns $\mathbf{a}_i$ and $\mathbf{a}_j$ of $\mathbf{A}$ can be clustered together. That is, $i, j \in I_h$ for a cluster I_h of an optimal solution. Hence, a solution can be seen as a partition of the family of *initial* clusters, i.e., inclusion maximal sets of indices $J \subseteq \{1,\ldots,n\}$ such that the columns $\mathbf{a}_i$ for $i \in J$ are the same. Since the number of initial clusters that are part of *composite* clusters of a solution, that is, clusters including at least two initial clusters, is at most $2B$ in any yes-instance, the color coding technique of Alon, Yuster, and Zwick [2] allows to highlight initial clusters that may be included in a single composite cluster of a solution. This way, the initial problem is reduced to selecting a single composite cluster of the minimum cost that contains a given number of initial clusters. To solve this problem, the result of Marx [30] about the enumeration of subhypergraphs becomes handy.

This scheme does not work for Capacitated Clustering because splitting of an initial cluster between clusters of a solution may be inevitable due to size constraints. This makes it impossible to select composite clusters independently from each other and destroys the approach used in [19,20].

The main insight that allows overcoming the above issues is the very specific structure of possible splitting of initial clusters (Lemma 3). For a clustering $\mathcal{I} = \{I_1,\ldots,I_k\}$ and the partition $\mathcal{J}$ of the column indices into initial clusters, we look at the structure of the intersection graph $G(\mathcal{I},\mathcal{J})$ defined by the partitions $\mathcal{I}$ and $\mathcal{J}$ of $\{1,\ldots,n\}$. The crucial fact we prove here is that there is an optimal solution such that this intersection graph is a forest. It can be seen that $G(\mathcal{I},\mathcal{J})$ has at most $3B$ vertices in connected components with at least three vertices for such a solution. This allows us to guess the structure of $G(\mathcal{I},\mathcal{J})$, that is, guess a forest F isomorphic to $G(\mathcal{I},\mathcal{J})$, by using brute force. Then for a given F, we find a solution $\mathcal{I}$ with $G(\mathcal{I},\mathcal{J})$ isomorphic to F by combining dynamic programming with color coding and enumeration of subhypergraphs of Marx.

Related work Clustering is one of the most common procedures in unsupervised machine learning. Capacitated Clustering is the variant of the popular k-median clustering with the Hamming norm. In many applications of clustering, constraints come naturally. For example, the lower bound on the size of a cluster ensures certain anonymity of data and is often required for data privacy [36]. There is a rich literature on approximation algorithms for various versions of capacitated clustering [1,7,6,8,10,14,26,12,9,29,37]. However, to the best of our knowledge, no parameterized algorithms for categorical clustering with constraints on the sizes of clusters were known prior to our work.

Several approximations and parameterized algorithms are known for the vanilla case of Categorical Clustering without constraints can be found in the literature. For the binary field, Categorical Clustering was introduced by Kleinberg, Papadimitriou, and Raghavan [24] as one of the examples of segmentation problems. The problem appears under different names in the literature [11,31]. Feige proved in [16] that the problem is NP-complete for every $k \geq 2$. We use several ideas from Feige's construction for our lower bounds. Ostrovsky and Rabani [34] gave a randomized PTAS for the binary Categorical Clustering problem which was recently improved to EPTAS in [18] and [4]. Fomin, Golovach and Simonov in [20] studied k-clusterings with various distance norms in Categorical Clustering. One of their results is that clustering with Hamming-distance (ℓ_0-distance) (but unbounded size of the alphabet Σ) is W[1]-hard parameterized by $m + B$. The following paper about the binary variant of Categorical Clustering is highly relevant to this paper. Fomin, Golovach, and Panolan [19] gave two parameterized algorithms for the binary case of Categorical Clustering with the running time $2^{\mathcal{O}(B\log B)} \cdot (nm)^{\mathcal{O}(1)}$ and $2^{\mathcal{O}(\sqrt{kB\log(k+B)\log k})} \cdot (nm)^{\mathcal{O}(1)}$.

Organization of the paper In Section 2, we introduce basic notions and notation used throughout the paper. We also show some auxiliary claims. In particular, we show that Capacitated Clustering is NP-complete for $k = 2$ and binary matrices even if the clusters are required to be of the same size. In Section 3, we show our main result by constructing an FPT algorithm for Capacitated Clustering parameterized by $B + |\Sigma|$. In Section 4, we discuss Balanced Clustering and Factor-Balanced Clustering. Further, in Section 5, we discuss kernelization for clustering problems with size constraints. We conclude in Section 6, by stating some open problems.

2. Preliminaries

In this section, we introduce the terminology used throughout the paper and obtain some auxiliary results.

2.1. Basic notions

Matrices and vectors All matrices and vectors considered in this paper are assumed to be over a finite alphabet Σ and we say that a matrix (vector) is *binary* if $\Sigma = \{0,1\}$. Therefore, to simplify notation, we omit Σ in the notation whenever it does not create confusion. We use m and n to denote the number of rows and columns, respectively, of input matrices if it does not create confusion. We write $\mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_n)$ to denote that $\mathbf{A}$ is a matrix with n columns $\mathbf{a}_1,\ldots,\mathbf{a}_n$. For a partition $\mathcal{I} = \{I_1,\ldots,I_k\}$ of $\{1,\ldots,n\}$, we say that $\{I_1,\ldots,I_k\}$ is a *k-clustering* for $\mathbf{A}$. For an inclusion maximal $J \subseteq \{1,\ldots,n\}$ such that the columns $\mathbf{a}_i$ are identical for all $i \in J$, we say that J is an *initial cluster*. We say that a cluster I_i of $\mathcal{I}$ is *simple* if $I_i \subseteq J$ for some initial cluster J and I_i is *composite*, otherwise, that is, if I_i contains some $h, j \in \{1,\ldots,n\}$ such that $\mathbf{a}_h$ and

$\mathbf{a}_j$ are distinct. For a vector $\mathbf{a} \in \Sigma^m$, we use $\mathbf{a}[i]$ to denote the i-th element of the vector for $i \in \{1, \ldots, m\}$. Thus, for two vectors $\mathbf{a}, \mathbf{b} \in \Sigma^m$, $d_H(\mathbf{a}, \mathbf{b}) = |\{i \in \{1, \ldots, m\} \mid \mathbf{a}[i] \neq \mathbf{b}[i]\}|$. Let a_{ij} for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$ be the elements of $\mathbf{A}$. For $I \subseteq \{1, \ldots, m\}$ and $J \subseteq \{1, \ldots, n\}$, we denote by $\mathbf{A}[I, J]$ the $|I| \times |J|$-submatrix of $\mathbf{A}$ with the elements a_{ij} where $i \in I$ and $j \in J$.

Parameterized complexity We refer to the books of Cygan et al. [13] and Downey and Fellows [15] for a detailed introduction to the field, see also the recent book of Fomin et al. on kernelization [17]. Here, we just informally sketch basic notions.

The input of a parameterized problem contains an integer value k that is referred to as a *parameter*. A parameterized problem is *fixed-parameter tractable* (FPT) if there is an algorithm solving it in $f(k) \cdot |I|^{\mathcal{O}(1)}$ time, where I is an input, k is a parameter, and $f(\cdot)$ is a computable function; the parameterized complexity class FPT consists of fixed-parameter tractable problems.

A *kernelization algorithm*, or simply a *kernel*, for a parameterized problem P is an algorithm that, given an instance (I, k) of P, in polynomial in $|I|$ and k time returns an instance (I', k') of P such that (i) (I, k) and $(I,' k')$ are equivalent, that is, (I, k) is a yes-instance if and only if (I', k') is a yes-instance, and (ii) $|I'| + k' \leq g(k)$ for some computable function $g(k)$. It is said that $g(\cdot)$ is the *size* of a kernel; if $g(\cdot)$ is a polynomial, then the kernel is polynomial. It is well-known that every FPT problem admits a kernel but, up to some reasonable complexity assumptions, there are FPT problems that have no polynomial kernels. The typical assumption is that NP $\not\subseteq$ coNP/poly (see [17] for details).

2.2. Solutions, clusters, and medians

Formally, for Categorical Clustering and its variants, a solution is formed by clusters $I_1, \ldots, I_k$ together with the corresponding medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$. However, given clusters $I_1, \ldots, I_k$, optimal medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$ can be computed by the easy *majority rule*. Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ and let $\{I_1, \ldots, I_k\}$ be a k-clustering. For every $i \in \{1, \ldots, k\}$, we compute $\mathbf{c}_i \in \Sigma^m$ as follows. For each $j \in \{1, \ldots, m\}$, we consider the multiset $R_{ij} = \{\mathbf{a}_h[j] \mid h \in I_i\}$ of elements of Σ. For each $s \in R_{ij}$, we compute the number of its occurrences in the multiset and find an element s^* that occurs most often (ties are broken arbitrarily). Then we set $\mathbf{c}_i[j] = s^*$. It is straightforward to verify that for every $\mathbf{c} \in \Sigma^m$, $\sum_{h \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_h) \leq \sum_{h \in I_i} d_H(\mathbf{c}, \mathbf{a}_h)$. Therefore, the choice of $\mathbf{c}_i$ is optimal. This gives the following observation.

Observation 1. *Given a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ and a k-clustering $\{I_1, \ldots, I_k\}$, a family of vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$ such that*

$$\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j)$$

is minimum can be computed in polynomial time by the majority rule.

For a k-clustering $\{I_1, \ldots, I_k\}$, we define the *cost* $\mathrm{cost}(I_1, \ldots, I_k)$ as the minimum value of $\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j)$ over all k-tuples of vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$. By Observation 1, we have that $\mathrm{cost}(I_1, \ldots, I_k)$ can be computed in polynomial time. Then the task of Categorical Clustering and its variants is reduced to finding a k-clustering of cost at most B (with the respective constraints of the cluster sizes). Thus, we may refer to a k-clustering as a solution without specifying medians.

Observe that given vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$, we can find a k-clustering $\{I_1, \ldots, I_k\}$ that minimizes $\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j)$ by the greedy procedure. For each $i \in \{1, \ldots, n\}$, we find $j \in \{1, \ldots, k\}$ such that $d_H(\mathbf{c}_j, \mathbf{a}_i)$ is minimum (ties are broken arbitrarily) and place i in the cluster I_j. Since

$$\sum_{i=1}^{n} \min\{d_H(\mathbf{c}_j, \mathbf{a}_i) \mid 1 \leq j \leq k\} \leq \sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j),$$

for every k-clustering $\{I_1, \ldots, I_k\}$, the described greedy procedure produces optimal partition of $\{1, \ldots n\}$ (some sets may be empty). However, the constructed k-clustering does not respect the size constraints of our problems. Still, given vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$, we can decide in polynomial time whether an instance of Capacitated Clustering has a solution with the medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$ using a reduction to the classical Minimum Weight Perfect Matching problem on bipartite graphs that is well-known to be solvable in polynomial time by the Hungarian method of Kuhn [25] (see also [27]).

Recall that a *matching* M of a graph G is a set of edges without common vertices. It is said that a matching M *saturates* a vertex v if M has an edge incident to v. A matching M is *perfect* if every vertex of G is saturated. The task of Minimum Weight Perfect Matching is, given a bipartite graph G and a weight function $w\colon E(G) \to \mathbb{Z}_{\geq 0}$, to find a perfect matching M (if it exists) such that its weight $w(M) = \sum_{e \in M} w(e)$ is minimum.

Lemma 1. *There is a polynomial time algorithm that, given an instance $(\mathbf{A}, \Sigma, k, B, p, q)$ of* Capacitated Clustering *and $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$, decides whether the instance has a solution with the family of medians $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$.*

Proof. Let $(\mathbf{A}, \Sigma, k, B, p, q)$ be an instance of Capacitated Clustering, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. Clearly, we can assume that the parameter k in the instance equals the number of vectors $\mathbf{c}_i$ as, otherwise, $\mathbf{c}_1, \ldots, \mathbf{c}_k$ cannot be the medians. We also assume without loss of generality that $p \leq \frac{n}{k} \leq q$; otherwise, $(\mathbf{A}, \Sigma, k, B, p, q)$ is a no-instance.

We construct the bipartite graph G as follows.

- For each $i \in \{1, \ldots, k\}$, construct a set of p vertices $W_i = \{v_1^i, \ldots, v_p^i\}$ and a set of $q - p$ vertices $W_i' = \{v_{p+1}^i, \ldots, v_q^i\}$; note that $W_i' = \emptyset$ if $p = q$. Let $V_i = W_i \cup W_i'$ for $i \in \{1, \ldots, k\}$ and denote $V = \bigcup_{i=1}^k V_i$; the block of vertices V_i corresponds to the median c_i.
- For each $i \in \{1, \ldots, n\}$, construct a vertex u_i corresponding to the column $\mathbf{a}_i$ of $\mathbf{A}$ and make u_i adjacent to the vertices of V. Denote $U = \{u_1, \ldots, u_n\}$.
- Construct a set of $s = kq - n$ vertices $U' = \{u_1', \ldots, u_s'\}$ that we call *fillers* and make the vertices of U' adjacent to the vertices of W_j' for all $j \in \{1, \ldots, k\}$; note that $U' = \emptyset$ if $n = qk$ and observe that $qk \geq n$ by our assumption about the instance of Capacitated Clustering.

Observe that G is a bipartite graph, where $U \cup U'$ and V form the bipartition. Note also that $|U \cup U'| = |V| = qk$.

We define the edge weights as follows.

- For every $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$, set $w(u_i v_h^j) = d_H(\mathbf{c}_j, \mathbf{a}_i)$ for $h \in \{1, \ldots, q\}$, that is, the weights of all edges joining u_i corresponding to $\mathbf{a}_i$ with the vertices of V_j corresponding to the median $\mathbf{c}_j$ are the same and coincide with the Hamming distance between $\mathbf{a}_i$ and $\mathbf{c}_j$.
- For every $i \in \{1, \ldots, s\}$ and $j \in \{1, \ldots, k\}$, set $w(u_i' v_h^j) = 0$ for $h \in \{p+1, \ldots, q\}$, that is, the edges incident to the fillers have zero weights.

We now show that G has a perfect matching of weight at most B if and only if there is a k-clustering $\{I_1, \ldots, I_k\}$ for A such that $p \leq |I_i| \leq q$ for all $i \in \{1, \ldots, k\}$ and $\sum_{i=1}^k \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq B$.

In the forward direction, assume G has a perfect matching $M \subseteq E(G)$ of weight at most B. We construct the clustering $\{I_1, \ldots, I_k\}$ as follows. For every $h \in \{1, \ldots, n\}$, u_h is saturated by M and, therefore, there are $i_h \in \{1, \ldots, k\}$ and $j_h \in \{1, \ldots q\}$ such that edge $u_h v_{j_h}^{i_h} \in M$. Consider $M' = \{u_h v_{j_h}^{i_h} \mid 1 \leq h \leq n\} \subseteq M$. We cluster the columns of $\mathbf{A}$ according to M'. Formally, we place h in I_{i_h} for each $h \in \{1, \ldots, n\}$. Clearly, $\{I_1, \ldots, I_k\}$ is a partition of $\{1, \ldots, n\}$. Observe that for each $i \in \{1, \ldots, k\}$, the vertices of W_i are adjacent only to the vertices of U. Since these vertices are saturated by M, we obtain that $|I_i| \geq p$ for every $i \in \{1, \ldots, k\}$. Since $|V_i| = q$, $|I_i| \leq q$ for all $i \in \{1, \ldots, k\}$. Now we upper bound the cost of the obtained k-clustering:

$$\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) = \sum_{h=1}^{n} d_H(\mathbf{c}_{i_h}, \mathbf{a}_h) = w(M') \leq w(M) \leq B.$$

For the reverse direction, consider a k-clustering $\{I_1, \ldots, I_k\}$ for A such that $p \leq |I_i| \leq q$ for all $i \in \{1, \ldots, k\}$ and $\sum_{i=1}^k \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq B$. Let $i \in \{1, \ldots, k\}$. Consider the cluster I_i and assume that $I_i = \{j_1, \ldots, j_{h_i}\}$. Recall that every vertex of V_i is adjacent to every vertex of U. Let $M_i = \{u_{j_1} v_1^i, \ldots, u_{j_{h_i}} v_{h_i}^i\}$. Clearly, M_i is a matching saturating the first $p \leq h_i \leq q$ vertices of V_i. In particular, the vertices of W_i are saturated. We construct M_i for every $i \in \{1, \ldots, k\}$ and set $M' = \bigcup_{i=1}^k M_i$. Since $\{I_1, \ldots, I_k\}$ is a partition of $\{1, \ldots, n\}$, M' is a matching saturating every vertex of U. Denote by V' the set of vertices of V that are not saturated by M'. Notice that $V' \subseteq \bigcup_{i=1}^k W_i'$ because the vertices of each W_i are saturated by M_i. Observe that every vertex of U' is adjacent to every vertex of W_i' for $i \in \{1, \ldots, k\}$, that is, $G[U' \cup V']$ is a complete bipartite graph. Because $|U'| = |V'| = s$, $G[U' \cup V']$ has a perfect matching M''. We set $M = M' \cup M''$. It is easy to see that M is a matching and, since M saturates every vertex of G, M is a perfect matching. To evaluate the weight of M, recall that the edges of G incident to the fillers have zero weights, that is, $w(M'') = 0$. Then

$$\begin{aligned} w(M) = w(M') = w\Big(\bigcup_{i=1}^{k} M_i\Big) &= \sum_{i=1}^{k} \sum_{e \in M_i} w(e) = \sum_{i=1}^{k} \big(w(u_{j_1} v_1^i) + \cdots + w(u_{j_{h_i}} v_{h_i}^i)\big) \\ &= \sum_{i=1}^{k} \sum_{j \in I_i} d_{\mathbf{H}}(\mathbf{c}_i, \mathbf{a}_j) \leq B, \end{aligned}$$

and we conclude that M is a perfect matching of weight at most B.

It is straightforward to see that the construction of the graph G from an instance $(\mathbf{A}, \Sigma, k, B, p, q)$ of Capacitated Clustering can be done in polynomial time. Then, because a perfect matching of minimum weight in G can be found in polynomial time [25,27], Capacitated Clustering can be solved in polynomial time. This completes the proof of the lemma. □

By Lemma 1, we have that solving our problems can be reduced to finding a family of medians $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ (notice that some medians may be the same).

2.3. Hardness of clustering

Since we are interested in the parameterized complexity of clustering problems, in the last part of this section, we argue that Balanced Clustering, Factor-Balanced Clustering, and Capacitated Clustering are NP-hard for very restricted instances.

In [16], Feige proved that Categorical Clustering is NP-complete for $k = 2$ and binary matrices, that is, for the case $\Sigma = \{0, 1\}$. This result immediately implies that Capacitated Clustering is also NP-complete for $k = 2$ and binary matrices. To see it, note that an instance $(\mathbf{A}, \Sigma, k, B)$ of Categorical Clustering is equivalent to the instance $(\mathbf{A}, \Sigma, k, B, p, q)$ of Capacitated Clustering for $p = 1$ and $q = n$. However, we would like to underline that Capacitated Clustering is NP-hard even if $p = q$. For this, we use some details of the hardness proof of Feige [16].

Feige proved that Categorical Clustering is NP-hard by showing a reduction from the Max-Cut problem [16]. In Max-Cut, we are given a graph G and a nonnegative integer ℓ, and the task is to find a *cut* $(S, \overline{S})$, that is, a partition of the vertex set into a set S and its complement $\overline{S} = V(G) \setminus S$ such that the size of the cut, i.e., the number of edges between S and $\overline{S}$ is at least ℓ. The reduction constructed by Feige has the property given in the following lemma.

Lemma 2 *([16]). There is a polynomial time reduction from* Max-Cut *to* Categorical Clustering *that computes from an instance* (G, ℓ) *of* Max-Cut *an instance* $(\mathbf{A}, \Sigma, 2, B)$ *of* Categorical Clustering*, where* $\Sigma = \{0, 1\}$*, such that the following holds: if* (G, ℓ) *is a yes-instance of* Max-Cut *with a cut* $(S, \bar{S})$ *of size at least* ℓ*, then* $(\mathbf{A}, \Sigma, 2, B)$ *is a yes-instance of* Categorical Clustering *that has a solution* $\{I_1, I_2\}$ *with the property that* $|I_1|/|I_2| = |S|/|\overline{S}|$.

Theorem 2. *For every fixed integer constant* $c \geq 0$*,* Capacitated Clustering *is* NP*-complete for* $k = 2$*, binary matrices and* $q - p \leq c$.

Proof. We show the theorem by a reduction from Max-Cut that is well-known to be NP-complete [21]. Given an instance (G, ℓ) of Max-Cut, we construct an auxiliary instance $(G', 2\ell)$ of Max-Cut, where G' is the union of two disjoint copies G_1 and G_2 of G. Then for the constructed instance $(G', 2\ell)$ we can use as a black box the algorithm of Fiege [16] from Lemma 2 to produce the instance $(\mathbf{A}, \Sigma, 2, B)$ of Categorical Clustering with $\Sigma = \{0, 1\}$. We further set $p = q = |V(G)|$ and consider the instance $(\mathbf{A}, \Sigma, 2, B, p, q)$ of Capacitated Clustering. Clearly, $q - p \leq c$. We show that (G, ℓ) is a yes-instance of Max-Cut if and only if $(\mathbf{A}, \Sigma, 2, B, p, q)$ is a yes-instance of Capacitated Clustering.

In the forward direction, assume that (G, ℓ) is a yes-instance of Max-Cut and let $(S, \bar{S})$ be a cut of size at least ℓ. Let S_1 and S_2 be the copies of S in G_1 and G_2, respectively. We now consider $S' \subseteq V(G')$ such that $S' = S_1 \cup (V(G_2) \setminus S_2)$. Clearly, $\overline{S'} = (V(G_1) \setminus S_1) \cup S_2$ and $(S', \overline{S'})$ is a cut of G' of size at least 2ℓ. Moreover, $|S'| = |S_1| + |V(G_2) \setminus S_2| = |S_2| + |V(G_1) \setminus S_1| = |\overline{S'}|$. Hence, $(G', 2\ell)$ is a yes-instance of Max-Cut with a solution $(S', \overline{S'})$ that has the property that $|S'| = |\overline{S'}|$. By Lemma 2, $(\mathbf{A}, \Sigma, 2, B)$ is a yes-instance of Categorical Clustering that has a solution $\{I_1, I_2\}$ such that $|I_1| = |I_2|$. This implies that $p \leq |I_1|, |I_2| \leq q$. Therefore, $\{I_1, I_2\}$ is also solution for the instance $(\mathbf{A}, \Sigma, 2, B, p, q)$ of Capacitated Clustering. Thus, $(\mathbf{A}, \Sigma, 2, B, p, q)$ is a yes-instance of Capacitated Clustering.

In the reverse direction, suppose that $(\mathbf{A}, \Sigma, 2, B, p, q)$ is a yes-instance of Capacitated Clustering. Then there is a 2-clustering $\{I_1, I_2\}$ for $\mathbf{A}$ of cost at most B. This means that $(\mathbf{A}, \Sigma, 2, B)$ is a yes-instance of Categorical Clustering. Because $(\mathbf{A}, \Sigma, 2, B)$ is obtained from $(G', 2\ell)$ by a polynomial reduction from Lemma 2, $(G', 2\ell)$ is a yes-instance of Max-Cut, that is, G' has a cut of size at least 2ℓ. Since G' is a disjoint union of two identical copies of G, each copy has a cut of size at least l. Therefore, (G, ℓ) is a yes-instance of Max-Cut. This completes the hardness proof. □

3. FPT algorithm for parameterization by B and the alphabet size

In this section, we show that Capacitated Clustering is FPT when parameterized by B and $|\Sigma|$. Our main result is Theorem 1 that we restate here.

Theorem 1. Capacitated Clustering *can be solved in* $2^{\mathcal{O}(B \log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ *time.*

Note that this result is tight in the sense that it is unlikely that the dependence on the alphabet size could be made polynomial. It was shown in [20], that Categorical Clustering is W[1]-hard when parameterized by B and the number of rows m of the input matrix if $\Sigma = \mathbb{Z}$, i.e., for an infinite alphabet. However, it is straightforward to see that this result holds for $\Sigma = \{0, \ldots, n-1\}$ because our measure is the Hamming distance. For each row of the input matrix, we can replace the original symbols by the symbols of $\Sigma = \{0, \ldots, n-1\}$ in such a way that the original symbols in the row are the same if and only if the new symbols are the same. Clearly, this replacement gives an equivalent instance. This immediately leads to the following proposition.

Proposition 1. Capacitated Clustering *is* W[1]*-hard when parameterized by* B *and* m.

The remaining part of the section contains the proof of Theorem 1. The proof is constructive. In Subsection 3.1, we introduce some notation and show technical claims that are used by the algorithm, and Subsection 3.2 contains the algorithm and its analysis.

3.1. Technical lemmata

Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be a matrix. Recall that an inclusion maximal $J \subseteq \{1, \ldots, n\}$ such that the columns $\mathbf{a}_i$ are identical for all $i \in J$ is called an initial cluster. Suppose that $\{I_1, \ldots, I_k\}$ is a k-clustering for $\mathbf{A}$. Recall that a cluster I_i is simple if $I_i \subseteq J$ for some initial cluster J and I_i is composite, otherwise, that is, if I_i contains some $h, j \in \{1, \ldots, n\}$ such that $\mathbf{a}_h$ and $\mathbf{a}_j$ are distinct.

We start by making the following observation about medians of sufficiently big (in B) clusters.

Observation 2. *Let $\{I_1, \ldots, I_k\}$ be a k-clustering for a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ of cost at most B, and let $|I_i| \geq B + 1$ for some $i \in \{1, \ldots, k\}$. Then for all vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$ such that $\sum_{h=1}^{k} \sum_{j \in I_h} d_H(\mathbf{c}_h, \mathbf{a}_j) \leq B$, $\mathbf{c}_i = \mathbf{a}_j$ for at least $|I_i| - B$ indices $j \in I_i$. Moreover, if $|I_i| \geq 2B + 1$, then $\mathbf{c}_i$ is unique.*

Proof. To show the first part of the claim, assume that $\mathbf{c}_i \in \Sigma^m$ is distinct from at least $B + 1$ columns $\mathbf{a}_j$ for $j \in I_i$. Then

$$B \geq \sum_{h=1}^{k} \sum_{j \in I_h} d_H(\mathbf{c}_h, \mathbf{a}_j) \geq \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \geq B + 1;$$

a contradiction. For the second part of the claim, note that if $|I_i| \geq 2B + 1$, then $\mathbf{c}_i$ should coincide with more than half of the columns $\mathbf{a}_j$ with $j \in I_i$ and, therefore, the choice of $\mathbf{c}_i$ is unique. □

We use the following simple observation about the number of composite clusters and the number of initial clusters having elements in the composite clusters of a solution.

Observation 3. *Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be a matrix with the partition $\mathcal{J} = \{J_1, \ldots, J_s\}$ of $\{1, \ldots, n\}$ into initial clusters. Let also $\mathcal{I} = \{I_1, \ldots, I_k\}$ be a k-clustering for $\mathbf{A}$ of cost at most B. Then $\mathcal{I}$ contains at most B composite clusters and $\mathcal{J}$ has at most $2B$ initial clusters with nonempty intersections with the composite clusters of $\mathcal{I}$.*

Proof. Let $\mathbf{c}_1, \ldots, \mathbf{c}_k$ be medians such that $\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq B$. Note that if I_i is a composite cluster for some $i \in \{1, \ldots, k\}$, then $\mathbf{c}_i$ is distinct from $\mathbf{a}_j$ for at least one $j \in I_i$ and $\sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \geq 1$. Therefore, $\mathcal{I}$ contains at most B composite clusters. For the second claim, notice that if $\mathcal{J}$ has $t \geq B$ initial clusters with nonempty intersections with composite clusters, then because $\mathcal{I}$ has at most B composite clusters, for at least $t - B$ of these initial clusters J_j, $\mathbf{a}_h \neq \mathbf{c}_i$ for $h \in J_j$ and all the medians $\mathbf{c}_i$ of composite clusters. Hence, $t \leq 2B$. □

Let $J \subseteq \{1, \ldots, n\}$ be an initial cluster. Due to size constraints, it may happen that a k-clustering $\{I_1, \ldots, I_k\}$ with several simple clusters $I_i \subseteq J$ provides a solution. This means, that we should partition a subset of J into blocks of bounded size. To verify whether we are able to create such a partition, we use the following observation.

Observation 4. *Let p and q be positive integers, $p \leq q$. A finite set X can be partitioned into h subsets such that each of them has size at least p and at most q if and only if $\left\lceil \frac{|X|}{q} \right\rceil \leq h \leq \left\lfloor \frac{|X|}{p} \right\rfloor$.*

Proof. If X can be partitioned into h subsets of size at least p and at most q, then, trivially, $ph \leq |X|$ and $qh \geq |X|$, i.e., $\left\lceil \frac{|X|}{q} \right\rceil \leq h \leq \left\lfloor \frac{|X|}{p} \right\rfloor$. If $\left\lceil \frac{|X|}{q} \right\rceil \leq h \leq \left\lfloor \frac{|X|}{p} \right\rfloor$, then X has h disjoint subsets $X_1, \ldots, X_h$ of size p. Then the remaining $|X| - ph$ elements can be greedily added to these subsets without exceeding the upper bound q on the size. □

Let $\mathcal{J} = \{J_1, \ldots, J_s\}$ be the partition of $\{1, \ldots, n\}$ into initial clusters. For a k-clustering $\mathcal{I} = \{I_1, \ldots, I_k\}$, we define the graph $G(\mathcal{I}, \mathcal{J})$ as the intersection graph of the sets of $\mathcal{I}$ and $\mathcal{J}$, that is, $G(\mathcal{I}, \mathcal{J})$ is the bipartite graph with the set of vertices $\mathcal{I} \cup \mathcal{J}$ such that for every $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, s\}$, I_i and J_j are adjacent if and only if $I_i \cap J_j \neq \emptyset$. We show that we can assume $G(\mathcal{I}, \mathcal{J})$ to be a forest. This can be proved using an ILP or flow formulation of the clustering problem with given medians. For simplicity, we provide a direct proof.

Lemma 3. *Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be a matrix with the partition $\mathcal{J} = \{J_1, \ldots, J_s\}$ of $\{1, \ldots, n\}$ into initial clusters. Also, let $\mathcal{I} = \{I_1, \ldots, I_k\}$ be a k-clustering for $\mathbf{A}$. Then there is a k-clustering $\mathcal{I}' = \{I'_1, \ldots, I'_k\}$ such that (i) $|I_i| = |I'_i|$ for all $i \in \{1, \ldots, k\}$, (ii) $\mathsf{cost}(I'_1, \ldots, I'_k) \leq \mathsf{cost}(I_1, \ldots, I_k)$, and (iii) $G(\mathcal{I}', \mathcal{J})$ is a forest.*

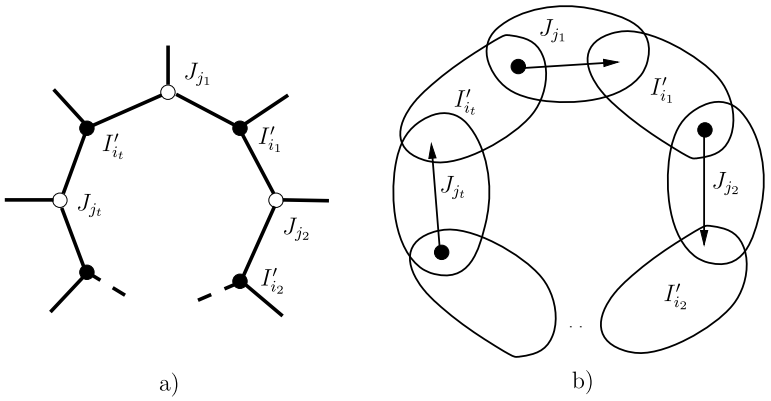

Fig. 1. A cycle in $G(\mathcal{I}', \mathcal{J})$ and the cluster rearrangement scheme.

Proof. Assume that $\mathcal{I}' = \{I'_1, \ldots, I'_k\}$ is a k-clustering for $\mathbf{A}$ satisfying conditions (i) and (ii) such that the number of edges of $G(\mathcal{I}', \mathcal{J})$ is minimum. Denote by $\mathbf{c}_1, \ldots, \mathbf{c}_k$ optimal medians for $I'_1, \ldots, I'_k$. We claim that $G(\mathcal{I}', \mathcal{J})$ is a forest.

The proof is by contradiction. Assume that $G(\mathcal{I}', \mathcal{J})$ has a cycle. This means that there are distinct $i_1, \ldots, i_t \in \{1, \ldots k\}$ and distinct $j_1, \ldots, j_t \in \{1, \ldots, s\}$ such that $I'_{i_h} \cap J_{j_h} \neq \emptyset$ and $I'_{i_h} \cap J_{j_{h+1}} \neq \emptyset$ for all $h \in \{1, \ldots, t\}$; here and further in the proof, we assume that $j_{t+1} = j_1$ and $i_{t+1} = i_1$ (see Fig. 1(a)).

For $h \in \{1, \ldots, s\}$, denote by $\mathbf{b}_h$ the vector coinciding with $\mathbf{a}_{h'}$ for $h' \in J_h$. We observe that either

$$\sum_{h=1}^{t}(d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h}) + d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}})) \geq 2\sum_{h=1}^{t} d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h}) \tag{1}$$

or

$$\sum_{h=1}^{t}(d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h}) + d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}})) \geq 2\sum_{h=1}^{t} d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}}) \tag{2}$$

because the sums of the left and right parts of inequalities (1) and (2) are the same.

We assume without loss of generality that (1) holds, as the second case is symmetric. This means that

$$\sum_{h=1}^{t} d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}}) \geq \sum_{h=1}^{t} d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h}). \tag{3}$$

We iteratively modify $\mathcal{I}'$ by moving a representative of J_{j_h} in $I_{i_{h-1}}$ to I_{i_h} for $h \in \{2, \ldots, t+1\}$, that is, representatives are moved cyclically without changing the cluster sizes (see Fig. 1(b)). We show that this procedure does not increase the clustering cost with respect to the medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$.

Formally, we construct the k-clusterings $\mathcal{I}^{(0)}, \mathcal{I}^{(1)}, \ldots$, where $\mathcal{I}^{(p)} = \{I_1^{(p)}, \ldots, I_k^{(p)}\}$ for $p = 0, 1, \ldots$, starting from $\mathcal{I}^{(0)} = \mathcal{I}'$ while $J_{j_{h+1}} \cap I_{i_h}^{(p)} \neq \emptyset$ for all $h \in \{1, \ldots, t\}$.

Assume that $\mathcal{I}^{(p)} = \{I_1^{(p)}, \ldots, I_k^{(p)}\}$ is constructed and $J_{j_{h+1}} \cap I_{i_h}^{(p)} \neq \emptyset$ for all $h \in \{1, \ldots, t\}$. For every $h \in \{1, \ldots, t\}$, let $i'_h \in J_{j_{h+1}} \cap I_{i_h}^{(p)}$. We define $\mathcal{I}^{(p+1)} = \{I_1^{(p+1)}, \ldots, I_k^{(p+1)}\}$ by setting

$$I_{i_h}^{(p+1)} = (I_{i_h}^{(p)} \setminus \{i'_h\}) \cup \{i'_{h-1}\}$$

for all $h \in \{1, \ldots, t\}$ assuming that $i'_0 = i'_t$, and we set $I_q^{(p+1)} = I_q^{(p)}$ for $q \in \{1, \ldots, k\} \setminus \{i_1, \ldots, i_t\}$. Clearly, $|I_i^{(p+1)}| = |I_q^{(p)}|$ for all $i \in \{1, \ldots, r\}$. We have that

$$\begin{aligned}
\Big(\sum_{i=1}^{k}\sum_{j\in I_i^{(p)}} d_H(\mathbf{c}_i, \mathbf{a}_j)\Big) - \Big(\sum_{i=1}^{k}\sum_{j\in I_i^{(p+1)}} d_H(\mathbf{c}_i, \mathbf{a}_j)\Big) &= \sum_{h=1}^{t}(d_H(\mathbf{c}_{i_h}, \mathbf{a}_{i'_h}) - d_H(\mathbf{c}_{i_h}, \mathbf{a}_{i'_{h-1}})) \\
&= \sum_{h=1}^{t}(d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}}) - d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h})) \\
&= \Big(\sum_{h=1}^{t}(d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_{h+1}})\Big) - \Big(\sum_{h=1}^{t} d_H(\mathbf{c}_{i_h}, \mathbf{b}_{j_h})\Big) \geq 0,
\end{aligned}$$

where the last inequality follows from (3). This means that the cost of the k-clustering $\mathcal{I}^{(p+1)}$ with respect to the medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$ is at most the cost of $\mathcal{I}^{(p)}$ with respect to the same medians.

The next k-clustering $\mathcal{I}^{(p+1)}$ is constructed from $\mathcal{I}^{(p)}$ if $J_{j_{h+1}} \cap I^{(p)}_{i_h} \neq \emptyset$ for all $h \in \{1, \ldots, t\}$. Thus, the sequence is finite, and for the last k-clustering $\mathcal{I}^{(q)}$, there is $h \in \{1, \ldots, t\}$ such that $J_{j_{h+1}} \cap I^{(q)}_{i_h} = \emptyset$, that is, $I^{(q)}_{i_h}$ and $J_{j_{h+1}}$ are not adjacent in $G(\mathcal{I}^{(q)}, \mathcal{J})$. Note that the rearrangement of elements of clusters does not create new adjacencies in $G(\mathcal{I}^{(q)}, \mathcal{J})$ because no cluster gets representatives of an initial cluster that had no representatives in it. We conclude that $G(\mathcal{I}^{(q)}, \mathcal{J})$ has less edges than $G(\mathcal{I}', \mathcal{J})$ but this contradicts the choice of $\mathcal{I}'$. Therefore, $G(\mathcal{I}', \mathcal{J})$ is a forest and $\mathcal{I}'$ satisfies conditions (i)–(iii) as required. □

Next, we show that, given a matrix $\mathbf{A}$, we can list all potential medians for a k-clustering of cost at most B in FPT when B and $|\Sigma|$ are parameters. We show this by making use of the nontrivial result of Marx [30] about the enumeration of subhypergraphs with bounded partial edge cover. This result already proved to be very useful for designing FPT algorithms for clustering problems [19,20].

Recall that a hypergraph $\mathcal{H}$ is a pair $(V, \mathcal{E})$, where V is a set of *vertices* and $\mathcal{E}$ is a family of subsets of V called *hyperedges*. Similarly to graphs, we denote by $V(\mathcal{H})$ the set of vertices and by $\mathcal{E}(\mathcal{H})$ the set of hyperedges. For a vertex v, we denote by $\mathcal{E}_{\mathcal{H}}(v)$ the set of hyperedges containing v, that is, $\mathcal{E}_{\mathcal{H}}(v) = \{E \in \mathcal{E}(\mathcal{H}) \mid v \in E\}$.

Let $\mathcal{G}$ be a hypergraph and let $U \subseteq V(\mathcal{G})$. We say that a hypergraph $\mathcal{H}$ *appears at U as a subhypergraph* if there is a bijection $\pi : V(\mathcal{H}) \to U$ with the property that for every $E \in \mathcal{E}(\mathcal{H})$, there is $E' \in \mathcal{E}(\mathcal{G})$ such that $\pi(E) = E' \cap U$.

A *fractional hyperedge cover* of a hypergraph $\mathcal{H}$ is a function $\varphi : \mathcal{E}(\mathcal{H}) \to [0, 1]$ such that for every vertex $v \in V(\mathcal{H})$, $\sum_{E \in \mathcal{E}_{\mathcal{H}}(v)} \varphi(E) \geq 1$, that is, the sum of the values assigned by f of the hyperedges containing v is at least one. The *fractional cover number* $\rho^*(\mathcal{H})$ of $\mathcal{H}$ is the minimum value $\sum_{E \in \mathcal{E}(\mathcal{H})} \varphi(E)$ taken over all fractional hyperedge covers φ of H.

Proposition 2 *([30]). Let $\mathcal{H}$ be a hypergraph with fractional cover number $\rho^*(\mathcal{H})$, and let $\mathcal{G}$ be a hypergraph whose hyperedges have size at most ℓ. There is an algorithm that enumerates, in $|V(\mathcal{H})|^{\mathcal{O}(|V(\mathcal{H})|)} \cdot \ell^{|V(\mathcal{H})|\rho^*(\mathcal{H})+1} \cdot |\mathcal{E}(\mathcal{G})|^{\rho^*(\mathcal{H})+1} \cdot |V(\mathcal{G})|^2$ time, every $U \subseteq V(\mathcal{G})$ where $\mathcal{H}$ appears at U as subhypergraph in $\mathcal{G}$.*

We apply this result similarly to [20] and, therefore, only briefly sketch the proof of the following lemma.

Lemma 4. *There is an algorithm that, given a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ and a nonnegative integer B, in $2^{\mathcal{O}(B \log B)} \cdot |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time outputs a set $\mathcal{M}(\mathbf{A}, B) \subseteq \Sigma^m$ of size $2^{\mathcal{O}(B \log B)} \cdot |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ such that for every k-clustering $\{I_1, \ldots, I_k\}$ for $\mathbf{A}$ of cost at most B, there are $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathcal{M}(\mathbf{A}, B)$ such that $\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq B$.*

Proof. Let $\mathcal{S}$ be the set of distinct columns of $\mathbf{A}$. Initially, we set $\mathcal{M}(\mathbf{A}, B) := \mathcal{S}$.

For every $\mathbf{s} \in \mathcal{S}$, we construct the hypergraph $\mathcal{G}_{\mathbf{s}}$ with the vertex set $\{1, \ldots, m\}$ with hyperedges corresponding to the columns of $\mathbf{A}$ at Hamming distance at most B from $\mathbf{s}$: for every $i \in \{1, \ldots, n\}$ such that $d_H(\mathbf{s}, \mathbf{a}_i) \leq B$, we introduce the hyperedge

$$E_i = \{j \mid 1 \leq j \leq m \text{ and } \mathbf{a}_i[j] \neq \mathbf{s}[j]\},$$

that is, the hyperedge contains indices, where $\mathbf{s}$ differs from $\mathbf{a}_i$. Note that $|E_i| \leq B$.

Consider an arbitrary k-clustering $\{I_1, \ldots, I_k\}$ for $\mathbf{A}$ of cost at most B. Let $i \in \{1, \ldots, k\}$ and let $\mathbf{s} \in \mathcal{S}$ be such that $\mathbf{s} = \mathbf{a}_j$ for some $j \in I_i$. Let also $\mathbf{c}_i \in \Sigma^m$ be an optimal median for I_i, that is, $\sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}^j)$ is minimum. Notice that if $|I_i| \geq B + 1$, then by Observation 2, every feasible median for I_i is a column of $\mathbf{A}$ and these columns are already placed in $\mathcal{M}(\mathbf{A}, B)$. Also, if $\mathbf{c}_i = \mathbf{s}$, then $\mathbf{c}_i \in \mathcal{M}(\mathbf{A}, B)$. Assume that $|I_i| \leq B$ and $\mathbf{c}_i \neq \mathbf{s}$. Clearly, $d_H(\mathbf{c}_i, \mathbf{s}) \leq B$. Moreover, for any $j \in I_i$, $d_H(\mathbf{s}, \mathbf{a}_j) \leq B$. This holds trivially if $\mathbf{s} = \mathbf{a}_j$. Otherwise, if $\mathbf{s} \neq \mathbf{a}_j$, we have that $d_H(\mathbf{s}, \mathbf{a}_j) \leq d_H(\mathbf{c}_i, \mathbf{s}) + d_H(\mathbf{c}_i, \mathbf{a}_j) \leq \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq B$. Let

$$D = \{j \mid 1 \leq j \leq m \text{ and } \mathbf{c}_i[j] \neq \mathbf{s}[j]\},$$

that is, D is the set of indices where $\mathbf{s}$ differs from the median $\mathbf{c}_i$.

We consider the hypergraph $\mathcal{H}_i$ with the vertex set D whose edges correspond to the columns $\mathbf{a}_j$ for $j \in I_i$. For each $j \in I_i$, we construct the hyperedge

$$F_j = \{h \mid h \in D \text{ and } \mathbf{a}_j[h] \neq \mathbf{s}[h]\},$$

that is, each hyperedge contains indices from D, where $\mathbf{s}$ differs from $\mathbf{a}_j$. We claim that the fractional cover number $\rho^*(\mathcal{H}_i) \leq 2$.

To show this, we define the function $\varphi(F) = \frac{2}{|\mathcal{E}(\mathcal{H}_i)|}$ for every hyperedge F of $\mathcal{H}_i$. We prove that φ is a fractional hyperedge cover of $\mathcal{H}_i$. Thus, we have to show that for every $j \in D$, $\sum_{F \in \mathcal{E}_{\mathcal{H}_i}(j)} \varphi(F) \geq 1$. This is equivalent to proving that for every $j \in D$, at least half of the hyperedges of $\mathcal{H}_i$ contain j. Assume that this is not the case, i.e., there is $j \in D$ such that more than half of hyperedges do not contain j. This mean that for more than half of columns $\mathbf{a}_h$ for $h \in I_i$, $\mathbf{s}[j] = \mathbf{a}_h[j] = s$.

However, by the definition of D, $\mathbf{s}[j] \neq \mathbf{c}_i[j]$ and, therefore, $\mathbf{c}_i[j] \neq s$. This contradicts the assumption that $\mathbf{c}_i$ is an optimal median for I_i because replacing the current value $\mathbf{c}_i[j]$ by s decreases the cost. Hence, φ is a fractional hyperedge cover. Then

$$\rho^*(\mathcal{H}_i) \leq \sum_{F \in \mathcal{E}(\mathcal{H}_i)} \varphi(F) = \sum_{F \in \mathcal{E}(\mathcal{H}_i)} \frac{2}{|\mathcal{E}(\mathcal{H}_i)|} = 2.$$

Observe that $\mathcal{H}_i$ appears in $\mathcal{G}_{\mathbf{s}}$ at D because for each $j \in I_i$, $d_H(\mathbf{s}, \mathbf{a}_j) \leq B$, that is, for every $j \in I_i$, $\mathcal{G}_{\mathbf{s}}$ contains the hyperedge E_j corresponding to $\mathbf{a}_j$; the mapping $\pi : V(\mathcal{H}_i) \to D$ is the identity function.

We obtain that $\mathcal{H}_i$ is a hypergraph with the fractional cover number at most 2 that appears in $\mathcal{G}_{\mathbf{s}}$ at D. Notice that, given $\mathbf{s}$ and D, we can list the vectors over Σ^m that differ from $\mathbf{s}$ in the indices from D and the total number of such vectors is at most $|\Sigma|^B$ because $|D| \leq B$. Then $\mathbf{c}_i$ appears in this list. This leads to the following algorithm. We consider all hypergraphs $\mathcal{H}$ on at most B vertices with at most B hyperedges. Then for each $\mathcal{H}$ and every $\mathbf{s} \in \mathcal{S}$, we use the algorithm of Marx from Proposition 2 to enumerate every $D \subseteq V(\mathcal{G}_{\mathbf{s}})$ where $\mathcal{H}$ appears in $\mathcal{G}_D$ as subhypergraph. Then for every D, we list the vectors that differ from $\mathbf{s}$ in the indices from D by brute force. Then these vectors are included in $\mathcal{M}(\mathbf{A}, B)$.

For given $\mathcal{H}$ and $\mathbf{s}$, the sets D can be enumerated in time $2^{\mathcal{O}(B \log B)} \cdot B^{2B+1} \cdot n^3 \cdot m^2$ by Proposition 2. Then generating the vectors that differ from $\mathbf{s}$ in D can be done in $|\Sigma|^B \cdot n^{\mathcal{O}(1)}$ time as we can assume that $|\Sigma| \leq n$. However, we need $2^{\mathcal{O}(B^2)}$ time to generate all hypergraphs with at most B vertices and at most B hyperedges. This gives the total running time $2^{\mathcal{O}(B^2)} \cdot |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ and the same bound on the size of $\mathcal{M}(\mathbf{A}, B)$.

The running time can be improved by proving that there is a subhypergraph $\mathcal{H}'_i$ of $\mathcal{H}$ with $V(\mathcal{H}'_i) = V(\mathcal{H}_i)$ and $\mathcal{E}(\mathcal{H}'_i) \subseteq \mathcal{E}(\mathcal{H}_i)$ of size $\mathcal{O}(\log B)$ (more precisely, of size at most $160 \ln B$) such that $\rho^*(\mathcal{H}'_i) \leq 4$. The proof is identical to the proof of Claim 18 of [20] (see also Proposition 6.3 of [30]) and we omit it here.

Then we consider all hypergraphs $\mathcal{H}$ with at most B vertices and at most $160 \ln B$ hyperedges. The total number of these hypergraphs is $2^{\mathcal{O}(B \log B)}$. Then, in the same way as above, for each $\mathcal{H}$ and every $\mathbf{s} \in \mathcal{S}$, we use the algorithm of Marx from Proposition 2 to enumerate every $D \subseteq V(\mathcal{G}_{\mathbf{s}})$ where $\mathcal{H}$ appears in $\mathcal{G}_D$ as subhypergraph. For every D, the vectors that differ from $\mathbf{s}$ in the indices from D are enumerated by brute force and each vector is added to $\mathcal{M}(\mathbf{A}, B)$ unless it is already included in the set. The total running time is $2^{\mathcal{O}(B \log B)} \cdot |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ and the number of vectors in $\mathcal{M}(\mathbf{A}, B)$ is $2^{\mathcal{O}(B \log B)} \cdot |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$. □

3.2. Algorithm

Let $(\mathbf{A}, \Sigma, k, B, p, q)$ be an instance of Capacitated Clustering with $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. First, we compute the partition $\mathcal{J} = \{J_1, \ldots, J_s\}$ of $\{1, \ldots, n\}$ into initial clusters.

3.2.1. Choosing potential medians

By the next step, we restrict the set of considered medians. For this, we apply Lemma 4 and construct the set $\mathcal{M} = \mathcal{M}(\mathbf{A}, B)$ of potential medians. Recall that this set has size $2^{\mathcal{O}(B \log B)} |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ and can be computed in $2^{\mathcal{O}(B \log B)} |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time. For a k-clustering $\mathcal{I} = \{I_1, \ldots, I_k\}$, we define the *minimum cost (with respect to $\mathcal{M}$)*, as

$$\min\{\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \mid \mathbf{c}_1, \ldots, \mathbf{c}_k \in \mathcal{M}\}.$$

If $(\mathbf{A}, \Sigma, k, B, p, q)$ is a yes-instance, then it has a solution such that the medians are in $\mathcal{M}$ by Lemma 4. Therefore, solving the problem is equivalent to finding a clustering of minimum cost at most B with respect to $\mathcal{M}$. Throughout this section, whenever we say that $\mathcal{I}$ is a clustering of minimum cost, we mean that the cost is minimum with respect to $\mathcal{M}$.

3.2.2. Structure of solutions

Further, we argue that we can consider solutions of a special structure whose nontrivial part involves bounded number of initial clusters.

By Lemma 3, if $(\mathbf{A}, \Sigma, k, B, p, q)$ is a yes-instance, then there is a solution $\mathcal{I} = \{I_1, \ldots, I_k\}$ to the instance such that the intersection graph $G(\mathcal{I}, \mathcal{J})$ of the initial clusters and the clusters of the solution is a forest. We call such a solution (or k-clustering) *acyclic*. To solve the problem, we check whether the considered instance has an acyclic solution. To simplify notation, we assume that all solutions considered further on are acyclic.

By Observation 3, any k-clustering for $\mathbf{A}$ of cost at most B has at most B composite clusters. We consecutively consider $t = 0, \ldots, \min\{B, k\}$, and for each t, we verify whether there is a solution $\mathcal{I} = \{I_1, \ldots, I_k\}$ with exactly t composite clusters. If we find such a solution, then we return the yes-answer and stop. Otherwise, if we have no solution for all the values of t, we report that $(\mathbf{A}, \Sigma, k, B, p, q)$ is a no-instance. From now on, we assume that nonnegative $t \leq \min\{B, k\}$ is fixed.

It is convenient to consider the special case $t = 0$ separately. If $t = 0$, then a solution $\mathcal{I}$ has no composite cluster, that is, the clusters of the solution form partitions of the initial clusters. Observe that $\mathrm{cost}(\mathcal{I}) = 0 \leq B$ in this case. By Observation 4, the initial clusters can be partitioned into k blocks of size at least p and at most q, if and only if there are positive integers

$h_1, \ldots, h_s$ such that $k = h_1 + \cdots + h_s$ and $\left\lceil \frac{|J_i|}{q} \right\rceil \leq h_i \leq \left\lfloor \frac{|J_i|}{p} \right\rfloor$ for every $i \in \{1, \ldots, s\}$. For every $i \in \{1, \ldots, s\}$, we verify whether $\left\lceil \frac{|J_i|}{q} \right\rceil \leq \left\lfloor \frac{|J_i|}{p} \right\rfloor$. If at least one of the inequalities does not hold, the required $h_1, \ldots, h_s$ do not exist. Otherwise, we observe that positive integers $h_1, \ldots, h_s$ such that $k = h_1 + \cdots + h_s$ and $\left\lceil \frac{|J_i|}{q} \right\rceil \leq h_i \leq \left\lfloor \frac{|J_i|}{p} \right\rfloor$ for every $i \in \{1, \ldots, s\}$ exist if and only if $\sum_{i=1}^{s} \left\lceil \frac{|J_i|}{q} \right\rceil \leq k \leq \sum_{i=1}^{s} \left\lfloor \frac{|J_i|}{p} \right\rfloor$. Then we verify the last inequality.

From now, we assume that $t \geq 1$. Note that we also can assume that $B \geq 1$ because for $B = 0$, no cluster of a solution can be composite.

By Observation 3, there are at most $2B$ initial clusters with nonempty intersections with the composite clusters of a solution $\mathcal{I}$. Since $G(\mathcal{I}, \mathcal{J})$ is a forest, it is easy to observe that at least $t + 1$ initial clusters have nonempty intersections with the composite clusters. We consider $\ell = t + 1, \ldots, 2B$, and for each ℓ, we check whether there is a solution $\mathcal{I} = \{I_1, \ldots, I_k\}$ such that exactly ℓ initial clusters have nonempty intersections with the composite clusters of $\mathcal{I}$. If we find such a solution, then we return the yes-answer and stop. Otherwise, if we have no solution for all the values of ℓ, we report that $(\mathbf{A}, \Sigma, k, B, p, q)$ is a no-instance. From now, we assume that positive $t + 1 \leq \ell \leq 2B$ is given.

Recall that we are looking for an acyclic solution $\mathcal{I} = \{I_1, \ldots, I_k\}$, that is, $G(\mathcal{I}, \mathcal{J})$ is required to be a forest. Let $\mathcal{I}$ be such a k-clustering. Let $\mathcal{I}' \subseteq \mathcal{I}$ be the set of composite clusters and let $\mathcal{J}' \subseteq \mathcal{J}$ be the set of initial clusters having nonempty intersections with the composite clusters. Recall that $|\mathcal{I}'| = t$ and $|\mathcal{J}'| = \ell$ by our assumptions. Note also that the leaves of $G(\mathcal{I}', \mathcal{J}')$ are initial clusters and every connected component of this forest contains at least three vertices.

We consider all forests F on $t + \ell$ vertices such that (i) each connected component of F has at least three vertices, and (ii) F admits a bipartition (U, W) of its vertex set with $|U| = t$ and $|W| = \ell$ such that the leaves of F are in W. Since $t \leq B$ and $\ell \leq 2B$, the number of such forests is $2^{\mathcal{O}(B)}$ [35] and they can be listed in $2^{\mathcal{O}(B)}$ time (see, e.g., [38]). Note that since the leaves are required to be in W, the bipartition (U, W) is unique. From now on, we assume that F together with the bipartition (U, W) is given.

3.2.3. Colorful solutions

Recall that we are looking for a solution such that exactly ℓ initial clusters have nonempty intersections with composite clusters of the solution. We use the *color coding* technique of Alon, Yuster, and Zwick [2] (see [13, Chapter 5] for the detailed introduction) to highlight the initial clusters with nonempty intersections with clusters of a potential solution. We first give a Monte Carlo algorithm with false negatives and then explain how to derandomize it. We color the initial clusters by ℓ colors uniformly at random. We say that a k-clustering $\mathcal{I} = \{I_1, \ldots, I_k\}$ of cost at most B is a *colorful* solution if the initial clusters with nonempty intersections with the clusters of $\mathcal{I}$ have distinct colors. As it is standard for color coding, the algorithm exploits the property that if there is a solution such that exactly ℓ initial clusters have nonempty intersections with the composite clusters of the solution, then the probability that these ℓ clusters get distinct colors in a random coloring is at least $\frac{\ell!}{\ell^\ell} \geq e^{-\ell} \geq e^{-2B}$. Therefore, with probability at least e^{-2B}, a yes-instance admits a colorful solution.

3.2.4. Finding colorful solutions

Our next task is to explain how to check whether there is a colorful solution for a given random coloring $\psi : \mathcal{J} \to \{1, \ldots, \ell\}$ such that $G(\mathcal{I}', \mathcal{J}')$, where $\mathcal{I}'$ is the set of composite clusters in the solution and $\mathcal{J}' \subseteq \mathcal{J}$ is the set of initial clusters having nonempty intersections with the composite clusters, is isomorphic to F. For this, we use dynamic programming over F. Recall that F is given together with the bipartition (U, W) of its vertex set, where the leaves are in W. To construct our dynamic programming algorithm, we formally define k-clusterings forming solutions as follows.

Definition 1 *(Feasible k-clustering).* For a given forest F with the bipartition (U, W) of its vertex set, we say that an acyclic k-clustering $\{I_1, \ldots, I_k\}$ for $\mathbf{A}$ is a *feasible* (with respect to F and the parameters t and ℓ) if the following holds:

(i) $p \leq |I_i| \leq q$ for $i \in \{1, \ldots, k\}$,
(ii) the set $\mathcal{I}' \subseteq \mathcal{I}$ of composite clusters has size t and the set $\mathcal{J}' \subseteq \mathcal{J}$ of initial clusters having nonempty intersections with the composite clusters has size ℓ,
(iii) the initial clusters in $\mathcal{J}'$ are colored by distinct colors by ψ, and
(iv) $G(\mathcal{I}', \mathcal{J}')$ is isomorphic to F with an isomorphism that bijectively maps $\mathcal{I}'$ to U and $\mathcal{J}'$ to W.

Then the problem of finding a colorful solution boils down to checking whether there is a feasible k-clustering of cost at most B.

To proceed with the algorithm, we need some auxiliary notation. For a set of colors $X \subseteq \{1, \ldots, \ell\}$, we use $\mathcal{J}(X) \subseteq \mathcal{J}$ to denote the subset of initial clusters with the colors from X and $C(X) \subseteq \{1, \ldots, n\}$ is used to denote the set of indices in the initial clusters with their colors in X, that is, $C(X) = \cup_{J \in \mathcal{J}(X)} J$. We also denote $\mathbf{A}(X) = \mathbf{A}[\{1, \ldots, m\}, C(X)]$, that is, $\mathbf{A}(X)$ is the submatrix of $\mathbf{A}$ with the columns $\mathbf{a}_i$ such that $i \in C(X)$.

It is common to do bottom-up dynamic programming over rooted trees. However, F may be disconnected. We argue that, given partial solutions for the connected components of F, we can combine them and solve the problem for F. Denote by $F_1, \ldots, F_f$ the connected components of F. Let $U_i = V(F_i) \cap U$ and $W_i = V(F_i) \cap W$ for $i \in \{1, \ldots, f\}$. Let also $t_i = |U_i|$ and $\ell_i = |W_i|$ for $i \in \{1, \ldots, f\}$.

For $i \in \{1, \ldots, f\}$, $X \subseteq \{1, \ldots \ell\}$ and a positive integer $h \leq k$, denote by $\omega_i(X, h)$ the minimum cost of an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to F_i and the parameters t_i and ℓ_i if $|X| = \ell_i$. We assume that $\omega_i(X, h) = +\infty$ if $|X| \neq \ell_i$ or no h-clustering is feasible. Thus, the functions $\omega_i(X, h)$ represent partial solutions for $F_1, \ldots, F_f$.

We show that if we are given the tables of values of $\omega_i(X, h)$, then we can verify whether there is a feasible k-clustering of cost at most B.

Lemma 5. *Given the values $\omega_i(X, h)$ for all $i \in \{1, \ldots, f\}$, $X \subseteq \{1, \ldots \ell\}$ and positive integers $h \leq k$, it can be decided in time $2^{\mathcal{O}(B)} \cdot n^2$ whether there is a feasible k-clustering for $\mathbf{A}$ of cost at most B with respect to F, t and ℓ.*

Proof. To give the intuition behind the proof, observe that a feasible k-clustering of cost at most B with respect to F, t and ℓ exists if and only if there are positive integers $h_1, \ldots, h_f$ such that $h_1 + \cdots + h_f = k$ and a partition $\{X_1, \ldots, X_f\}$ of $\{1, \ldots, \ell\}$ such that

$$\omega_1(X_1, h_1) + \cdots + \omega_f(X_f, h_f) \leq B$$

because in a feasible clustering the initial clusters in $\mathcal{J}'$ are colored by distinct colors. This leads to the following dynamic programming algorithm

For $j \in \{1, \ldots, f\}$, $X \subseteq \{1, \ldots, \ell\}$, let $F^{(j)}$ be the disjoint union of $F_1, \ldots, F_j$, $t^{(j)} = t_1 + \cdots + t_j$ and $\ell^{(j)} = \ell_1 + \cdots + \ell_j$. For $j \in \{1, \ldots, f\}$, $X \subseteq \{1, \ldots, \ell\}$ and positive integer h, denote by $w^{(j)}(X, h)$ the minimum cost of an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to $F^{(j)}$, $t^{(j)}$ and $\ell^{(j)}$ if $|X| = \ell^{(j)}$; we also assume that $w^{(j)}(X, h) = +\infty$ if $|X| \neq \ell^{(j)}$ or there is no feasible h-clustering. Notice that $\omega_1(X, h) = w^{(1)}(X, h)$ and $w^{(f)}(X, h)$ is the minimum cost of an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to F, t and ℓ. Thus, $w^{(f)}(\{1, \ldots, \ell\}, r) \leq B$ if and only if there is a feasible k-clustering for $\mathbf{A}$ of cost at most B with respect to F, t and ℓ.

We compute the values of $w^{(j)}(X, h)$ for $j = 1, 2, \ldots, f$. As we observed, $w^{(1)}(X, h) = \omega_1(X, h)$. To compute $w^{(j)}(X, h)$ for $j \geq 2$, we use the following recurrence:

$$w^{(j)}(X, h) = \min\{\omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h') \mid 1 \leq h' < h \text{ and } \emptyset \neq Y \subset X\}; \tag{4}$$

we also assume that $w^{(j)}(X, h) = +\infty$ if the set in the right part of (4) is empty.

The correctness of (4) is proved in the standard way by showing the two opposite inequalities. Let $X \subseteq \{1, \ldots, \ell\}$. To simplify notation, assume that $\{J_1, \ldots, J_{s'}\}$ are initial clusters with colors from X. Let also $h \leq k$ be a positive integer.

Suppose that $|X| = \ell^{(j)}$ and $\{I_1, \ldots, I_h\}$ is an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to $F^{(j)}$, $t^{(j)}$ and $\ell^{(j)}$ of minimum cost. Let $\mathcal{I}' \subseteq \{I_1, \ldots, I_h\}$ be the set of composite clusters and let $\mathcal{J}' \subseteq \{J_1, \ldots, J_{s'}\}$ be the set of initial clusters having nonempty intersections with the composite clusters. Recall that $|\mathcal{I}'| = t^{(j)}$, $|\mathcal{J}'| = \ell^{(j)}$, and the initial clusters in $\mathcal{J}'$ are colored by distinct colors. Consider an isomorphism α that bijectively maps the vertices of $G(\mathcal{I}', \mathcal{J}')$ to the vertices of F with the property that the vertices of $\mathcal{I}'$ are mapped to $\bigcup_{i=1}^{(j)} U_i$ and $\mathcal{J}'$ are mapped to $\bigcup_{i=1}^{(j)} W_i$. Then ℓ_j clusters of $\mathcal{J}'$ are mapped to W_j. Denote by $Y \subset X$ the set of their colors. Clearly, $|Y| = \ell_j$ and $|X \setminus Y| = \ell^{(j)} - \ell_j = \ell^{j-1}$. Notice that the clusters of $\mathcal{I}'$ that are mapped to U_j are composed of elements of initial clusters with colors from Y and no other composite cluster contains an element of an initial cluster with a color from Y. To simplify notation, assume that the clusters $I_1, \ldots, I_{h'}$ contain elements of the initial clusters with the colors from Y and $I_{h'+1}, \ldots, I_h$ are the clusters containing elements of the initial clusters with the colors from $X \subseteq Y$. Then we have that $\{I_1, \ldots, I_{h'}\}$ is a feasible h'-clustering for $\mathbf{A}(Y)$ with respect to F_j, t_j and ℓ_j. Similarly, we obtain that $\{I_{h'+1}, \ldots, I_h\}$ is a feasible $(h - h')$-clustering for $\mathbf{A}(X \setminus Y)$ with respect to $F^{(j-1)}$, $t^{(j-1)}$ and $\ell^{(j-1)}$. Thus, $w^{(j)}(X, h) \geq \omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h')$ and, therefore,

$$w^{(j)}(X, h) \geq \min\{\omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h') \mid 1 \leq h' < h \text{ and } \emptyset \neq Y \subset X\}. \tag{5}$$

If either $|X| \neq \ell^{(j)}$ or there is no an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to $F^{(j)}$, $t^{(j)}$ and $\ell^{(j)}$, then $w^{(j)}(X, h) = +\infty$ and (5) is trivial.

To show the opposite inequality, let nonempty $Y \subseteq X$ and positive $h' < h$ be such that the right part of (4) is minimum. If $\omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h') = +\infty$, then the required inequality holds trivially. Assume that this is not the case. Then $|Y| = \ell_j$, $|X \setminus Y| = \ell^{(j-1)}$, there is an h'-clustering $\mathcal{I}^{(1)}$ for $\mathbf{A}(Y)$ of cost $\omega_j(Y, h')$ that is feasible with respect to F_j, t_j and ℓ_j, and there is an $(h - h')$-clustering $\mathcal{I}^{(2)}$ for $\mathbf{A}(X \setminus Y)$ of cost $w^{(j-1)}(X \setminus Y, h - h')$ that is feasible with respect to $F^{(j-1)}$, $t^{(j-1)}$ and $\ell^{(j-1)}$. Consider $\mathcal{I} = \mathcal{I}^{(1)} \cup \mathcal{I}^{(2)}$ and observe that this is an h-clustering for $\mathbf{A}(X)$ that is feasible with respect to $F^{(j)}$, $t^{(j)}$ and $\ell^{(j)}$. This means that $w^{(j)}(X, h) \leq \omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h')$. By the choice of Y and h',

$$w^{(j)}(X, h) \leq \min\{\omega_j(Y, h') + w^{(j-1)}(X \setminus Y, h - h') \mid 1 \leq h' < h \text{ and } \emptyset \neq Y \subset X\}. \tag{6}$$

Combining (5) and (6), we obtain that the recurrence (4) holds.

Finally, we compute $w^{(f)}(X, h)$ for all $X \subseteq \{1, \ldots, \ell\}$ and all positive $h \leq r$. In particular, we find $w^{(f)}(\{1, \ldots, \ell\}, k)$ and verify whether this value is at most B.

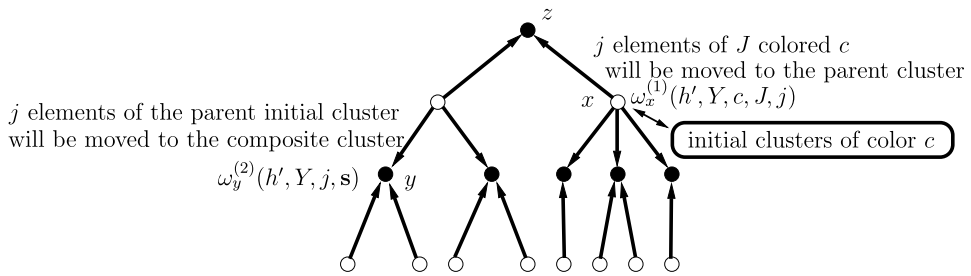

Fig. 2. The general scheme of dynamic programming over T. The vertices of U corresponding to composite clusters are shown by black bullets and the vertices of W corresponding to initial clusters are white. The arrows show which initial clusters are contributing to composite clusters. Note that J is a (part of) initial cluster of color c and the remaining initial clusters of color c (including the rest of the cluster containing J) are split into simple clusters.

To evaluate the running time, note that to compute the table of values of $w^{(j)}(X,h)$ by (4), we consider all nonempty X of size at most ℓ and the nonempty subsets $Y \subset X$. This means that we consider at most 3^ℓ pairs of sets. Also, we consider all positive $h \leq k$ and $h' \leq h$, that is, at most k^2 pairs of integers. Since $\ell \leq 2B$ and $k \leq n$, the computations can be done in $2^{\mathcal{O}(B)} \cdot n^2$ time. Since $f \leq t \leq B$, the total running time is $2^{\mathcal{O}(B)} \cdot n^2$. □

The final step is to compute the partial solutions for $F_1, \ldots, F_f$. By Lemma 5, we have to compute the tables of values of $\omega_i(X,h)$ for all $i \in \{1, \ldots, f\}$, nonempty $X \subseteq \{1, \ldots, \ell\}$ and positive $h \leq k$. For this, we use the fact that $F_1, \ldots, F_f$ are trees and this allows us to use dynamic programming over these trees.

Lemma 6. *Let T be a tree with a bipartition (U, W) of its vertex set such that $t' = |U| \leq t$, $\ell' = |W| \leq \ell$ and the leaves of T are in W. For a given $X \subseteq \{1, \ldots, \ell\}$ with $|X| = \ell'$ and positive $h \leq k$, the minimum cost of a feasible h-clustering for $\mathbf{A}(X)$ with respect to T, t' and ℓ' can be found in $2^{\mathcal{O}(B \log B)} |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time.*

Proof. We select a vertex $z \in U$ as a root of T. This selection defines a parent-child relation on the set of vertices. For a vertex $x \in V(T)$, we denote by T_x the subtree of T induced by the descendants of x (including the vertex itself). For $x \in V(T)$, let $t_x = V(T_x) \cap U$ and $\ell_x = V(T_x) \cap W$. For every $x \in V(T)$, we compute the tables of auxiliary values depending on whether $x \in U$ or $x \in W$.

For a set of colors $Z \subseteq X$, $J \in \mathcal{J}(Z)$ and $J' \subseteq J$, we use $\mathcal{J}(Z)/J'$ to denote the set of clusters obtained from the initial clusters of $\mathcal{J}(Z)$ by the replacement of J by $J'' = J \setminus J'$ if $J' \subset J$ and $\mathcal{J}(Z)/J' = \mathcal{J}(Z) \setminus \{J\}$ if $J' = J$. We assume that the clusters of $\mathcal{J}(Z)/J'$ have the inherited colors. We also write $\mathbf{A}(Z)/J'$ to denote the submatrix of $\mathbf{A}(Z)$ obtained by the deletion of the columns with the indices from J'. Note that $\mathcal{J}(Z)/J'$ is the set of initial clusters for $\mathbf{A}(Z)/J'$.

Suppose that $x \in W$. For every positive integer $h' \leq h$, every $Y \subseteq X$, every $c \in Y$, every $J \in \mathcal{J}(Y)$ and every nonnegative integer $j \leq |J|$, we define $\omega_x^{(1)}(h', Y, c, J, j)$. For technical reasons, it is convenient to define this function for leaves separately.

Definition 2 *(Partial solution for a leaf $x \in W$).* Let x be a leaf. We define $\omega_x^{(1)}(h', \{c\}, c, J, j)$ as the minimum cost of an h'-clustering for $\mathbf{A}(Y)/J'$, where $J' \subseteq J$ of size j, such that all the clusters are simple, and $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ if $Y \neq \{c\}$.

Definition 3 *(Partial solution for an internal $x \in W$).* If x is an internal vertex of T, then $\omega_x^{(1)}(h', Y, c, J, j)$ is the minimum cost of an h'-clustering $\mathcal{I} = \{I_1, \ldots, I_{h'}\}$ for $\mathbf{A}(Y)/J'$, where $J' \subset J$ of size j, such that

(i) $p \leq |I_i| \leq q$ for $i \in \{1, \ldots, h'\}$,
(ii) the set $\mathcal{I}' \subseteq \mathcal{I}$ of composite clusters has size t_x, and the set $\mathcal{J}' \subseteq \mathcal{J}(Y)/J'$ of initial clusters having nonempty intersections with the composite clusters has size ℓ_x,
(iii) $|Y| = \ell_x$ and the initial clusters in $\mathcal{J}'$ are colored by distinct colors by ψ,
(iv) $G(\mathcal{I}', \mathcal{J}')$ is isomorphic to T_x with an isomorphism α that bijectively maps $\mathcal{I}'$ to U_x, $\mathcal{J}'$ to W_x, and
(v) $J \setminus J' \in \mathcal{J}'$, $\alpha(J \setminus J') = x$ and $\psi(J \setminus J') = c$.

In both cases, we assume that $\omega_x^{(1)}(h, Y, c, J, j) = +\infty$ if there is no such an h'-clustering.

Informally, $\omega_x^{(1)}(h', Y, c, J, j)$ is the minimum cost of an h'-clustering for $\mathbf{A}(Y)/J'$ that is feasible with respect to T_x, t_x and ℓ_x with the additional assumption that we take j elements of J colored by c to include to the composite cluster that corresponds to the parent of x (see Fig. 2). Observe that the value of $\omega_x^{(1)}(h', Y, c, J, j)$ does not depend on the choice of J'. Notice also that we have the special case when $U_x = \emptyset$, i.e. when x is a leaf because we have no composite clusters in this case. Then we form h' simple clusters from the initial clusters $\mathcal{J}(Y)/J'$.

Now we define the function $\omega_x^{(2)}(h', Y, j, \mathbf{s})$ for $x \in U$ for every positive integer $h' \leq h$, every $Y \subseteq X$, every nonnegative integer $j \leq q$, and every $\mathbf{s} \in \mathcal{M}$.

Definition 4 *(Partial solution for $x \in U$).* $\omega_x^{(2)}(h', Y, j, \mathbf{s})$ is the minimum cost of an h'-clustering $\mathcal{I} = \{I_1, \ldots, I_{h'}\}$ for $\mathbf{A}(Y)$ such that

(i) the cost of I_1 is computed with respect to the median $\mathbf{s}$, that is, the cost equals $\sum_{i \in I_1} d_H(\mathbf{s}, \mathbf{a}_i)$,
(ii) $p - j \leq |I_1| \leq q - j$ and $p \leq |I_i| \leq q$ for $i \in \{2, \ldots, h'\}$,
(iii) for the set of composite clusters $\mathcal{I}' \subseteq \mathcal{I}$, $\mathcal{I}'' = \mathcal{I}' \cup \{I_1\}$ has size t_x, and the set $\mathcal{J}' \subseteq \mathcal{J}(Y)$ of initial clusters having nonempty intersections with the clusters from $\mathcal{I}''$ has size ℓ_x,
(iv) $|Y| = \ell_x$ and the initial clusters in $\mathcal{J}'$ are colored by distinct colors by ψ,
(v) $G(\mathcal{I}'', \mathcal{J}')$ is isomorphic to T_x with an isomorphism α that bijectively maps $\mathcal{I}''$ to U_x, $\mathcal{J}'$ to W_x, and $\alpha(I_1) = x$.

In the same way as above for other functions, it is assumed that $\omega_x^{(2)}(h', Y, j, \mathbf{s}) = +\infty$ if there is no such an h'-clustering.

Informally, $\omega_x^{(2)}(h', Y, j, \mathbf{s})$ is the minimum cost of an h'-clustering for $\mathbf{A}(Y)$ that is feasible with respect to T_x, t_x and ℓ_x, where the specific cluster I_1 associated with x is required to have $\mathbf{s}$ as its median and "misses" j elements (see Fig. 2). Notice that it is not required that $\mathbf{s}$ is optimal for I_1. However, in the future, I_1 is going to be complemented by j elements of an initial cluster corresponding to the parent of x, unless x is a root. Note also that I_1 is not a composite cluster if x has a unique child, but because I_1 is expected to be complemented by other elements, I_1 is counted as a composite cluster in the definition of $\omega_x^{(2)}(h', Y, j, \mathbf{s})$.

Now we explain how to compute the table of values of $\omega_x^{(1)}(h', Y, c, J, j)$ and $\omega_x^{(2)}(h', Y, j, \mathbf{s})$. First, we compute $\omega_x^{(1)}(h', Y, c, J, j)$ for leaves.

Claim 3.1. *For every leaf x of T, $\omega_x^{(1)}(h', Y, c, J, j)$ can be computed in $\mathcal{O}(n)$ time.*

Proof. If $Y \neq \{c\}$, $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ by the definition. Assume that $Y = \{c\}$. Let $J' \subseteq J$ be a set of size j. We compute $\hat{\mathcal{J}} = \mathcal{J}(Y)/J'$ in $\mathcal{O}(n)$ time. Then $\omega_x^{(1)}(h', Y, c, J, j) = 0$ if every set in $\hat{\mathcal{J}}$ can be partitioned into clusters of size at least p and at most q in such a way that the total number of clusters is h', and $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ otherwise. We apply Observation 4. First, we verify whether every $\hat{J} \in \hat{\mathcal{J}}$ can be partitioned into clusters of size at least p and at most q by checking whether $\left\lceil \frac{|\hat{J}|}{q} \right\rceil \leq \left\lfloor \frac{|\hat{J}|}{p} \right\rfloor$. If this holds, then we observe that we can obtain exactly h' clusters in total if and only if $\sum_{\hat{J} \in \hat{\mathcal{J}}} \left\lceil \frac{|\hat{J}|}{q} \right\rceil \leq h' \leq \sum_{\hat{J} \in \hat{\mathcal{J}}} \left\lfloor \frac{|\hat{J}|}{p} \right\rfloor$. Since checking of these conditions can be done in $\mathcal{O}(n)$ time, the total running time is $\mathcal{O}(n)$. □

Next, we explain how to compute $\omega_x^{(1)}(h', Y, c, J, j)$ for internal vertices if the tables of values of $\omega_y^{(2)}(\cdot, \cdot, \cdot, \cdot)$ are given for all children y of x. This is done by an auxiliary dynamic programming algorithm.

Claim 3.2. *Let $x \in W$ be an internal vertex of T and assume that the table of values of $\omega_y^{(2)}(\cdot, \cdot, \cdot, \cdot)$ is computed for every child y of x. Then $\omega_x^{(1)}(h', Y, c, J, j)$ can be computed in $2^{\mathcal{O}(B \log B)} |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time.*

Proof. Let $h' \leq h$, $Y \subseteq X$, $c \in Y$, $J \in \mathcal{J}(Y)$ and $j \leq |J|$. If $j = |J|$, then we immediately set $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ because we have no proper $J' \subset J$ of size j. Also, if $|Y| \neq \ell_x$ or $\psi(J) \neq c$, then $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ by definition. Assume that $j < |J|$, $J' \subset J$ of size j, $\psi(J) = c$ and $|Y| = \ell_x$. Let $\hat{J} = J \setminus J'$. We denote by $y_1, \ldots, y_f$ the children of x in T.

Consider the initial clusters of color c. By the definition of $\omega_x^{(1)}(h', Y, c, J, j)$, we are interested in an h'-clustering, where the initial clusters of color c distinct from J are split into simple clusters and, possibly, some parts of J also form simple clusters. For nonnegative integers $\hat{h} \leq h'$ and $\hat{j} \leq |\hat{J}|$, we define $w(\hat{h}, \hat{j})$ to be 0 if the initial clusters of $\hat{\mathcal{J}} = \mathcal{J}(\{c\})/(J \setminus J'')$, where $J'' \subseteq \hat{J}$ of size $\hat{j}$ can be partitioned into $\hat{h}$ simple clusters of size at least p and at most q, and we set $w(\hat{h}, \hat{j}) = +\infty$ otherwise. To compute $w(\hat{h}, \hat{j})$, we use Observation 4 similarly to the proof of Claim 3.1. Namely, we verify whether every $\tilde{J} \in \hat{\mathcal{J}}$ can be partitioned into clusters of size at least p and at most q by checking whether $\left\lceil \frac{|\tilde{J}|}{q} \right\rceil \leq \left\lfloor \frac{|\tilde{J}|}{p} \right\rfloor$, and then we check whether $\sum_{\tilde{J} \in \hat{\mathcal{J}}} \left\lceil \frac{|\tilde{J}|}{q} \right\rceil \leq \hat{h} \leq \sum_{\tilde{J} \in \hat{\mathcal{J}}} \left\lfloor \frac{|\tilde{J}|}{p} \right\rfloor$. Since $\hat{h} \leq h' \leq h$, the values of $w(\hat{h})$ can be computed in $\mathcal{O}(n^2)$ time.

Observe that by the definition of $\omega_x^{(1)}(h', Y, c, J, j)$, the elements of $\hat{J}$ should be included in f composite clusters associated with the children of x in an h'-clustering for $\mathbf{A}(Y)/J'$. In particular, if $|\hat{J}| < f$, it cannot be done and $\omega_x^{(1)}(h', Y, c, J, j) = +\infty$ by the definition. From now, we assume that $|\hat{J}| \geq f$.

For $i \in \{1, \ldots, f\}$, denote by $T^{(i)}$ the subtree of T induced by $\{x\} \cup \bigcup_{i'=1}^{i} V(T_{y_{i'}})$, set $U^{(i)} = U \cap V(T^{(i)})$ and $W^{(i)} = W \cap V(T^{(i)})$. Let also $t^{(i)} = |U^{(i)}|$ and $\ell^{(i)} = |W^{(i)}|$ for $i \in \{1, \ldots, f\}$. For each $i \in \{1, \ldots, f\}$, each nonnegative $\hat{h} \leq h'$, each positive $\hat{j} \leq |J| - j$, and every $c \in Z \subseteq Y$, define the auxiliary values $w^{(i)}(\hat{h}, \hat{j}, Z)$.

Definition 5 *(Auxiliary partial solution).* $w^{(i)}(\hat{h},\hat{j},Z)$ is the minimum cost of $\hat{h}$-clustering $\mathcal{I}=\{I_1,\ldots,I_{\hat{h}}\}$ for $\mathbf{A}(Z)/(J\setminus J'')$, where $J''\subseteq\hat{J}$ of size $\hat{j}$, such that

(i) $p\leq|I_{i'}|\leq q$ for $i'\in\{1,\ldots,\hat{h}\}$,
(ii) the set $\mathcal{I}'\subseteq\mathcal{I}$ of composite clusters has size $t^{(i)}$, and the set $\mathcal{J}'\subseteq\mathcal{J}(Y)/(J\setminus J'')$ of initial clusters having nonempty intersections with the composite clusters has size $\ell^{(i)}$,
(iii) $|Z|=\ell^{(i)}$ and the initial clusters in $\mathcal{J}'$ are colored by distinct colors by ψ,
(iv) $G(\mathcal{I}',\mathcal{J}')$ is isomorphic to $T^{(i)}$ with an isomorphism α that bijectively maps $\mathcal{I}'$ to $U^{(i)}$, $\mathcal{J}'$ to $W^{(i)}$, and
(v) $J''\in\mathcal{J}'$, $\alpha(J'')=x$ and $\psi(J'')=c$.

We also follow the same convention as above that $w^{(i)}(\hat{h},\hat{j},Z)=+\infty$ if either there is no $\hat{h}$-clustering satisfying (i)–(v). Observe that, by the definition, $\omega_x^{(1)}(h',Y,c,J,j)=w^{(f)}(h',|J|-j,Y)$. Therefore, we compute the tables of values of $w^{(i)}(\cdot,\cdot,\cdot)$ for $i=1,\ldots,f$.

To initiate the computation of $w^{(i)}(\cdot,\cdot,\cdot)$, it is convenient to formally define this function for $i=0$. We set

$$w^{(0)}(\hat{h},\hat{j},Z)=\begin{cases}w(\hat{h},\hat{j}) & \text{if } Z=\{c\},\\ +\infty & \text{otherwise.}\end{cases}$$

For $\mathbf{s}\in M$, denote $d(\mathbf{s})=d_H(\mathbf{s},\mathbf{a}_j)$ for $j\in J$. Then to compute $w^{(i)}(\hat{h},\hat{j},Z)$ for $i\geq 1$, we use the following recurrence:

$$w^{(i)}(\hat{h},\hat{j},Z)=\min\{\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\},\tag{7}$$

where the minimum in the right part is taken over all integers $1\leq\hat{h}'\leq\hat{h}$ and $0<\hat{j}'\leq\hat{j}$, all sets $\hat{Z}$ such that $c\notin\hat{Z}\subset Z$, and all $\mathbf{s}\in\mathcal{M}$. We assume that $w^{(i)}(\hat{h},\hat{j},Z)=+\infty$ if the set in the right part is empty.

We prove the correctness of (7) by showing the opposite inequalities between the left and the right part.

If $w^{(i)}(\hat{h},\hat{j},Z)=+\infty$, then

$$w^{(i)}(\hat{h},\hat{j},Z)\geq\min\{\omega^{(2)}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}.$$

Suppose that $w^{(i)}(\hat{h},\hat{j},Z)<+\infty$. Consider $\hat{h}$-clustering $\mathcal{I}$ for $\mathbf{A}(Z)/(J\setminus J'')$ of cost $w^{(i)}(\hat{h},\hat{j},Z)$ satisfying (i)–(v). Let $I\in\mathcal{I}$ be the composite cluster such that $\alpha(I)=y_i$. Since $\alpha(J'')=x$, I contains elements of J''. Let $\hat{J}''=I\cap J''$ and $\hat{j}'=\hat{J}''$. Denote by $\mathbf{s}\in\mathcal{M}$ the median of I_1. Consider $\hat{\mathcal{J}}''=\alpha^{-1}(V(T_{y_i}))\cap\mathcal{J}'$, that is, the set of initial clusters having nonempty intersections with the composite clusters that are mapped by α to the nodes of T_{y_i}. The coloring ψ colors these clusters by distinct colors and we define $\hat{Z}$ to be the set of colors of the clusters of $\hat{\mathcal{J}}''$; note that $c\notin\hat{Z}$. Denote by $\hat{h}'$ the number of clusters in $\mathcal{I}$ containing elements of the initial clusters with colors in $\hat{Z}$ and let $\mathcal{I}_1$ be the set of these clusters; observe that $I\in\mathcal{I}_1$. Let $\mathcal{I}_2=\mathcal{I}\setminus\mathcal{I}_1$.

By the definition of the values of $w^{(i-1)}(\cdot,\cdot,\cdot)$, we obtain that the cost of clustering for $\mathcal{I}_2$ is at least $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$. The cluster I contains $\hat{j}'$ elements of J. Since $\mathbf{s}$ is its median, these $\hat{j}'$ elements contribute $\hat{j}'d(\mathbf{s})$ to its cost. Then, by the definition of $\omega^{(2)}_{y_i}(\cdot,\cdot,\cdot,\cdot)$, we have that the cost of clustering for $\mathcal{I}_1$ is at least $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})$. This means that $w^{(i)}(\hat{h},\hat{j},Z)\geq\omega^{(2)}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',(Z\setminus\hat{Z})\cup\{c\})$ and

$$w^{(i)}(\hat{h},\hat{j},Z)\geq\min\{\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}.\tag{8}$$

For the opposite direction, assume that integers $\hat{h}'$, $\hat{j}'$, a set $\hat{Z}$, and a median $\mathbf{s}$ are chosen in such a way that the value of $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$ is minimum. If the value is $+\infty$, then $w^{(i)}(\hat{h},\hat{j},Z)\leq\omega^{(2)}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$ as required. Assume that $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})<+\infty$ and $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})<+\infty$.

By the definition of $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})$, there is an $\hat{h}'$-clustering $\mathcal{I}_1$ for $\mathbf{A}(\hat{Z})$ of cost $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})$ satisfying conditions (i)–(v) of the definition. In particular, $\mathcal{I}_1$ contains a special cluster I with the median $\mathbf{s}$ such that $p-\hat{j}'\leq|I|\leq q-\hat{j}'$ and I is mapped to the root y_i of T_{y_i} by the isomorphism α.

Let $J''\subseteq\hat{J}$ of size $\hat{j}$ and let $\hat{J}''\subseteq J''$ of size $\hat{j}'$. By the definition of $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$, there is an $(\hat{h}-\hat{h}')$-clustering $\mathcal{I}_2$ for $\mathbf{A}(Z\setminus\hat{Z})/((J\setminus J'')\cup\hat{J}'')$ of cost $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$ satisfying conditions (i)–(v) of the definition of $w^{(i-1)}(\cdot,\cdot,\cdot)$.

Observe that the clusters of $\mathcal{I}_1$ and $\mathcal{I}$ are pairwise disjoint and include all elements of the initial clusters with their colors in Z except $\hat{j}'+j$ elements of J. We construct the $\hat{h}$-clustering $\mathcal{I}$ for $\mathbf{A}(Z)/(J\setminus J'')$ as follows. First, we modify the cluster $I\in\mathcal{I}_1$ by setting $I:=I\cup\hat{J}''$. Note that we increase the cost of the cluster by at most $\hat{j}'d(\mathbf{s})$. Then we take the union of $\mathcal{I}_1$ and $\mathcal{I}_2$. The definitions of the values $\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})$ and $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$ imply that $\mathcal{I}$ satisfies conditions (i)–(v) for $w^{(i)}(\hat{h},\hat{j},Z)$. Therefore, $w^{(i)}(\hat{h},\hat{j},Z)\leq\omega^{(2)}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s})+\hat{j}'d(\mathbf{s})+w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$.

By the choice of $\hat{h}'$, $\hat{j}'$, $\hat{Z}$, and $\mathbf{s}$,

$$w^{(i)}(\hat{h},\hat{j},Z) \leq \min\{\omega^{(2)}_{y_i}(\hat{h}',\hat{Z},\hat{j}',\mathbf{s}) + \hat{j}'d(\mathbf{s}) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}. \tag{9}$$

Then (8) and (9) imply (7).

We use (7) to compute the table of values of $w^{(f)}(\cdot,\cdot,\cdot)$. Then $\omega^{(1)}_x(h',Y,c,J,j) = w^{(f)}(h',|J|-j,Y)$ by the definition.

To evaluate the running time, notice that the initial table $w^{(f)}(\cdot,\cdot,\cdot)$ can be computed in $2^{\mathcal{O}(B)}\cdot n^2$, since $w(\hat{h})$ can be computed in $\mathcal{O}(n^2)$ time and then the table is constructed for at most n values of $\hat{j}$ and at most 2^ℓ sets Z. To compute the table $w^{(i)}(\cdot,\cdot,\cdot)$ from $w^{(i-1)}(\cdot,\cdot,\cdot)$ by (7) for $i\in\{1,\ldots,f\}$, we consider all pairs of integers $\hat{h}'\leq\hat{h}$, all pairs of sets Z and $\hat{Z}\subset Z$ and all $\mathbf{s}\in\mathcal{M}$. Since $\hat{h}\leq n$, $Z\subseteq\{1,\ldots,\ell\}$ and $\ell\leq 2B$, and $|\mathcal{M}| = 2^{\mathcal{O}(B\log B)}|\Sigma|^B\cdot(mn)^{\mathcal{O}(1)}$, $w^{(i)}(\cdot,\cdot,\cdot)$ can be computed in $2^{\mathcal{O}(B\log B)}|\Sigma|^B\cdot(mn)^{\mathcal{O}(1)}$. Then the total running time is $2^{\mathcal{O}(B\log B)}|\Sigma|^B\cdot(mn)^{\mathcal{O}(1)}$. □

Further, we show how to compute $\omega^{(2)}_x(h',Y,j,\mathbf{s})$ if the tables of values of $\omega^{(1)}_y(\cdot,\cdot,\cdot,\cdot,\cdot)$ are already computed. Similarly to the proof of Claim 3.2, we also use an auxiliary dynamic programming algorithm.

Claim 3.3. *Let $x\in U$ be an internal vertex of T and assume that the table of values of $\omega^{(1)}_y(\cdot,\cdot,\cdot,\cdot,\cdot)$ is computed for every child y of x. Then $\omega^{(2)}_x(h',Y,j,\mathbf{s})$ can be computed in $2^{\mathcal{O}(B)}\cdot n^{\mathcal{O}(1)}$ time.*

Proof. Let $h'\leq h$, $Y\subseteq X$, $j\leq q$, and let $\mathbf{s}\in\mathcal{M}$. If $|Y|\neq\ell_x$, then $\omega^{(2)}_x(h',Y,j,\mathbf{s}) = +\infty$ by definition. Assume that $|Y|=\ell_x$. In the same way as in the proof of Claim 3.2, denote by $y_1,\ldots,y_f$ the children of x in T. For $i\in\{1,\ldots,f\}$, let $T^{(i)}$ be the subtree of T induced by $\{x\}\cup\bigcup_{i'=1}^{i}V(T_{y_{i'}})$, set $U^{(i)} = U\cap V(T^{(i)})$ and $W^{(i)} = W\cap V(T^{(i)})$. Let also $t^{(i)} = |U^{(i)}|$ and $\ell^{(i)} = |W^{(i)}|$ for $i\in\{1,\ldots,f\}$. For an initial cluster J, we denote by $d(J) = d_H(\mathbf{s},\mathbf{a}_i)$ for $i\in J$. Similarly to the proof of Claim 3.2, we compute some auxiliary values.

For each $i\in\{1,\ldots,f\}$, every positive integer $\hat{h}\leq h'$, every nonnegative integer $\hat{j}\leq q$, and every nonempty $Z\subseteq X$, we define the auxiliary value as follows.

Definition 6 *(Auxiliary partial solution).* $w^{(i)}(\hat{h},\hat{j},Z)$ is the minimum cost of an $\hat{h}$-clustering $\mathcal{I} = \{I_1,\ldots,I_{\hat{h}}\}$ for $\mathbf{A}(Z)$ such that

(i) the cost of I_1 is computed with respect to the median $\mathbf{s}$, that is, the cost equals $\sum_{i\in I_1} d_H(\mathbf{s},\mathbf{a}_i)$,
(ii) $|I_1| = \hat{j}$ and $p\leq|I_{i'}|\leq q$ for $i'\in\{2,\ldots,\hat{h}\}$,
(iii) for the set of composite clusters $\mathcal{I}'\subseteq\mathcal{I}$, $\mathcal{I}'' = \mathcal{I}'\cup\{I_1\}$ has size $t^{(i)}$, and the set $\mathcal{J}'\subseteq\mathcal{J}(Z)$ of initial clusters having nonempty intersections with the clusters from $\mathcal{I}''$ has size $\ell^{(i)}$,
(iv) $|Z| = \ell^{(i)}$ and the initial clusters in $\mathcal{J}'$ are colored by distinct colors by ψ,
(v) $G(\mathcal{I}'',\mathcal{J}')$ is isomorphic to $T^{(i)}$ with an isomorphism α that bijectively maps $\mathcal{I}''$ to $U^{(i)}$, $\mathcal{J}'$ to $W^{(i)}$, and $\alpha(I_1) = x$.

We assume that $w^{(i)}(\hat{h},\hat{j},Z) = +\infty$ if there is no such a $\hat{h}$-clustering.

Notice that the parameter $\hat{j}$ defines the size of a selected cluster I_1. Then, by the definition, we have that

$$\omega^{(2)}_x(h',Y,j,\mathbf{s}) = \min\{w^{(f)}(h',\hat{j},Y)\mid p-j\leq\hat{j}\leq q-j\} \tag{10}$$

assuming that $\omega^{(2)}_x(h',Y,j,\mathbf{s}) = +\infty$ if the set in the right part is empty.

We compute the tables of values of $w^{(i)}(\cdot,\cdot,\cdot)$ for $i=1,\ldots,f$.

First, we observe that

$$w^{(1)}(\hat{h},\hat{j},Z) = \min\{\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)\mid c\in Z, J\in\mathcal{J}(Z)\}; \tag{11}$$

as before, $w^{(1)}(\hat{h},\hat{j},Z) = +\infty$ if the set in the right part is empty.

To see that $w^{(1)}(\hat{h},\hat{j},Z)\geq\min\{\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)\mid c\in Z, J\in\mathcal{J}(Z)\}$, assume that $w^{(1)}(\hat{h},\hat{j},Z)<+\infty$; otherwise, the inequality is trivial. Let $\mathcal{I} = \{I_1,\ldots,I_{\hat{h}}\}$ be an $\hat{h}$-clustering for $\mathbf{A}(Z)$ satisfying conditions (i)–(v). Since y_1 is the unique child of x in $T^{(1)}$, I_1 consists of $\hat{j}$ elements of some initial cluster J. Let c be the color assigned to J by ψ. Then, by the definition of $\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j})$, $\{I_2,\ldots,I_{\hat{h}}\}$ is an $(\hat{h}-1)$-clustering for $\mathbf{A}(Z)/J'$ for $J'\subseteq J$ of size $\hat{j}$ that satisfies all the condition of the definition of $\omega^{(1)}_{y_1}(\cdot,\cdot,\cdot,\cdots,\cdot)$. Therefore, the cost of $\{I_2,\ldots,I_{\hat{h}}\}$ is an $(\hat{h}-1)$ is at least $\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j})$. The median of I_1 is $\mathbf{s}$ and I_1 contains $\hat{j}$ elements of J. Therefore, the cost of I_1 is $\hat{j}d(J)$. We conclude that $w^{(1)}(\hat{h},\hat{j},Z)\geq\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)$. Therefore, $w^{(1)}(\hat{h},\hat{j},Z)\geq\min\{\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)\mid c\in Z, J\in\mathcal{J}(Z)\}$.

Now we prove that $w^{(1)}(\hat{h},\hat{j},Z)\leq\min\{\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)\mid c\in Z, J\in\mathcal{J}(Z)\}$. If the right part of (11) is $+\infty$, then the inequality is trivial. Assume that this is not the case and let $c\in Z$ and $J\in\mathcal{J}(Z)$ be such that the right part of (11) achieves the minimum value for them. Then there is an $(\hat{h}-1)$-clustering $\mathcal{I}$ for $\mathbf{A}(Z)/J'$, where $J'\subseteq J$ has size $\hat{j}$,

with the cost $\omega^{(1)}_{y_1}(\hat{h}-1, Z, c, J, \hat{j})$ that satisfies all the condition of the definition of $\omega^{(1)}_{y_1}(\cdot,\cdot,\cdot,\cdots,\cdot)$. Then we construct a new cluster $I = J'$ with the median $\mathbf{s}$. Clearly, the cost is $\hat{j}d(J)$. It is straightforward to verify that $\mathcal{I}\cup\{I\}$ satisfies (i)–(v). Therefore, $w^{(1)}(\hat{h},\hat{j},Z) \le \omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J)$ and $w^{(1)}(\hat{h},\hat{j},Z) \le \min\{\omega^{(1)}_{y_1}(\hat{h}-1,Z,c,J,\hat{j}) + \hat{j}d(J) \mid c \in Z,\ J \in \mathcal{J}(Z)\}$.

Combining the two inequalities, we conclude that (11) holds.

To compute $w^{(1)}(\hat{h},\hat{j},Z)$ for $i \ge 2$, we show that

$$w^{(i)}(\hat{h},\hat{j},Z) = \min\{\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}, \tag{12}$$

where the minimum is taken over all positive integers $\hat{h}' < \hat{h}$, $\hat{j}' < \hat{j}$, all nonempty sets $\hat{Z} \subset Z$, all $c \in Z$, and $J \in \mathcal{J}(Z)$. As it is standard in our paper, $w^{(i)}(\hat{h},\hat{j},Z) = +\infty$ if the set in the right part of (12) is empty.

We prove (12) by demonstrating the opposite inequalities between the left and the right part.

If $w^{(i)}(\hat{h},\hat{j},Z) = +\infty$, then $w^{(i)}(\hat{h},\hat{j},Z) \ge \min\{\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}$. Assume that this is not the case. Then there is an $\hat{h}$-clustering $\mathcal{I}$ for $\mathbf{A}(Z)$ of cost $w^{(i)}(\hat{h},\hat{j},Z)$ satisfying (i)–(v). In particular, there is $I \in \mathcal{I}$ such that $|I| = \hat{j}$, $\alpha(I) = x$ and $\mathbf{s}$ is its median. Let $J \in \mathcal{J}(Z)$ be the initial cluster such that $\alpha(J) = y_i$. Denote by c its color. By definition, $J \cap I \ne \emptyset$. Let $J' = I \cap J$ and $\hat{j}' = |J'|$. Consider $\hat{\mathcal{J}}' = \alpha^{-1}(V(T_{y_i})) \cap \mathcal{J}'$, that is, the set of initial clusters intersecting composite clusters that are mapped by α to the vertices of T_{y_i}. Note that $J \in \hat{\mathcal{J}}'$. By definition, these clusters are colored by distinct colors by ψ. Denote by $\hat{Z}$ the set of their colors. Clearly, $c \in \hat{Z}$. Let $\mathcal{I}_1 \subseteq \mathcal{I}\setminus\{I\}$ be the set of clusters in $\mathcal{I}$ having nonempty intersections with the initial clusters from $\hat{\mathcal{J}}'$; note that $I \notin \mathcal{I}_1$ by definition. Set $\hat{h}' = |\mathcal{I}_1|$. Let $\mathcal{I}_2 = \mathcal{I}\setminus\mathcal{I}_1$.

Observe that $\mathcal{I}_1$ is an $\hat{h}'$-clustering for $\mathbf{A}(\hat{Z})/J'$. Moreover, $\mathcal{I}_1$ satisfies all the conditions of the definition of $\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}')$. This implies that the cost of $\mathcal{I}_1$ is at least $\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}')$. Consider the clustering $\hat{\mathcal{I}}_2$ obtained from $\mathcal{I}_2$ by the replacement of I by $\hat{I} = I\setminus J'$. Notice that the clusters of $\hat{\mathcal{I}}_2$ contain only elements of initial clusters with colors from $Z\setminus\hat{Z}$. Also, we have that $|\hat{I}| = \hat{j}-\hat{j}'$ and $|\hat{\mathcal{I}}_2| = \hat{h}-\hat{h}'$ because $i \ge 2$ and $I \ne J'$. Then it is straightforward to verify that $\hat{\mathcal{I}}_2$ is $(\hat{h}-\hat{h}')$-clustering for $\mathbf{A}(Z\setminus\hat{Z})$ satisfying (i)–(v) for $w^{(i-1)}(\cdot,\cdot,\cdot)$. Therefore, the cost of $\hat{\mathcal{I}}_2$ is at least $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$. Finally, recall that $J' \subset I$. Since $\mathbf{s}$ is the median of I, the contribution of J' to the cost is $\hat{j}'d(J)$. We conclude that $w^{(i)}(\hat{h},\hat{j},Z) \ge \omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$. Hence,

$$w^{(i)}(\hat{h},\hat{j},Z) \ge \min\{\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}. \tag{13}$$

The opposite inequality is trivial if the right part of (12) equals $+\infty$. Assume that this is not the case and suppose that positive integers $\hat{h}' < \hat{h}$, $\hat{j}' < \hat{j}$, a set $\hat{Z} \subset Z$, $c \in Z$, and $J \in \mathcal{J}(Z)$ are chosen in such a way that the right part of (12) achieves the minimum value for them.

By the definition of $\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}')$, there is an $\hat{h}'$-clustering $\mathcal{I}_1$ for $\mathbf{A}(\hat{Z})/J'$ of cost $\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}')$ satisfying conditions (i)–(v) of the definition, where $J' \subseteq J$ of size $\hat{j}' = |J'|$. In particular, c is a color of J.

We also have that, by definition of $w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$, there is an $(\hat{h}-\hat{h}')$-clustering for $\mathbf{A}(Z\setminus\hat{Z})$ satisfying conditions (i)-(v) of the definition. In particular, there is a special cluster $I \in \mathcal{I}_2$ of size $\hat{j}-\hat{j}'$ with the median $\mathbf{s}$.

We construct the clustering $\mathcal{I}$ for $\mathbf{A}(Z)$ as follows. First, we modify the cluster $I \in \mathcal{I}_2$ by replacing it by $I' = I \cup J'$. Then we take the union of $\mathcal{I}_1$ and the modified $\mathcal{I}_2$. It is straightforward to verify that $\mathcal{I}$ is a $\hat{h}$-clustering for $\mathbf{A}(Z)$ satisfying (i)–(v) for $w^{(i)}(\hat{h},\hat{j},Z)$. Since I' is obtained by adding $\hat{j}'$ elements of J, the cost of $\mathcal{I}$ is $\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$. Therefore, $w^{(i)}(\hat{h},\hat{j},Z) \le \omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})$ and, by the choice of $\hat{h}'$, $\hat{j}'$, $\hat{Z}$, c and J,

$$w^{(i)}(\hat{h},\hat{j},Z) \le \min\{\omega^{(1)}_{y_i}(\hat{h}',\hat{Z},c,J,\hat{j}') + \hat{j}'d(J) + w^{(i-1)}(\hat{h}-\hat{h}',\hat{j}-\hat{j}',Z\setminus\hat{Z})\}. \tag{14}$$

By (13) and (14), we conclude that the recurrence (12) holds. Then we compute the tables of values of $w^{(i)}(\cdot,\cdot,\cdot)$ for $i = 1,\ldots,f$ using (11) and (12). Finally, we apply (10) to compute $\omega^{(2)}_x(h',Y,j,\mathbf{s})$.

Clearly, the table of values of $w^{(1)}(\cdot,\cdot,\cdot)$ can be computed in $2^{\mathcal{O}(B)}\cdot n^3$ time because we consider $\hat{h},\hat{j} \le n$ and at most 2^ℓ sets Z, and then go through at most ℓ values of c and at most n sets $\mathcal{J}$. To compute the tables of values of $w^{(i)}(\cdot,\cdot,\cdot)$ for $i \ge 2$, we consider all pairs of integers $\hat{h}' < \hat{h}$, all pairs $\hat{j}' < \hat{j}$, all nonempty sets $\hat{Z} \subset Z$, all $c \in Z$, and $J \in \mathcal{J}(Z)$. Since $\hat{h}',\hat{h},\hat{j}',\hat{j} \le n$, the number of pairs of set $\hat{Z} \subset Z$ is at most 3^ℓ, the number of the choices of c is at most ℓ and the number of the choices of J is at most n, we have that the total running time is $2^{\mathcal{O}(B)}\cdot n^{\mathcal{O}(1)}$ because $\ell \le 2B$. □

Claims 3.1–3.3 allow us to compute the table of values of $\omega^{(2)}_z(\cdot,\cdot,\cdot,\cdot)$ for the root z of T bottom-up starting from the leaves (recall that $z \in U$). To make the final step of our algorithm, observe that the minimum cost of a feasible h-clustering for $\mathbf{A}(X)$ with respect to T, t' and ℓ' is

$$\min\{\omega^{(2)}_x(h,X,0,\mathbf{s}) \mid \mathbf{s} \in \mathcal{M}\}$$

by the definition of these values.

The tree T has $\ell' + t' \leq 3B$ vertices. The table of values of either $\omega_x^{(1)}(\cdot,\cdot,\cdot,\cdot,\cdot)$ or $\omega_x^{(2)}(\cdot,\cdot,\cdot,\cdot,\cdot)$ constructed for every node x has size $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ and can be constructed in $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time by Claims 3.1–3.3. Therefore, the total running time is $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$. □

Using Lemmas 5 and 6, we are able to check whether the considered instance has a colorful solution.

3.2.5. Putting all together

Now we are ready to put all ingredients of our algorithm together.

Lemma 7. Capacitated Clustering *can be solved in* $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ *time by a Monte Carlo algorithm with false negatives.*

Proof. Let $(\mathbf{A}, \Sigma, k, B, p, q)$ be an instance of Capacitated Clustering with $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. We start with computing the partition $\mathcal{J} = \{J_1, \ldots, J_s\}$ of $\{1, \ldots, n\}$ into initial clusters and this step can be done in polynomial time. By the next step, we construct the set $\mathcal{M} = \mathcal{M}(\mathbf{A}, B)$ of potential medians of size $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ in $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time using Lemma 4.

Then we consider all nonnegative $t \leq \min\{B, k\}$ to guess the number of composite clusters. If $t = 0$, then the problem is solved in polynomial time. If $t \geq 1$, then we proceed and guess the number ℓ of initial clusters having nonempty intersections with composite clusters, where $t + 1 \leq \ell \leq B$. For the chosen values of t and ℓ, we consider all forests F on $t + \ell$ vertices to guess the structure of $G(\mathcal{I}', \mathcal{J}')$. Recall that we have $2^{\mathcal{O}(B)}$ forests (see [35]) that can be listed in $2^{\mathcal{O}(B)}$ time (see [38]).

Further, we color the elements of $\mathcal{J}$ uniformly at random by ℓ colors and, given t, ℓ, F and a random coloring $\psi : \mathcal{J} \to \{1, \ldots, \ell\}$, check whether there is a feasible k-clustering of cost at most B. Recall that if there is a solution with t composite clusters such that exactly ℓ initial clusters have nonempty intersections with the composite clusters of the solution, then the probability that these ℓ clusters are assigned distinct colors in a random coloring ψ is at least e^{-2k}. Then the probability that some initial clusters having nonempty intersections with the composite clusters of the solution obtain the same color is at most $1 - e^{-2B}$. This implies that if we try e^{2B} random colorings, then the probability that for every coloring, some initial clusters having nonempty intersections with the composite clusters of the solution are of the same color is at most $(1 - e^{-2B})^{e^{2B}} \leq e^{-1}$. This implies that it is sufficient to consider $N = \lceil e^{2B} \rceil$ random coloring ψ. For each coloring, we verify the existence of a colorful solution. If a colorful solution exists for ψ, then we report that $(\mathbf{A}, \Sigma, r, k, p, q)$ admits a required solution and stop. Otherwise, if we fail to find a colorful solution for every ψ, we report that there is in solution and the probability of an incorrect answer is at most $e^{-1} < 1$.

For given t, ℓ, F and a random coloring $\psi : \mathcal{J} \to \{1, \ldots, \ell\}$, we use Lemmas 5 and 6 for verifying whether a feasible k-clustering exists. For the connected components $F_1, \ldots, F_f$ of F, we apply Lemma 6 and compute the values $\omega_i(X, h)$ for all $i \in \{1, \ldots, f\}$, $X \subseteq \{1, \ldots \ell\}$ and positive integers $h \leq k$. By Lemma 6, this can be done in $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time. Given these values, we apply Lemma 5 to check in $2^{\mathcal{O}(B)} \cdot n^2$ time whether there is a feasible k-clustering for $\mathbf{A}$ of cost at most B with respect to F, t and ℓ.

The overall running time of this algorithm is $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ and this concludes the proof. □

3.2.6. Derandomization

Our algorithm can be derandomized by standard tools [2] (see also [13, Chapter 5]). More precisely, we replace random colorings by functions from a perfect hash family.

Let s and ℓ be positive integers such that $s \geq \ell$. A set $\mathcal{F}$ of functions $\xi : \{1, \ldots, s\} \to \{1, \ldots, \ell\}$ is said to be an (s, ℓ)-*perfect hash family* if for every $X \subseteq \{1, \ldots, s\}$ of size ℓ, there is $\xi \in \mathcal{F}$ such that $\xi|_X$ is a bijection between X and $\{1, \ldots, \ell\}$.

We use the result of Naor, Schulman, and Srinivasan [32] (see also [13, Chapter 5]).

Proposition 3. *For every* $s \geq \ell \geq 1$, *there is an* (s, ℓ)*-perfect hash family* $\mathcal{F}$ *of size* $e^{\ell}\ell^{\mathcal{O}(\log \ell)} \cdot \log s$ *that can be constructed in* $e^{\ell}\ell^{\mathcal{O}(\log \ell)} \cdot s \log s$ *time.*

We consider our set of initial clusters $\mathcal{J} = \{J_1, \ldots, J_s\}$ and construct an (s, ℓ)-perfect hash family $\mathcal{F}$. Since $\ell \leq 2B$ and $s \leq n$, $|\mathcal{F}| = e^{2B}(2B)^{\mathcal{O}(\log B)} \cdot \log n$ and $\mathcal{F}$ can be constructed in $e^{2B}(2B)^{\mathcal{O}(\log B)} \cdot n \log n$ time by Proposition 3. For every $\xi \in \mathcal{F}$, we define the coloring $\psi_\xi : \mathcal{J} \to \{1, \ldots, \ell\}$ by setting $\psi_\xi(J_i) = \xi(i)$ for $i \in \{1, \ldots, s\}$.

If $(\mathbf{A}, \Sigma, k, B, p, q)$ admits a solution with t composite clusters such that exactly ℓ initial clusters have nonempty intersections with the composite clusters, then there is $\xi \in \mathcal{F}$ such that ψ_ξ colors these initial clusters by distinct colors by the definition of an (s, ℓ)-perfect hash family. Then our randomized algorithm can be modified as follows: instead of trying $N = \lceil e^{2B} \rceil$ random coloring ψ we try all $\xi \in F$, we verify the existence of a colorful solution with respect to ψ_ξ. We obtain that we can solve Capacitated Clustering for $(\mathbf{A}, \Sigma, k, B, p, q)$ in $2^{\mathcal{O}(B\log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ deterministic time and this concludes the proof of Theorem 1.

4. Clustering with size constraints

In this section, we discuss other variants of Categorical Clustering with cluster size constraints: Balanced Clustering and Factor-Balanced Clustering. We also discuss the special case of Capacitated Clustering for $p = q = n/k$ that is equivalent to Balanced Clustering for $\delta = 0$ and to Factor-Balanced Clustering for $\alpha = 1$. We refer to this problem as Equal Clustering.

Recall that by Theorem 2, Capacitated Clustering is NP-complete for $k = 2$ and $p = q = n/2$, that is, Equal Clustering is NP-complete for $k = 2$. Using the same arguments as in the proof of Theorem 2, we can show the following more general claim.

Theorem 3. *For every fixed $\alpha \geq 1$ ($\delta \geq 0$, respectively), Factor-Balanced Clustering (Balanced Clustering, respectively) is NP-complete for $k = 2$ and binary matrices.*

From the positive side, we observe that Balanced Clustering and Factor-Balanced Clustering admit Turing reductions to Capacitated Clustering, that is, Capacitated Clustering is the most general among the considered problems. For this, we make the following straightforward observation.

Observation 5. *An instance $(\mathbf{A}, \Sigma, k, B, \delta)$ of Balanced Clustering (an instance $(\mathbf{A}, \Sigma, k, B, \alpha)$ of Factor-Balanced Clustering, respectively) is a yes-instance if and only if there is a nonnegative integer p such that $\frac{n}{k} - \delta \leq p \leq \frac{n}{k}$ ($\frac{n}{\alpha k} \leq p \leq \frac{n}{k}$, respectively) and for $q = p + \delta$ ($q = \alpha p$, respectively), $(\mathbf{A}, \Sigma, k, B, p, q)$ is a yes-instance of Capacitated Clustering.*

Thus, given an algorithm $\mathcal{A}$ for Capacitated Clustering, we can solve Balanced Clustering for $(\mathbf{A}, \Sigma, k, B, \delta)$ as follows. We consider all p starting from $\max\{1, \lceil \frac{n}{k} \rceil - \delta\}$ up to $\lfloor \frac{n}{k} \rfloor$, and use $\mathcal{A}$ to solve Capacitated Clustering for $(\mathbf{A}, \Sigma, k, B, p, \min\{n, p + \delta\})$. If $\mathcal{A}$ returns "yes" for one of the values of p, we conclude that $(\mathbf{A}, \Sigma, k, B, \delta)$ is a yes-instance of Balanced Clustering and stop. Otherwise, if $\mathcal{A}$ always returns "no", $(\mathbf{A}, \Sigma, k, B, \delta)$ is a no-instance. Clearly, Factor-Balanced Clustering can be solved in similar way. This allows to obtain the following corollary of Theorem 1.

Corollary 1. Balanced Clustering *and* Factor-Balanced Clustering *are solvable in time* $2^{\mathcal{O}(B \log B)} |\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$.

5. Kernelization for clustering with size constraints

In this section, we discuss kernelization for clustering problems with size constraints. In [19, Theorem 3], Fomin, Golovach and Panolan proved that Categorical Clustering does not admit a polynomial kernel when parmeterized by B, unless NP $\subseteq$ coNP/poly. This immediately implies the following proposition.

Proposition 4. Capacitated Clustering (Balanced Clustering *and* Factor-Balanced Clustering, *respectively) has no polynomial kernel when paramterized by B, unless* NP $\subseteq$ coNP/poly, *even if* $\Sigma = \{0, 1\}$.

Also, by Theorems 2 and 3 the problems are already NP-hard for $k = 2$. Thus, for kernelization, we have to consider more restrictive parameterizations. Up to now, we have only partial results. In particular, we can show Balanced Clustering admits a polynomial kernel when parameterzied by B, k and δ.

We start with some auxiliary results. First, we observe that if there is an initial cluster J of size at least $B + 1$, then at least one median should be the same as a column of the input matrix with its index in J.

Observation 6. *Let $\{I_1, \ldots, I_k\}$ be a k-clustering for a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ of cost at most B and let $J \subseteq \{1, \ldots, n\}$ be an initial cluster with $|J| \geq B + 1$. Then there is $i \in \{1, \ldots, k\}$ such that an optimal median of I_i coincides with $\mathbf{s} = \mathbf{a}_j$ for $j \in J$.*

Proof. For the sake of contradiction, assume that medians $\mathbf{c}_1, \ldots, \mathbf{c}_k$ for the clusters $\{I_1, \ldots, I_k\}$, respectively, are distinct from $\mathbf{s}$. Then

$$\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \geq \sum_{i=1}^{k} \sum_{j \in J \cap I_i} d_H(\mathbf{c}_i, \mathbf{s}) \geq |J| > B$$

contradicting that the cost of $\{I_1, \ldots, I_k\}$ is at most B. □

Our next lemma shows that if there is a clustering such that a median $\mathbf{c}_i$ coincides with a column $\mathbf{a}_j$, then we can either collect all the elements of the initial cluster J containing j in the same cluster of a solution or form a cluster of a solution out of its elements.

Lemma 8. *Let $\{I_1,\ldots,I_k\}$ be a k-clustering for a matrix $\mathbf{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_n)$ with optimal medians $\mathbf{c}_1,\ldots,\mathbf{c}_k$, respectively. Let also $\mathbf{S}\subseteq\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}$ be the set of medians coinciding with columns of $\mathbf{A}$. Then there is a k-clustering $\{I'_1,\ldots,I'_k\}$ for $\mathbf{A}$ such that*

(i) $|I'_i|=|I_i|$ *for all* $i\in\{1,\ldots,k\}$,
(ii) $\sum_{i=1}^{k}\sum_{j\in I'_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq\sum_{i=1}^{k}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)$, *and*
(iii) *for every* $\mathbf{s}\in\mathbf{S}$ *and the initial cluster* J *such that* $\mathbf{s}=\mathbf{a}_j$ *for* $j\in J$, *there is* $i\in\{1,\ldots,k\}$ *such that either* $J\subseteq I'_i$ *or* $I'_i\subset J$.

Proof. Let $\mathbf{c}_1,\ldots,\mathbf{c}_k$ be optimal medians for $I_1,\ldots,I_k$, respectively. Assume without loss of generality that $\mathbf{S}=\{\mathbf{c}_1,\ldots,\mathbf{c}_t\}$, and denote by $J_1,\ldots,J_t$ the initial clusters such that for every $i\in\{1,\ldots,t\}$, $\mathbf{a}_j=\mathbf{c}_i$ for $j\in J_i$. Let $\mathcal{I}'=\{I'_1,\ldots,I'_k\}$ be a k-clustering for $\mathbf{A}$ such that (a) $|I'_i|=|I_i|$ for all $i\in\{1,\ldots,k\}$, (b) $\sum_{i=1}^{k}\sum_{j\in I'_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq\sum_{i=1}^{k}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)$, and (c) $\sum_{i=1}^{t}|I'_i\cap J_i|$ is maximum. We claim that $\mathcal{I}'$ satisfies conditions (i)–(iii) of the lemma. Clearly, (i) and (ii) are fulfilled by conditions (a) and (b) of the choice of $\mathcal{I}'$. To show (iii), we prove that either $J_i\subseteq I'_i$ or $I'_i\subset J_i$ for every $i\in\{1,\ldots,t\}$.

Assume to the contrary that there is $i\in\{1,\ldots,t\}$ such that neither $J_i\subseteq I'_i$ nor $I'_i\subset J_i$. Then there is a cluster I'_j for $j\in\{1,\ldots,k\}$ such that $j\neq i$, $I'_j\cap J\neq\emptyset$, and there is $h\in I'_i$ such that $h\notin J_i$. Let $\ell\in I'_j\cap J_i$. Consider the k-clustering $\mathcal{I}''=\{I''_1,\ldots,I''_k\}$ such that $I''_i=(I'_i\cup\{\ell\})\setminus\{h\}$, $I''_j=(I'_j\cup\{h\})\setminus\{\ell\}$, and $I''_h=I'_h$ for $h\in\{1,\ldots,k\}$ such that $h\neq i,j$. In words, we exchange the elements h and ℓ between I'_i and I'_j. Then

$$\sum_{p=1}^{k}\sum_{q\in I'_p} d_H(\mathbf{c}_p,\mathbf{a}_q)-\sum_{p=1}^{k}\sum_{q\in I''_p} d_H(\mathbf{c}_p,\mathbf{a}_q)=d_H(\mathbf{c}_i,\mathbf{a}_h)+d_H(\mathbf{c}_j,\mathbf{a}_\ell)-d_H(\mathbf{c}_i,\mathbf{a}_\ell)-d_H(\mathbf{c}_j,\mathbf{a}_h),$$

and since $\mathbf{c}_i=\mathbf{a}_\ell$, we obtain that

$$\sum_{p=1}^{k}\sum_{q\in I'_p} d_H(\mathbf{c}_p,\mathbf{a}_q)-\sum_{p=1}^{k}\sum_{q\in I''_p} d_H(\mathbf{c}_p,\mathbf{a}_q)=d_H(\mathbf{a}_\ell,\mathbf{a}_h)+d_H(\mathbf{c}_j,\mathbf{a}_\ell)-d_H(\mathbf{c}_j,\mathbf{a}_h)\geq 0$$

by the triangle inequality. This means that

$$\sum_{p=1}^{k}\sum_{q\in I''_p} d_H(\mathbf{c}_p,\mathbf{a}_q)\leq\sum_{p=1}^{k}\sum_{q\in I'_p} d_H(\mathbf{c}_p,\mathbf{a}_q)\leq\sum_{p=1}^{k}\sum_{q\in I_p} d_H(\mathbf{c}_p,\mathbf{a}_q). \tag{15}$$

Since $|I''_i|=|I'_i|$ for all $i\in\{1,\ldots,r\}$, $\mathcal{I}''$ satisfies condition (a) of the choice of $\mathcal{I}'$. Condition (b) is satisfied because of (15). However, $|I''_i\cap J|=|(I'_1\cap I)\cup\{\ell\}|=|I'_i\cap J|+1$. Because $\mathcal{I}''$ was obtained by the exchange h and ℓ between I'_i and I'_j, $I'_p\cap J_p\subseteq I''_p\cap J_p$ for $p\in\{1,\ldots,t\}$. We obtain that $\sum_{p=1}^{t}|I'_p\cap J_p|<\sum_{p=1}^{t}|I''_p\cap J_p|$ contradicting (c). Therefore, either $J_p\subseteq I'_p$ or $I'_p\subset J_p$ for every $p\in\{1,\ldots,t\}$ as it claimed. □

The following lemma is used to find medians if the sizes of clusters in a solution are sufficiently big.

Lemma 9. *Let $\mathcal{I}=\{I_1,\ldots,I_k\}$ be a k-clustering for a matrix $\mathbf{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_n)$ of cost at most B such that $s\leq|I_i|\leq s+\delta$ for all $i\in\{1,\ldots,k\}$, where δ is a nonnegative integer and an integer $s\geq 2B+1+(k-1)\delta$. Then for every initial clusters $J\subseteq\{1,\ldots,n\}$, the following is fulfilled for $\mathbf{c}=\mathbf{a}_j$ for $j\in J$:*

(i) *if* $|J| \mod s\geq B+1+(k-1)\delta$, *then exactly* $\left\lceil\frac{|J|}{s}\right\rceil$ *clusters of* $\mathcal{I}$ *have optimal medians coinciding with* $\mathbf{c}$ *(the other medians are different),*
(ii) *if* $|J| \mod s\leq B+(k-1)\delta$, *then exactly* $\left\lfloor\frac{|J|}{s}\right\rfloor$ *clusters of* $\mathcal{I}$ *have optimal medians coinciding with* $\mathbf{c}$.

Proof. We start with proving (i). Let $|J| \mod s\geq B+1+(k-1)\delta$. We show that (i) holds for J by induction on $p=\left\lfloor\frac{|J|}{s}\right\rfloor$.

The base case is $p=0$. Then $\left\lceil\frac{|J|}{s}\right\rceil=1$. As $|J| \mod s\geq B+1+(k-1)\delta$ and $\left\lfloor\frac{|J|}{s}\right\rfloor=0$, $B+1\leq|J|\leq s$. By Observation 6, there is a cluster in $\mathcal{I}$ whose optimal median is $\mathbf{c}$. Thus, at least one optimal median coincides with $\mathbf{c}$. Without loss of generality, we assume that $\mathbf{c}$ is the median of I_1. We now show that $\mathbf{c}_i\neq\mathbf{c}$ for $i\in\{2,\ldots,k\}$. Assume to the contrary that there exists $h\in\{2,\ldots,k\}$ such that $\mathbf{c}_h=\mathbf{c}$. By Lemma 8, there is a k-clustering $\mathcal{I}=\{I'_1,\ldots,I'_k\}$ for $\mathbf{A}$ such that $|I'_i|=|I'_i|$ for all $i\in\{1,\ldots,k\}$, $\sum_{i=1}^{k}\sum_{j\in I'_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\leq\sum_{i=1}^{r}\sum_{j\in I_i} d_H(\mathbf{c}_i,\mathbf{a}_j)$, and $J\subseteq I'_1$. Then

$$\sum_{i=1}^{k}\sum_{j\in I'_i} d_H(\mathbf{c}_i,\mathbf{a}_j)\geq\sum_{j\in I'_h} d_H(\mathbf{c}_h,\mathbf{a}_j)=\sum_{j\in I'_h} d_H(\mathbf{c},\mathbf{a}_j)\geq|I'_h|\geq s\geq B+1,$$

contradicting that $\mathrm{cost}(\mathcal{I}) \leq B$. We conclude that exactly one median coincides with $\mathbf{c}$, that is, (i) holds for $p = 0$.

Now let $p \geq 1$ and assume that the claim holds when p is smaller. Note that $k \geq 2$ in this case. We observe that, because $|J| \mod s \geq B + 1 + (k-1)\delta$, $|J| \geq sp + B + 1 + (k-1)\delta$. By Observation 6, there is a cluster in $\mathcal{I}$ whose optimal median is $\mathbf{c}$. Without loss of generality, we assume that $\mathbf{c}$ is the median of I_1. Then by Lemma 8, there is a k-clustering $\{I'_1, \ldots, I'_k\}$ for $\mathbf{A}$ such that $|I'_i| = |I_i|$ for all $i \in \{1, \ldots, k\}$, $\sum_{i=1}^{k} \sum_{j \in I'_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq \sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j)$, and $I'_1 \subset J$. Consider $\mathbf{A}' = \mathbf{A}[\{1, \ldots, m\}, \{1, \ldots, n\} \setminus I_1]$, that is, $\mathbf{A}'$ is obtained from $\mathbf{A}$ by the deletion of the columns with their indices in I_1. Notice that $\mathcal{I}' = \{I'_2, \ldots, I'_k\}$ is an $(k-1)$-clustering for $\mathbf{A}'$ of cost at most B. Moreover, because $|I'_i| = |I_i| \geq s \geq 2B + 1$, $\mathbf{c}_2, \ldots, \mathbf{c}_k$ are unique optimal medians for $I'_2, \ldots, I'_k$, respectively, by Observation 2. Let $J' = J \setminus I'_1$. Since $|I'_1| \leq s + \delta$,

$$|J'| = |J| - |I'_1| \geq sp + B + 1 + (k-1)\delta - s - \delta = s(p-1) + B + 1 + (k-2)\delta \geq B + 1 + (k-2)\delta.$$

By our inductive hypothesis, exactly $\left\lceil \frac{|J'|}{s} \right\rceil$ clusters of $\mathcal{I}'$ have optimal medians coinciding with $\mathbf{c}$. As $|I_1| \geq s$, $\left\lfloor \frac{|J'|}{s} \right\rfloor \leq p - 1$. Because $|J'| \geq s(p-1) + B + 1 + (k-2)\delta$, $\left\lfloor \frac{|J'|}{s} \right\rfloor \geq p - 1$. Hence, $\left\lfloor \frac{|J'|}{s} \right\rfloor = p - 1$ and $\left\lceil \frac{|J'|}{s} \right\rceil = p$. Since $\mathbf{c}_2, \ldots, \mathbf{c}_k$ are optimal medians, exactly p of them are equal to $\mathbf{c}$. Together with the median $\mathbf{c}_1 = \mathbf{c}$, exactly $p + 1$ medians in $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ are equal to $\mathbf{c}$. Then exactly $\left\lceil \frac{|J|}{s} \right\rceil = p + 1$ clusters of $\mathcal{I}$ have optimal medians coinciding with $\mathbf{c}$. This completes the proof of (i).

To show (ii), we first claim that for every initial cluster J, there are at least $p = \left\lfloor \frac{|J|}{s} \right\rfloor$ clusters in $\mathcal{I}$, whose optimal medians are equal to $\mathbf{c}$, where $\mathbf{c} = \mathbf{a}_j$ for $j \in J$. The proof is by induction on p.

The claim is trivial if $p = 0$. Let $p \geq 1$ and assume that the claim holds when p is smaller. Since $p \geq 1$, $|J| \geq s \geq B + 1$. By Observation 6, there is a cluster in $\mathcal{I}$ whose optimal median is $\mathbf{c}$. Without loss of generality, we assume that $\mathbf{c}$ is the median of I_1. Then by Lemma 8, there is an r-clustering $\{I'_1, \ldots, I'_k\}$ for $\mathbf{A}$ such that $|I'_i| = |I_i|$ for all $i \in \{1, \ldots, k\}$, $\sum_{i=1}^{k} \sum_{j \in I'_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \leq \sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j)$, and either $J \subseteq I'_1$ or $I'_1 \subset J$.

Suppose that $J \subseteq I'_1$. Then $|J| \leq |I'_i| \leq s + \delta < 2s$. This means that $p = 1$ and our claim holds, as $\mathbf{c} = \mathbf{c}_1$.

Assume from now that this is not the case, that is, $I'_i \subset J$. Then we argue similarly to the proof of (i). Consider $\mathbf{A}' = \mathbf{A}[\{1, \ldots, m\}, \{1, \ldots, n\} \setminus I_1]$, that is, $\mathbf{A}'$ is obtained from $\mathbf{A}$ by the deletion of the columns with their indices in I_1. Notice that $\mathcal{I}' = \{I'_2, \ldots, I'_k\}$ is an $(r-1)$-clustering for $\mathbf{A}'$ of cost at most B. Moreover, because $|I'_i| = |I_i| \geq s \geq 2B + 1$, $\mathbf{c}_2, \ldots, \mathbf{c}_k$ are unique optimal medians for $I'_2, \ldots, I'_k$, respectively, by Observation 2. Let $J' = J \setminus I'_1$.

If $\left\lfloor \frac{|J'|}{s} \right\rfloor \geq p - 1$, then by the inductive assumption, there are at least $p - 1$ clusters in $\mathcal{I}'$, whose optimal medians coincide with $\mathbf{c}$. Thus, at least $p - 1$ medians from $\{\mathbf{c}_2, \ldots, \mathbf{c}_k\}$ are equal to $\mathbf{c}$. Taking into account $\mathbf{c}_1 = \mathbf{c}$, we have that at least p medians from $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ are equal to $\mathbf{c}$, as required.

Let $\left\lfloor \frac{|J'|}{s} \right\rfloor \leq p - 2$. Note that $p \geq 2$ in this case. Since $|I'_1| \leq s + \delta$ and $|J| \geq ps$, we obtain that $|J'| = |J| - |I'_1| \geq (p-2)s + (s - \delta)$. Thus, $\left\lfloor \frac{|J'|}{s} \right\rfloor = p - 2$ and

$$|J'| \mod s \geq s - \delta \geq 2B + 1 + (k-1)\delta \geq B + 1 + (k-2)\delta.$$

By the already proven (i), we have that there are at least $\left\lceil \frac{|J'|}{s} \right\rceil = p - 1$ clusters in $\mathcal{I}'$, whose optimal medians coincide with $\mathbf{c}$. Since $\mathbf{c}_1 = \mathbf{c}$, we again obtain that at least p medians from $\{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ are equal to $\mathbf{c}$. This concludes the proof of our auxiliary claim.

To finish the proof of (ii), assume that $|J| \mod s \leq B + (k-1)\delta$. We already have that at least $p = \left\lfloor \frac{|J|}{s} \right\rfloor$ clusters of $\mathcal{I}$ have optimal medians coinciding with $\mathbf{c}$. It remains to show that there are at most p such clusters. Assume to the contrary that at least $p + 1$ medians are equal to $\mathbf{s}$ and assume without loss of generality that $\mathbf{c} = \mathbf{c}_1 = \ldots = \mathbf{c}_{p+1}$. Then

$$\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \geq \sum_{i=1}^{p+1} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j j) = \sum_{j \in I_1 \cup \ldots \cup I_{p+1}} d_H(\mathbf{c}, \mathbf{a}_j)$$
$$\geq \sum_{j \in (I_1 \cup \ldots \cup I_{p+1}) \setminus J} d_H(\mathbf{c}, \mathbf{a}_j) \geq |(I_1 \cup \ldots \cup I_{p-1}) \setminus I|.$$

We know that $|I_i| \geq s$ for $i \in \{1, \ldots, k\}$. Then $|I_1 \cup \ldots \cup I_{p+1}| \geq s(p+1)$. Since $|J| \mod s \leq B + (k-1)\delta$, $|J| \leqslant ps + B + (k-1)\delta$. This implies $|(I_1 \cup \ldots \cup I_{p+1}) \setminus J| \geq s - B - (k-1)\delta \geqslant B + 1$. Hence $\sum_{j \in (I_1 \cup \ldots \cup I_{P+1}) \setminus J} d_H(\mathbf{c}, \mathbf{a}_j) \geq B + 1 > B$ contradicting that $\mathrm{cost}(\mathcal{I}) \leq B$. This proves that exactly $\left\lfloor \frac{|J|}{s} \right\rfloor$ clusters of $\mathcal{I}$ have optimal medians coinciding with $\mathbf{c}$. □

Lemma 9 allows us to compute optimal medians and solve Balanced Clustering if the average size of clusters is sufficiently big.

Lemma 10. Balanced Clustering *can be solved in polynomial time for instances* $(\mathbf{A}, \Sigma, k, B, \delta)$ *with* $\frac{n}{k} \geq 2B + 1 + \delta k$.

Proof. Let $(\mathbf{A}, \Sigma, k, B, \delta)$ be an instance of Balanced Clustering with $\frac{n}{k} \geq 2B+1+\delta k$. Clearly, we can assume that $\delta \leq n-1$. If $(\mathbf{A}, \Sigma, k, B, \delta)$ is a yes-instance, then there is an integer s such that $\frac{n}{k} - \delta \leq s \leq \frac{n}{k}$ and $s \leqslant |I_i| \leqslant s+\delta$ for a solution $\{I_1, \ldots, I_k\}$ to the instance.

Then we consider all integers s such that $\frac{n}{k} - \delta \leq s \leq \frac{n}{k}$. For each value of s, we check whether there is a solution $\{I_1, \ldots, I_k\}$ for the considered instance with $s \leq |I_i| \leq s+\delta$, for all $i \in \{1, \ldots, k\}$. If yes, we return the yes-answer, otherwise, if we fail to find a solution for every s, then the algorithm returns the no-answer.

Let s be fixed. For each initial cluster J, we compute $\left\lfloor \frac{|J|}{s} \right\rfloor$ and $|J| \mod s$. Using these two values, we find the medians coinciding with $\mathbf{c}$ such that $\mathbf{c} = \mathbf{a}_j$ for $j \in J$ using Lemma 9. Denote by $\mathcal{C}$ the obtained collection of medians. If $|\mathcal{C}| \neq k$, then we discard the current choice of s. Otherwise, $\mathcal{C}$ contains exactly k potential medians and we combine Observation 5 and Lemma 1 to decide whether $(\mathbf{A}, \Sigma, k, B, \delta)$ admits a solution with these medians.

Since we consider at most $\delta + 1 \leq n$ values of s and the algorithm from Lemma 1 is polynomial, the total running time of our algorithm is polynomial. □

In [19], Fomin et al. proved that Categorical Clustering admits a polynomial kernel when parameterized by B and k for binary matrices. As one of the steps of their kernelization algorithm (see Theorem 2 of [19]), they show that the number of rows in the output matrix can be reduced to $\mathcal{O}(B(B+r))$. Formally, the proof is done for the binary case, that is, for $\Sigma = \{0, 1\}$. However, the reduction rule used in [19] works for arbitrary alphabet Σ because to apply the rule, we only should be able to compute the Hamming distances between pairs of columns of $\mathbf{A}$ and check whether two rows of certain submatrices are the same. We state this result in the following lemma.

Lemma 11 *([19]). There is a polynomial algorithm that, given an instance $(\mathbf{A}, \Sigma, k, B)$ of Categorical Clustering with $m \times n$ matrix $\mathbf{A}$, produces an equivalent instance $(\mathbf{A}', \Sigma, k, B)$ with $m' \times n$ matrix $\mathbf{A}'$ such that the following holds:*

- $m' = \mathcal{O}(B(B+k))$.
- *$\{I_1, \ldots, I_k\}$ is a solution for $(\mathbf{A}, \Sigma, k, B)$ if and only if it is also a solution for $(\mathbf{A}', \Sigma, k, B)$.*

Now we are ready to show a polynomial kernel for Balanced Clustering.

Theorem 4. Balanced Clustering *admits a kernel, where the output matrix has $\mathcal{O}(B(B+k))$ rows and $O(k(B+\delta k))$ columns, and is a matrix over an alphabet of size at most $B+k$.*

Proof. Let $(\mathbf{A}, \Sigma, k, B, \delta)$ be an instance of Balanced Clustering with $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$.

Suppose $\frac{n}{k} \geq 2B + 1 + \delta k$. Then, by Lemma 10, the problem can be solved in polynomial time. We do it and return a trivial yes or no-instance, respectively. For example, we can return either the matrix $(0, 0)$ or $(0, 1)$, respectively, and set $k = 1$, $B = 0$ and $\delta = 0$. Assume from now that $\frac{n}{k} \leq 2B + \delta k$, that is, $n \leq 2Bk + \delta k^2$.

If $\mathbf{A}$ has at least $B + k + 1$ pairwise distinct columns, then for every k-clustering $\{I_1, \ldots, I_k\}$ and every $\mathbf{c}_1, \ldots, \mathbf{c}_k \in \Sigma^m$, $\sum_{i=1}^{k} \sum_{j \in I_i} d_H(\mathbf{c}_i, \mathbf{a}_j) \geq B + 1$ because at least $B + 1$ columns of $\mathbf{A}$ are distinct from each median. Thus, $(\mathbf{A}, \Sigma, k, B, \delta)$ is a no-instance in this case, and we return a trivial no-instance of Balanced Clustering.

Assume from now that the number of pairwise distinct columns is at most $B + k$. If $|\Sigma| > B + k$, then we can replace every symbol of Σ by a symbol of $\Sigma' = \{0, \ldots, B + k - 1\}$ maintaining the following property: for each row of $\mathbf{A}$, the same symbols of Σ are replaced by the same symbols of Σ'. It is straightforward to verify that this replacement produces an equivalent instance because we are using the Hamming distances. From now, we assume that $|\Sigma| \leq B + k$.

Given $(\mathbf{A}, \Sigma, k, B, \delta)$, we consider the instance $(\mathbf{A}, \Sigma, k, B)$ of Categorical Clustering. We use the algorithm from Lemma 11 and denote by $(\mathbf{A}', \Sigma, k, B)$ the output instance. Then we construct the instance $(\mathbf{A}', \Sigma, k, B, \delta)$ of Balanced Clustering and output it.

We show that $(\mathbf{A}, \Sigma, k, B, \delta)$ is a yes-instance of Balanced Clustering if and only if $(\mathbf{A}', \Sigma, k, B, \delta)$ is a yes-instance.

For the forward direction, suppose $(\mathbf{A}, \Sigma, k, B, \delta)$ is a yes-instance of Balanced Clustering. Let $\mathcal{I} = \{I_1, \ldots, I_k\}$ be a solution to the instance. Clearly, $\mathcal{I}$ is a solution for the instance $(\mathbf{A}, \Sigma, k, B)$ of Categorical Clustering. By Lemma 11, $\mathcal{I}$ is a solution for $(\mathbf{A}', \Sigma, k, B)$. Then $\mathcal{I}$ is a solution for $(\mathbf{A}', \Sigma, k, B, \delta)$. For the opposite direction, the arguments are similar. Let $\mathcal{I} = \{I_1, \ldots, I_k\}$ be a solution for $(\mathbf{A}', \Sigma, k, B, \delta)$. Then this is a solution for the instance $(\mathbf{A}', \Sigma, k, B)$ of Categorical Clustering and, by Lemma 11, a solution for $(\mathbf{A}', \Sigma, k, B)$. Finally, $\mathcal{I}$ is a solution of $(\mathbf{A}, \Sigma, k, B, \delta)$.

Recall that $n = \mathcal{O}(k(B + \delta k))$ and note that $\mathbf{A}'$ has $\mathcal{O}(B(B+k))$ rows by Lemma 11. Since $|\Sigma| \leq B + k$, we conclude that the output matrix has $\mathcal{O}(B(B+r))$ rows and $O(k(B+\delta k))$ columns, and is a matrix over an alphabet of size at most $B+k$.

It is easy to see that our kernelization algorithm is polynomial and this concludes the proof. □

6. Conclusion

We proved that Capacitated Clustering can be solved in $2^{\mathcal{O}(B \log B)}|\Sigma|^B \cdot (mn)^{\mathcal{O}(1)}$ time. This also implies that the same holds for Balanced Clustering and Factor-Balanced Clustering. The natural question is whether it is possible to improve

the dependence on B. We do not know the answer to this question even for the special case of Equal Clustering. Also, besides the considered size constraints, it may be interesting to consider other variants. For example, in Capacitated Clustering, the size constraints p and q are universal for all clusters. However, one may consider the case when the cluster sizes are given by individual constraints.

Another important direction of research is the investigation of kernelization for clustering problems with size constraints and we only initiated this study in Section 5. We list only several possible directions of research. In particular, Theorem 4 leads to the question of whether Factor-Balanced Clustering admits a polynomial kernel when parameterized by k and B with the assumption that α is a fixed constant. A more general question is whether there are polynomial kernels for Capacitated Clustering, Balanced Clustering, and Factor-Balanced Clustering parameterized by k and B. Notice that Categorical Clustering has a polynomial kernel for this parmeterization [19, Theorem 2]. Another direction of research is to investigate kernels of other types. Are there polynomial *Turing kernels* and do these problems admit polynomial *lossy kernels*, that is, approximative kernels? (We refer to the book [17] for the definition of the notions.)

CRediT authorship contribution statement

Fedor V. Fomin: Conceptualisation, Methodology, Investigation, Writing. **Petr A. Golovach:** Conceptualisation, Methodology, Investigation, Writing. **Nidhi Purohit:** Conceptualisation, Methodology, Investigation, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] Gagan Aggarwal, Rina Panigrahy, Tomás Feder, Dilys Thomas, Krishnaram Kenthapadi, Samir Khuller, An Zhu, Achieving anonymity via clustering, ACM Trans. Algorithms 6 (3) (2010).
[2] Noga Alon, Raphael Yuster, Uri Zwick, Color-coding, J. ACM 42 (4) (1995) 844–856, https://doi.org/10.1145/210332.210337.
[3] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al., Modern Information Retrieval, vol. 463, ACM Press, New York, 1999.
[4] Frank Ban, Vijay Bhattiprolu, Karl Bringmann, Pavel Kolev, Euiwoong Lee, David P. Woodruff, A PTAS for ℓ_p-low rank approximation, in: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, SIAM, 2019, pp. 747–766.
[5] Arindam Banerjee, Joydeep Ghosh, Clustering with balancing constraints, in: Constrained Clustering: Advances in Algorithms, Theory, and Applications, CRC Press, 2008, pp. 171–200.
[6] Jaroslaw Byrka, Krzysztof Fleszar, Bartosz Rybicki, Joachim Spoerhase, Bi-factor approximation algorithms for hard capacitated k-median problems, in: Piotr Indyk (Ed.), Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015, SIAM, 2015, pp. 722–736.
[7] Jaroslaw Byrka, Bartosz Rybicki, Sumedha Uniyal, An approximation algorithm for uniform capacitated k-median problem with $1+\varepsilon$ capacity violation, in: Quentin Louveaux, Martin Skutella (Eds.), Integer Programming and Combinatorial Optimization - 18th International Conference, IPCO 2016, Liège, Belgium, June 1-3, 2016, Proceedings, in: Lecture Notes in Computer Science, vol. 9682, Springer, 2016, pp. 262–274.
[8] Moses Charikar, Sudipto Guha, Éva Tardos, David B. Shmoys, A constant-factor approximation algorithm for the k-median problem, J. Comput. Syst. Sci. 65 (1) (2002) 129–149.
[9] Danny Z. Chen, Jian Li, Hongyu Liang, Haitao Wang, Matroid and knapsack center problems, Algorithmica 75 (1) (2016) 27–52.
[10] Julia Chuzhoy, Yuval Rabani, Approximating k-median with non-uniform capacities, in: Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005, SIAM, 2005, pp. 952–958.
[11] Rudi Cilibrasi, Leo van Iersel, Steven Kelk, John Tromp, The complexity of the single individual SNP haplotyping problem, Algorithmica (ISSN 0178-4617) 49 (1) (2007) 13–36, https://doi.org/10.1007/s00453-007-0029-z.
[12] Vincent Cohen-Addad, Jason Li, On the fixed-parameter tractability of capacitated clustering, in: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, Stefano Leonardi (Eds.), 46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece, in: LIPIcs, vol. 132, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, pp. 41:1–41:14.
[13] Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, Saket Saurabh, Parameterized Algorithms, Springer, ISBN 978-3-319-21274-6, 2015.
[14] H. Gökalp Demirci, Shi Li, Constant approximation for capacitated k-median with $(1+\varepsilon)$-capacity violation, in: 43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy, 2016, pp. 73:1–73:14.
[15] Rodney G. Downey, Michael R. Fellows, Fundamentals of Parameterized Complexity, Texts in Computer Science, Springer, ISBN 978-1-4471-5558-4, 2013.
[16] Uriel Feige, NP-hardness of hypercube 2-segmentation, CoRR, arXiv:1411.0821 [abs], 2014.
[17] Fedor V. Fomin, Daniel Lokshtanov, Saket Saurabh, Meirav Zehavi, Kernelization, Cambridge University Press, Cambridge, ISBN 978-1-107-05776-0, 2019, Theory of parameterized preprocessing.
[18] Fedor V. Fomin, Petr A. Golovach, Daniel Lokshtanov, Fahad Panolan, Saket Saurabh, Approximation schemes for low-rank binary matrix approximation problems, ACM Trans. Algorithms 16 (1) (2020) 12:1–12:39, https://doi.org/10.1145/3365653.
[19] Fedor V. Fomin, Petr A. Golovach, Fahad Panolan, Parameterized low-rank binary matrix approximation, Data Min. Knowl. Discov. 34 (2) (2020) 478–532, https://doi.org/10.1007/s10618-019-00669-5.

[20] Fedor V. Fomin, Petr A. Golovach, Kirill Simonov, Parameterized k-clustering: tractability island, J. Comput. Syst. Sci. 117 (2021) 50–74, https://doi.org/10.1016/j.jcss.2020.10.005.
[21] Michael R. Garey, David S. Johnson, Computers and Intractability, A Guide to the Theory of NP-Completeness, W.H. Freeman and Company, New York, 1979.
[22] Soheil Ghiasi, Ankur Srivastava, Xiaojian Yang, Majid Sarrafzadeh, Optimal energy aware clustering in sensor networks, Sensors 2 (7) (2002) 258–269.
[23] Gaurav Gupta, Mohamed Younis, Load-balanced clustering of wireless sensor networks, in: IEEE International Conference on Communications (ICC), vol. 3, IEEE, 2003, pp. 1848–1852.
[24] Jon Kleinberg, Christos Papadimitriou, Prabhakar Raghavan, Segmentation problems, J. ACM (ISSN 0004-5411) 51 (2) (2004) 263–280, https://doi.org/10.1145/972639.972644.
[25] H.W. Kuhn, The Hungarian method for the assignment problem, Nav. Res. Logist. Q. (ISSN 0028-1441) 2 (1955) 83–97, https://doi.org/10.1002/nav.3800020109.
[26] Shi Li, On uniform capacitated k-median beyond the natural LP relaxation, ACM Trans. Algorithms 13 (2) (2017) 22:1–22:18.
[27] László Lovász, Michael D. Plummer, Matching Theory, AMS Chelsea Publishing, Providence, RI, ISBN 978-0-8218-4759-6, 2009.
[28] Patrick J. Lynch, Sarah Horton, Sarah Horton, Web Style Guide: Basic Design Principles for Creating Web Sites, Universities Press, 1999.
[29] Mikko I. Malinen, Pasi Fränti, Balanced k-means for clustering, in: Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings, in: Lecture Notes in Computer Science, vol. 8621, Springer, 2014, pp. 32–41.
[30] Dániel Marx, Closest substring problems with small distances, SIAM J. Comput. 38 (4) (2008) 1382–1410, https://doi.org/10.1137/060673898.
[31] Pauli Miettinen, Taneli Mielikäinen, Aristides Gionis, Gautam Das, Heikki Mannila, The discrete basis problem, IEEE Trans. Knowl. Data Eng. 20 (10) (2008) 1348–1362, https://doi.org/10.1109/TKDE.2008.53.
[32] Moni Naor, Leonard J. Schulman, Aravind Srinivasan, Splitters and near-optimal derandomization, in: Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 1995, pp. 182–191.
[33] A.C. Nielsen, Category Management: Positioning Your Organization to Win, McGraw-Hill, Chicago, 1992.
[34] Rafail Ostrovsky, Yuval Rabani, Polynomial-time approximation schemes for geometric min-sum median clustering, J. ACM (ISSN 0004-5411) 49 (2) (2002) 139–156, https://doi.org/10.1145/506147.506149.
[35] Richard Otter, The number of trees, Ann. Math. (2) (ISSN 0003-486X) 49 (1948) 583–599, https://doi.org/10.2307/1969046.
[36] Clemens Rösner, Melanie Schmidt, Privacy preserving clustering with constraints, in: 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
[37] Diego Vallejo-Huanga, Paulina Morillo, César Ferri, Semi-supervised clustering algorithms for grouping scientific articles, in: International Conference on Computational Science (ICCS), in: Procedia Computer Science, vol. 108, Elsevier, 2017, pp. 325–334.
[38] Robert Alan Wright, L. Bruce Richmond, Andrew M. Odlyzko, Brendan D. McKay, Constant time generation of free trees, SIAM J. Comput. 15 (2) (1986) 540–548, https://doi.org/10.1137/0215039.
[39] Yinghui Yang, Balaji Padmanabhan, Segmenting customer transactions using a pattern-based clustering approach, in: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, 2003, pp. 411–418, https://ieeexplore.ieee.org/xpl/conhome/8854/proceeding.