



Algorithms and complexity of range clustering

Dorit S. Hochbaum 

Department of Industrial Engineering and Operations Research, University of California, Berkeley, California,

Correspondence

Dorit S. Hochbaum, Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720.

Email: hochbaum@ieor.berkeley.edu

Funding information

This research was supported by the National Science Foundation, CMMI-1760102.

We introduce a novel criterion in clustering that seeks clusters with limited *range* of values associated with each cluster's elements. In clustering or classification the objective is to partition a set of objects into subsets, called clusters or classes, consisting of similar objects so that different clusters are as dissimilar as possible. We propose a number of objective functions that employ the range of the clusters as part of the objective function. Several of the proposed objectives mimic objectives based on sums of similarities. These objective functions are motivated by image segmentation problems, where the diameter, or range of values associated with objects in each cluster, should be small. It is demonstrated that range-based problems are in general easier, in terms of their complexity, than the analogous similarity-sum problems. Several of the problems we present could therefore be viable alternatives to existing clustering problems which are NP-hard, offering the advantage of efficient algorithms.

KEYWORDS

clustering, range, image segmentation, complexity, minimum cut, partitioning

1 | INTRODUCTION

The typical clustering and classification problem is to partition a set of objects into subsets, called clusters, so that each subset consists of “similar” elements, and the different clusters are as dissimilar as possible. We introduce novel clustering criteria that seek clusters with limited *range* of values associated with each cluster's objects (or elements). Each of the objects to be clustered has a scalar value associated with it and the *range* of a cluster is the difference between the maximum and the minimum values of the elements in the cluster. We show that for a partition into k clusters the problem of minimizing the maximum range of a cluster and the problem of minimizing the sum of the ranges of the clusters are solvable in polynomial time. Other problems explored here are the k -normalized range sum, the k -range cut and the k -normalized range cut, and we provide particularly efficient algorithms for the first two and demonstrate that the last one is NP-complete.

A good model for separation between clusters is the minimum cut, applied to the graph with nodes representing objects, and edges between pairs of nodes associated with weights of the similarity between the respective nodes. The minimum 2-cut problem is to find a partition of the graph into two nonempty components so that the sum of similarity weights on edges with one endpoint in one cluster and the second endpoint in the other cluster, is minimum. Thus the similarity between the two resulting clusters of the bipartition is minimum. This notion of cut extends to multiple clusters, as in the minimum k -cut problem where the objective is to find a partition into k nonempty components so that the sum of weights of edges with endpoints in different components, or the inter-similarity, is minimum [7]. The concept of a cut plays an important role in image segmentation problems, where the goal is to partition the image into meaningful objects. The problems of k -range cut and k -normalized range cut combine the range objective with a form of the cut objective. The problems' formulations are given in Tables 1 and 2, and the complexity results and run times of the algorithms are given in Table 3.

TABLE 1 Formulations of bipartition range-clustering problems

Problem name	Formulation
min range sum	$\min_{\emptyset \subset S \subset V} \text{range}(S) + \text{range}(\bar{S})$
min max range	$\min_{\emptyset \subset S \subset V} \max\{\text{range}(S), \text{range}(\bar{S})\}$
min normalized range sum	$\min_{\emptyset \subset S \subset V} \frac{\text{range}(S)}{f(S)} + \frac{\text{range}(\bar{S})}{f(\bar{S})}$
min range cut	$\min_{\emptyset \subset S \subset V} \text{range}(S) + \text{range}(\bar{S}) + C(S, \bar{S})$
min normalized range cut	$\min_{\emptyset \subset S \subset V} \frac{\text{range}(S)}{f(S)} + \frac{\text{range}(\bar{S})}{f(\bar{S})} + C(S, \bar{S})$

TABLE 2 Formulations of k -partition range-clustering problems for $k \geq 3$

Problem name	Formulation
min k -range sum	$\min_{(S_1, \dots, S_k)} \sum_{i=1}^k \text{range}(S_i)$
min max k range	$\min_{(S_1, \dots, S_k)} (\max_{i \in \{1, \dots, k\}} \text{range}(S_i))$
min k -normalized range sum	$\min_{(S_1, \dots, S_k)} \sum_{i=1}^k \frac{\text{range}(S_i)}{f(S_i)}$
min k -range cut	$\min_{(S_1, \dots, S_k)} \sum_{i=1}^k \text{range}(S_i) + \sum_{i=1}^{k-1} \sum_{j=i+1}^k C(S_i, S_j)$
min k -normalized range cut	$\min_{(S_1, \dots, S_k)} \sum_{i=1}^k \frac{\text{range}(S_i)}{f(S_i)} + \sum_{i=1}^{k-1} \sum_{j=i+1}^k C(S_i, S_j)$

TABLE 3 Summary of the range-clustering problems and algorithmic results

min problem	Bipartition, $k = 2$	k -Partition, $k \geq 3$
k -Range sum	Polynomial time, $O(n)$	Polynomial time, $O(n)$
max k -range	$O(\log n)$	Polynomial time, $O(n \log^3 n)$
k -Normalized range sum ^a	Polynomial time, $O(n)$	Polynomial time, $O(n^2 k)$
k -Range cut	Polynomial time, $O\left(mn^2 \log \frac{n^2}{m}\right)$	Polynomial $O(n^{k^2})$ for k fixed NP-complete for general k
k -Normalized range cut	NP-complete, even if $f(S) = S $	NP-complete, even if $f(S) = S $

^a $f(|S|)$ is monotone nondecreasing in $|S|$.

The motivation for our study of range problems originates in image segmentation. In a typical image segmentation set-up there are similarity weights assigned to each pair of *adjacent* pixels (which are the objects for the image clustering/segmentation) (9, 18). However, it is often the case that each pixel has in addition some scalar value associated with it, such as its color intensity, or its texture (computed with respect to a neighborhood of the pixel). The goal in this type of clustering is then not only to have the pixels similar to each other, but also to have the scalar values associated with pixels in the same cluster close enough to each other. This is the case, for instance, in segmenting knee cartilage as in [14], where pixels of cartilage tissue have a distinctive texture. The goal then is to ensure that all pixels within each segmented object have only a limited variability in their range of texture values.

We devise a family of *range-based* clustering problems that are analogous to commonly considered goals in clustering. We demonstrate that, in general, range problems are easier to solve (in terms of complexity) than their respective total similarity problems. One such problem is the NP-hard *normalized cut* problem [18], defined below in Equation 1.2. In contrast, we demonstrate here that an analogous range objective problem similar to normalized cut is polynomial time solvable.

The term range-clustering was previously used in [17]. However, the context there is to provide an improved computation of similarities, rather than to generate meaningful clusters as is the case here. We believe that here is the first time that the concept of range is utilized as a clustering criterion.

To formalize the discussion and problem definitions, we introduce relevant graph notation and other preliminaries.

1.1 | Notations and preliminaries

Let $G = (V, E)$ be an undirected graph where the nodes of the graph correspond to elements (also referred to as objects) to be clustered. We use the standard notation of n for the number of nodes $|V|$ and m for the number of edges $|E|$. We refer to the elements as the set $\{v_1, \dots, v_n\}$ and the corresponding nodes in the graph as $\{1, \dots, n\}$. There are edge weights w_{ij} associated with each edge $[i, j] \in E$ representing the “similarity” of nodes i and j . This similarity weight is in turn equivalent to the penalty of not assigning the respective pair of elements v_i and v_j to the same cluster. Higher similarity is associated with higher weights. A bi-partition of a graph is called a *cut*, $(S, \bar{S}) = \{[ij] | i \in S, j \in \bar{S}\}$, where $\bar{S} = V \setminus S$. The *capacity of a cut* (S, \bar{S}) is the sum of

the weights of the edges, with one endpoint in S and the other in \bar{S} : $C(S, \bar{S}) = \sum_{i \in S, j \in \bar{S}, [i, j] \in E} w_{ij}$. A minimum capacity cut, or minimum cut, is a bipartition of the nodes into two nonempty sets (S, \bar{S}) that minimizes $C(S, \bar{S})$. A bipartition resulting from a minimum cut has S and \bar{S} as dissimilar as possible in terms of the sum of similarities between their elements.

Two versions of the minimum cut problem are 2-cut and s, t -cut. The former is the bipartition of V into two nonempty sets, S and its complement \bar{S} , (S, \bar{S}) . For designated nodes $s, t \in V$, (S, \bar{S}) is said to be an s, t -cut, if $s \in S$ and $t \in \bar{S}$. A minimum 2-cut or s, t -cut are those bipartitions that minimize $C(S, \bar{S})$. When there is no ambiguity, we will simply refer to the *min-cut*.

A partition of the set of elements, V , into more than two sets, k -partition, is a collection of $k \geq 3$ disjoint nonempty sets $\{S_i\}_{i=1}^k$ so that $\cup_{i=1}^k S_i = V$. A k -partition is denoted by (S_1, \dots, S_k) . We refer to a partition into k clusters as k -clustering.

The *weighted degree* of node i is the sum of weights adjacent to i , $d_i = \sum_{j|[i, j] \in E} w_{ij}$. The weight of a subset of nodes $B \subseteq V$, referred to as the *volume* of B , is the sum of weighted degrees of nodes in B , $d(B) = \sum_{i: v_i \in B} d_i$.

The concept of *shrinking* of nodes is utilized here. Consider a graph $G = (V, E)$ with edge weights w_{ij} associated with each edge $[i, j] \in E$, and a specific pair of nodes p and q . The process of shrinking node p into q is to remove node p from the graph, and appending to node q all edges formerly adjacent to p . This can be done in two equivalent methods. In the first, we let $N(p) = \{v|[p, v] \in E\}$, remove the set of edges $\{[p, v]|v \in N(p)\}$ and add the set of edges $\{[q, v]|v \in N(p)\}$ with their respective weights. The second method is to simply add an edge $[p, q]$ of infinite capacity (weight). This ensures that p and q are always together in the same set when the minimum cut criterion is used.

The input to range-clustering problems includes scalar values associated with the elements. Respectively, the input graphs for range-clustering include node weights that are distinct scalar values associated with the nodes of V , $\{\alpha_1, \dots, \alpha_n\}$, where α_i are rational numbers. These scalars are necessarily rational and of finite number of significant digits (so the input is finite). The assumption of distinct scalar values is important for the algorithms presented here. This is because equal values of the scalars may lead to an exponential number of equal-valued solutions, all of which may have to be explored by the algorithms. This assumption of distinct scalar values is however shown next to hold without loss of generality.

If there are equal-valued scalars, a standard *perturbation process* is applied: This is done by adding different powers of a small enough ϵ to each of the equal values. (Such a process is applied in linear programming algorithms to avoid degeneracy and cycling.) For instance, the value of ϵ can be selected to be the smallest resolution of any α_i . That is, multiply all α_i by a large enough number, say M , so they are all integers, and then set $\epsilon = \frac{1}{2}$. For $\alpha_i = \alpha_j$, we set their perturbed values to $\tilde{\alpha}_i = \alpha_i + \epsilon^i$ and $\tilde{\alpha}_j = \alpha_j + \epsilon^j$. The perturbed values are then all distinct, and it is easy to see that an optimal solution to the perturbed problem, in terms of the range, is one of the optimal solutions to the unperturbed problem. If the problem involves similarity weights as well, the values of ϵ would depend on the smallest resolution among the scalar values of α as well as the values of w_{ij} .

Therefore, it is assumed without loss of generality that $\alpha_1 < \alpha_2 < \dots < \alpha_n$. We let α_i be associated with the element $v_i \in V$ so that v_1 is the element with the smallest value α_1 and v_n is the element with the largest value α_n .

For any nonempty set of elements $B \subseteq V$ the maximum, minimum and range of B are defined as:

$$\begin{aligned} \max(B) &= \max_{i: v_i \in B} \{\alpha_i\} \\ \min(B) &= \min_{i: v_i \in B} \{\alpha_i\} \\ \text{range}(B) &= \max(B) - \min(B). \end{aligned}$$

Note that the range of an empty set is undefined, and is not relevant here. This is because we are interested only in partitions of the set of elements into nonempty clusters of bounded range. Hence, an empty cluster is not considered. Note also that the range of a singleton is 0.

We say that a set $S \subseteq V$ is a *subset of an interval* $I = [a, b]$ if $S \subseteq \{j|a \leq \alpha_j \leq b\}$ and $\min(S) = a$, $\max(S) = b$. We denote “ S a subset of I ” by $S \subseteq I$.

The complexity model used here is the real computation model which allows arithmetic operations on real numbers, regardless of the number of significant digits, to count as a single operation [1].

We next introduce our range-based clustering problems and review related known clustering objectives that utilize similarities. Highlights of the differences in complexity between the range-based and known clustering objective are discussed.

1.2 | Range-based clustering problems and related clustering problems utilizing cuts and similarities

We introduce here a new collection of range-based problems and discuss related known clustering problems that utilize cuts and similarities. First we present 2-clustering range problems. The list of the names and formulations of the range-based clustering problems for bipartitioning problems is given in Table 1.

The simplest case of bipartition range-based problems considered is the min range sum. This objective function seeks to minimize the range of S and \bar{S} simultaneously, and it is shown here to be solvable in polynomial time. A weighted version of

the problem is the min weighted range sum which permits one to emphasize the limited range of S , more than that of \bar{S} . The min weighted range sum is also solved in polynomial time, as shown in section 2. The min max range problem is to minimize the bottleneck range between the two sets of the bipartition.

Many commonly used clustering models utilize the notion of minimum cut. The input to such problems is a graph $G = (V, E)$ and similarity weights associated with the edges. Bipartition clustering is to partition the set of elements V into two nonempty disjoint sets, S and its complement \bar{S} . The capacity, or weight, of the cut between S and \bar{S} , $C(S, \bar{S})$, signifies the degree of similarity between S and \bar{S} . To generate a set S that is highly dissimilar to its complement one seeks a minimum cut partition into S and \bar{S} that minimizes $C(S, \bar{S})$. This in turn also maximizes the total similarity within the two sets.

It has long been observed that a minimum cut in a graph with edge similarity weights tends to create a bipartition that has one side very small in size, containing a singleton in extreme cases [19]. This is so because the number of edges between a single node and the rest of the graph tends to be much smaller than between two comparable-sized sets. To correct for such unbalanced bipartitions, Shi and Malik, in the context of image segmentation [18], proposed the *normalized cut* as an alternative criterion to minimum cut. The normalized cut (NC) optimization problem is to find a bipartition of V , (S, \bar{S}) , minimizing:

$$(NC) \quad \min_{\emptyset \subset S \subset V} \frac{C(S, \bar{S})}{d(S)} + \frac{C(S, \bar{S})}{d(\bar{S})}. \quad (1.1)$$

The normalized cut problem (NC) was shown to be NP-hard in [18] by a reduction from set partitioning. Because set partitioning is weakly NP-hard, this only proves that normalized cut is at least weakly NP-hard. The problem is however strongly NP-hard with a reduction from the *balanced cut problem*, which is sketched below for an easier problem. The essence of the difficulty of NC derives from the fact that in the objective function of NC, the one ratio term with the smaller value of $d()$ is at least $\frac{1}{2}$ of the objective value. Therefore, this objective function drives the segment S and its complement to be approximately of equal sizes. Indeed, it is shown in [12] that the problem of minimizing the first term of NC, $\min_{\emptyset \subset S \subset V} \frac{C(S, \bar{S})}{d(S)}$, is polynomial time solvable.

The following problem is also a form of normalizing the cut with respect to the size of the sets. This problem is associated with finding the graph expander ratio and is known to be NP-hard [13]. It is referred to as *size-normalized cut*.

$$(size-NC) \quad \min_{\emptyset \subset S \subset V} \frac{C(S, \bar{S})}{|S|} + \frac{C(S, \bar{S})}{|\bar{S}|} \quad (1.2)$$

Note that like NC, the objective function of size-NC drives the segment S and its complement to be approximately of equal sizes.

The balanced cut problem is to find a bipartition cut of minimum weight such that both sets in the bipartition contain half of the nodes of the graph, (or $\min\{|S|, |\bar{S}|\} \leq c |V|$ for a constant $c \in (0, \frac{1}{2})$). This problem is also known as *minimum cut into bounded sets*, proved NP-complete in [5]. For an edge weighted graph G with total sum of edge weights equal to M , we scale all the edge weights by M . Since the numerator then is very small, an optimal solution is attained for half the nodes in the source set and the other half in the sink set, while minimizing the cut value, which is an optimal solution for the respective balanced cut problem. Therefore the size-NC is a strongly NP-hard problem.

The min normalized range sum is a range-based objective analogous to NC and size-NC. Unlike the normalized cut problem, this problem is shown here to be polynomial time solvable for $f(|S|)$ monotone nondecreasing in $|S|$.

For the next two range problems, min range cut and min normalized range cut, the input consists of a graph $G = (V, E)$, edge weights (similarities) w_{ij} for all $[i, j] \in E$, and scalars α_i associated with each node $i \in V$. The problems min range cut and min normalized range cut are generated by adding a minimum cut capacity term to the objective functions of min range cut and min normalized range cut respectively. The min range cut problem is shown here to be polynomial time solvable, whereas the min normalized range cut is proved to be NP-hard, even for $f(|S|) = |S|$. Thus in the latter problem, the addition of the minimum cut term to the objective changes its status from a polynomial problem to an NP-hard problem.

A sample of k -clustering problems for $k \geq 3$ that utilize similarities include:

1. The minimum k -cut problem seeks a k partition so that the total similarity between all k clusters is minimized. This similarity is measured by the sum of weights of edges that have endpoints in different clusters. The minimum k -cut problem is NP-hard, but can be solved in polynomial time for fixed k [7]. The minimum 2-cut problem is a polynomial time solvable special case of the k -cut problem where $k = 2$.
2. Clustering into k subsets of constrained size: Feo et al. [3] considered this problem in the context of VLSI (Very Large Scale Integrated Circuits) design application and gave a heuristic solution. This problem is NP-hard even for partition into *two* clusters. In this formulation, the requirement that the size of the sets is bounded, say by a fraction (such as half) of the total number of elements, makes the problem equivalent to the balanced cut problem, which is NP-hard.
3. The extension of normalized cut (NC) to k -clustering is $\min_{(S_1, S_2, \dots, S_k)} \sum_{i=1}^k \frac{C(S_i, \bar{S}_i)}{d(S_i)}$, which is obviously NP-hard as well.

The respective formulations of k -partition range-based problems for $k \geq 3$, that generalize the 2-clustering objectives, are presented in Table 2. The min k -normalized range sum problem is analogous to the k -normalized cut. The problems of min k -range cut and min k -normalized range cut are generated by adding the k -cut objective to the respective objectives of min k -range sum and min k -normalized range sum respectively.

Among our results for k -clustering range problems, the min max k range, min k -range sum and min k -normalized range sum problems are shown to be polynomial time solvable. The min k -normalized range cut problem is proved to be NP-hard. The min k -range cut problem is polynomial for fixed k , but proved to be NP-hard for general k . The complexity results are summarized in Table 3 under the assumption that the input scalar values are given sorted. This assumption is made in order to highlight the fact that the running time for solving some of the problems is faster than that required to sort the n values, $O(n \log n)$.

1.2.1 | Paper overview

The remainder of the paper is organized as follows: sections 2 to 4 present polynomial time algorithms for solving the min range sum, the min max range, and the min normalized range sum problems, respectively. In section 5 we describe a polynomial time algorithm for solving the min range cut problem which uses a parametric cut procedure. In section 6 the min normalized range cut problem is proved to be NP-hard. Section 7 provides algorithms and associated complexity results for the respective k -clustering problems for $k \geq 3$. An experiment to test the applicability of a range-based algorithm for image segmentation tasks is provided in section 8. Concluding remarks are given in section 9.

2 | THE MINIMUM RANGE SUM PROBLEM

We let $Z(S)$ be the objective value for the min range sum problem with the set S , and let S^* be an optimal solution to the problem:

$$Z(S^*) = \min_{\emptyset \subset S \subset V} Z(S) = \min_{\emptyset \subset S \subset V} \text{range}(S) + \text{range}(\bar{S}).$$

Using the definition of $\text{range}(\cdot)$, this problem may be written as:

$$\min_{\emptyset \subset S \subset V} \max(S) - \min(S) + \max(\bar{S}) - \min(\bar{S})$$

Recall that the elements are indexed according to their respective values of α , $\alpha_1 < \dots < \alpha_n$. Therefore element v_1 is associated with α_1 and element v_n with α_n . Without loss of generality, we can assume $v_1 \in S$; thus $\min(S) = \alpha_1$. The next lemma proves that in an optimal solution the element with the largest value, v_n , belongs to \bar{S} :

Lemma 1. *In every optimal solution to the min range sum problem, $v_1 \in S$ and $v_n \in \bar{S}$.*

Proof. If $v_1 \in \bar{S}$, then S and \bar{S} may be interchanged with no change to the value of the objective function. When $v_1, v_n \in S$, the objective value can be no lower than $\alpha_n - \alpha_1$. However, the solution $S' = V \setminus v_n$ and $\bar{S}' = v_n$ has an objective value that is at most $\alpha_{n-1} - \alpha_1$ with $v_n \in \bar{S}'$. So a solution that does not contain both v_1 and v_n in S is strictly better. ■

Therefore, we can assume that $\max(\bar{S}) = \alpha_n$ and $\min(S) = \alpha_1$. A more general lemma proves that there exists an optimal solution with $\max(S) \leq \min(\bar{S})$ (Figure 1):

Lemma 2. *There is an optimal solution to the min range sum problem with $\max(S) \leq \min(\bar{S})$.*

Proof. Suppose not. Let (S, \bar{S}) be a feasible solution with $\max(S) = \alpha_j > \alpha_\ell = \min(\bar{S})$. Thus $1 < \ell < j < n$. Construct another feasible solution by letting $S' = S \setminus \{v_j, v_{j-1}, \dots, v_\ell\}$ and $\bar{S}' = \bar{S} \cup \{v_j, v_{j-1}, \dots, v_\ell\}$. The solution (S', \bar{S}') is feasible since S' and \bar{S}' are nonempty. Note that the maximum and minimum elements of S' are the same as those of S , and thus the range for both is the same. However, the objective value $Z(S')$ can only be smaller than $Z(S)$:

$$\begin{aligned} Z(S') &= \max(S') - \min(S') + \max(\bar{S}') - \min(\bar{S}') = \alpha_{\ell-1} - \alpha_1 - (\alpha_n - \alpha_\ell) = \\ &= \max(S) - (\alpha_j - \alpha_{\ell-1}) - \min(S) + \max(\bar{S}) - \min(\bar{S}) = \\ &= Z(S) - (\alpha_j - \alpha_{\ell-1}) \leq Z(S). \end{aligned}$$

Thus there is an optimal solution in which $\text{range}(S)$ and $\text{range}(\bar{S})$ form two nonoverlapping intervals. ■

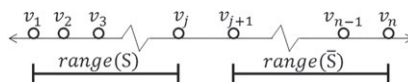


FIGURE 1 The elements of V arranged on the real line with $\text{range}(S)$ and $\text{range}(\bar{S})$ for $S = \{v_1, v_2, \dots, v_j\}$ and $\bar{S} = \{v_{j+1}, v_{j+2}, \dots, v_n\}$

Consider the elements of V arranged on the real line, where each element v_i is placed at position α_i . We observe that the largest gap between two consecutive elements plays a role in the min range sum partition:

Corollary 1. If p^* corresponds to the index of the largest value in the set of gaps $\mathcal{G} = \{\alpha_{p+1} - \alpha_p, p = 1, \dots, n\}$, then an optimal solution to the *min range sum* problem is $S = \{v_1, v_2, \dots, v_{p^*}\}$ and $\bar{S} = \{v_{p^*+1}, v_{p^*+2}, \dots, v_n\}$ with an optimal value of $\alpha_{p^*} - \alpha_1 + \alpha_n - \alpha_{p^*+1}$.

Proof. It was shown in Lemma 1 that $\min(S) = \alpha_1$ and $\max(\bar{S}) = \alpha_n$. Thus the objective is $Z(S^*) = \min_{\emptyset \subset S \subset V} \max(S) - \alpha_1 + \alpha_n - \min(\bar{S})$. This problem is then equivalent to the maximization problem: $\max_{\emptyset \subset S \subset V} \min(\bar{S}) - \max(S)$ which is to maximize the gap between the largest element of S and the smallest element of \bar{S} as stated. ■

Finding the largest gap requires $O(n)$ arithmetic operations and comparisons. This is generalized for k -partitions in section 7.1 with the same run time, albeit with a more involved algorithm.

We comment that the algorithm for min range sum can be extended to a weighted version of the problem, *weighted minimum range sum*. The *weighted minimum range sum* problem, for $\gamma \in (0, 1)$, is:

$$\min_{\emptyset \subset S \subset V} Z(S) = \min_{\emptyset \subset S \subset V} \text{range}(S) + \gamma \text{range}(\bar{S}).$$

This weighted problem is also solved in $O(n)$ time. This is because Lemma 2 holds also for the weighted variant. The algorithm is to compare the n values of $\alpha_n - \alpha_{p+1} + \gamma(\alpha_p - \alpha_1)$ and select the smallest.

3 | MIN MAX RANGE

Recall that the problem of min max range is $\min_{\emptyset \subset S \subset V} \max\{\text{range}(S), \text{range}(\bar{S})\}$. This problem is solved relatively easily: Let $\alpha\left(\frac{1}{2}\right) = \frac{\alpha_1 + \alpha_n}{2}$. Next, identify the index i^* such that,

$$\alpha_{i^*} \leq \alpha\left(\frac{1}{2}\right) < \alpha_{i^*+1}.$$

We now let $S = \{1, \dots, i^*\}$ and $\bar{S} = \{i^* + 1, \dots, n\}$. The minimum of $\max\{\text{range}(S), \text{range}(\bar{S})\}$ is attained for $\max\{\alpha_{i^*} - \alpha_1, \alpha_n - \alpha_{i^*}\}$. Since finding i^* in the sorted array can be done with binary search in $O(\log n)$ steps, this is the complexity of solving the min max range problem.

4 | MIN NORMALIZED RANGE SUM

In the min normalized range sum problem, the range of each segment is divided by a monotone nondecreasing function of the number of elements in that segment. The problem is to find a partition (S, \bar{S}) that attains the minimum for the problem:

$$\min_{\emptyset \subset S \subset V} \frac{\text{range}(S)}{f(|S|)} + \frac{\text{range}(\bar{S})}{f(|\bar{S}|)}.$$

This can be rewritten as:

$$\min_{\emptyset \subset S \subset V} \frac{\max(S) - \min(S)}{f(|S|)} + \frac{\max(\bar{S}) - \min(\bar{S})}{f(|\bar{S}|)}.$$

We note that since the values of $\max(S), \min(S), \max(\bar{S}), \min(\bar{S})$ are all in the set $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$, all possible combinations may be enumerated in polynomial time. The proof of Lemma 3, stated next, is similar to that of Lemma 1:

Lemma 3. *There exists an optimal solution to the min normalized range sum problem with $v_1 \in S$ and $v_n \in \bar{S}$.*

Proof. If $v_1 \in \bar{S}$, then S and \bar{S} may be interchanged with no change to the value of the objective function. Due to monotonicity of $f(\cdot)$, and the nonemptiness requirement on S , we know $f(n-1) \geq f(|S|)$. Therefore, when $v_1, v_n \in S$, the objective value can be no lower than $\frac{\alpha_n - \alpha_1}{f(n-1)}$. This is attained for instance for the solution $S = V \setminus v_i$ and $\bar{S} = v_i$ for $1 < i < n$.

However the solution $S' = V \setminus v_n$ and $\bar{S}' = v_n$ has an objective value that is at most $\frac{\alpha_{n-1} - \alpha_1}{f(n-1)}$ with $v_n \in \bar{S}'$. ■

A generalization of Lemma 2 states that in an optimal solution the two intervals representing $\text{range}(S)$ and $\text{range}(\bar{S})$ do not overlap:

Lemma 4. For any feasible solution (S, \bar{S}) to the min normalized range sum problem with $\max(S) > \min(\bar{S})$, there is a feasible solution (S', \bar{S}') with $\max(S') \leq \min(\bar{S}')$ and a lower objective function value.

Proof. For a feasible solution (S, \bar{S}) with $\max(S) > \min(\bar{S})$, define $S' = \{v_1, \dots, v_{|S|}\}$ and $\bar{S}' = \{v_{|S|+1}, \dots, v_n\}$. Note that since $|S| = |S'|$ and $|\bar{S}| = |\bar{S}'|$, the denominators stay the same. However, $\text{range}(S) \geq \text{range}(S')$ and $\text{range}(\bar{S}) \geq \text{range}(\bar{S}')$, so the objective function value is lower for (S', \bar{S}') . ■

This implies that for an optimal solution (S, \bar{S}) with $\max(S) = \alpha_i$, $\min(\bar{S}) = \alpha_{i+1}$. It is therefore sufficient to enumerate the $n - 1$ nonoverlapping bipartitions in order to solve the min normalized range sum problem. With the element indexed i^* so that:

$$i^* = \arg \min_{i=1, \dots, n-1} \frac{\alpha_i - \alpha_1}{f(i)} + \frac{\alpha_n - \alpha_{i+1}}{f(n-i)},$$

the optimal solution is $S = \{v_1, \dots, v_{i^*}\}$ and $\bar{S} = \{v_{i^*+1}, \dots, v_n\}$ with an objective function value of $\frac{\alpha_{i^*} - \alpha_1}{f(i^*)} + \frac{\alpha_n - \alpha_{i^*+1}}{f(n-i^*)}$. The complexity of this algorithm is $O(n)$.

5 | MIN RANGE CUT

The min range cut problem is to find the partition (S, \bar{S}) that solves:

$$\min_{\theta \subset S \subset V} \text{range}(S) + \text{range}(\bar{S}) + C(S, \bar{S}).$$

As a result of adding the cut to the objective function, some of the properties from the previous sections no longer apply. For example, we may not assume that the elements with the largest and smallest values, v_1 and v_n , are in different clusters as the weight on the edge between them, $w_{1,n}$, could be infinite hence forcing them to be into the same cluster.

We associate with any partition into S and \bar{S} , two respective intervals, $I(S) = I_1 = [\min(S), \max(S)]$ and $I(\bar{S}) = I_2 = [\min(\bar{S}), \max(\bar{S})]$. Since (S, \bar{S}) is a partition, the endpoints of these two intervals are four distinct values such that two of the four endpoints must be α_1 and α_n , and the remaining two endpoints, α_p and α_q , are selected from $\{\alpha_2, \dots, \alpha_{n-1}\}$. We say that the pair of intervals I_1 and I_2 is feasible if $\{\alpha_1, \alpha_2, \dots, \alpha_n\} \subseteq I_1 \cup I_2$. Given α_p and α_q for $2 \leq p < q \leq n - 1$, then, unless $p = q - 1$, there is one feasible choice of the two intervals, as $[\alpha_1, \alpha_q]$ and $[\alpha_p, \alpha_n]$. If $p = q - 1$ then both choices of the two intervals: the pair $[\alpha_1, \alpha_{p+1}]$, $[\alpha_p, \alpha_n]$, and the pair $[\alpha_1, \alpha_p]$, $[\alpha_{p+1}, \alpha_n]$, are feasible in that their union contains all the values. For the given selection of α_p and α_q there is a second choice of the interval pair $[\alpha_1, \alpha_n]$ and $[\alpha_p, \alpha_q]$. It follows that there are $O(n^2)$ feasible choices of I_1 and I_2 pairs.

Recall that a subset S of interval I is denoted by $S \subseteq I$. The min range cut problem is equivalent to the problem of finding a pair of feasible intervals minimizing the objective. We call this problem the min interval-range cut problem:

$$\min_{I_1, I_2 \text{ feasible}} \min_{\theta \subset S \subseteq I_1, \theta \subset \bar{S} \subseteq I_2} \text{range}(I_1) + \text{range}(I_2) + C(S, \bar{S}).$$

The min interval-range cut problem can be equivalently rewritten as:

$$\min_{I_1, I_2 \text{ feasible}} \text{range}(I_1) + \text{range}(I_2) + \min_{\theta \subset S \subseteq I_1, \theta \subset \bar{S} \subseteq I_2} C(S, \bar{S}).$$

Any feasible solution to min interval-range cut is a feasible solution for the min range cut problem and vice versa. It is possible that for a feasible pair of intervals, the objective value of the interval-range cut with an implied pair of sets S and \bar{S} has the range of S or the range of \bar{S} strictly greater than the range of the respective intervals and therefore larger than the respective objective value for range cut. This occurs when either one of the sets S and \bar{S} in the optimal solution for I_1 and I_2 , is strictly contained in either I_1 or I_2 and hence has smaller range. In that case the selection of I_1, I_2 , while feasible, is not optimal, as there exists another selection pair of feasible intervals I'_1, I'_2 , such that $I'_1 \subseteq I_1$, and $I'_2 \subseteq I_2$, with a strictly better objective value.

Proposition 1. The optimal solution of min range cut problem is identical to the optimal solution of min interval-range cut problem.

Proof. Let S^* and \bar{S}^* be an optimal solution to min range cut. Then, $I_1^* = [\alpha_1, \max(S^*)]$ and $I_2^* = [\min(\bar{S}^*), \alpha_n]$ are the arguments of the optimal solution to min interval-range cut. ■

In order to find an optimal solution for min range cut, it is therefore sufficient to enumerate all $O(n^2)$ pairs of feasible intervals, and for each pair of intervals I_1 and I_2 , to solve the respective min-cut problem $\min_{\theta \subset S \subseteq I_1, \theta \subset \bar{S} \subseteq I_2} C(S, \bar{S})$. We now elaborate on how to solve the min-cut problem for a given feasible pair of intervals I_1 and I_2 .

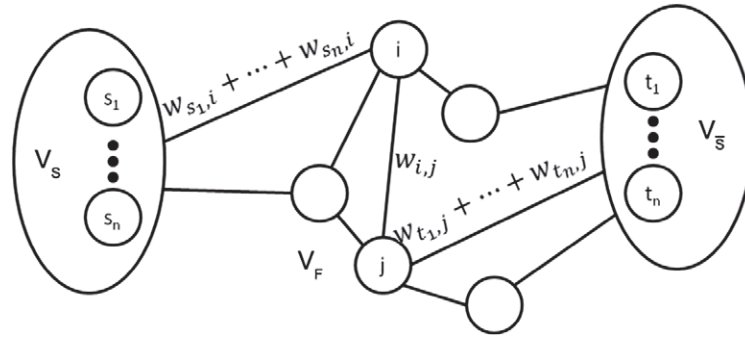


FIGURE 2 The s, t graph G' for $(V_S, V_{\bar{S}}, V_F)$ induced by the intervals pair I_1, I_2

The elements of V are partitioned into $(V_S, V_{\bar{S}}, V_F)$, as follows:

$$\begin{aligned} V_F &= \{i \in V \mid i \in I_1 \cap I_2\} \\ V_S &= \{i \in V \mid i \in I_1 \setminus V_F\} \\ V_{\bar{S}} &= \{i \in V \mid i \in I_2 \setminus V_F\}. \end{aligned}$$

By construction, any solution to the range cut problem for this given pair of intervals satisfies $V_S \subseteq S$ and $V_{\bar{S}} \subseteq \bar{S}$. Because the endpoints of the two intervals are distinct, both V_S and $V_{\bar{S}}$ are nonempty. It remains to partition V_F and allocate its elements to S and \bar{S} so that the cut value is minimized. Namely,

$$\min_{\emptyset \subset S \subseteq I_1, \emptyset \subset \bar{S} \subseteq I_2} C(S, \bar{S}) = \min_{S \subseteq V_F} C(S, \bar{S}).$$

In cases where $V_F = \emptyset$ the partition into S and \bar{S} is pre-determined and there is no need to solve the min-cut problem. This happens for the pair of intervals $[\alpha_1, \alpha_p], [\alpha_{p+1}, \alpha_n]$. We next construct a graph G' in which a minimum s, t -cut partition provides an optimal solution to the range cut problem for a given feasible pair of intervals.

Consider the graph $G = (V, E)$ with the edge weights w_{ij} associated with each edge $[i, j] \in E$. Let the graph $G' = (\{s, t\} \cup V_F, E_{s,t})$ be constructed as follows: The nodes corresponding to V_S are “shrunk” into a source node s , and the nodes corresponding to $V_{\bar{S}}$ are “shrunk” into a sink node t . The graph G' is illustrated in Figure 2. It is easy to see that shrinking a node i with the source node s is equivalent, in terms of min-cut in the resulting graph, to adding an arc (s, i) of infinite capacity from s to i . Similarly, shrinking a node j with the sink node t is equivalent to adding an arc (j, t) of infinite capacity from j to t .

A minimum s, t -cut in the graph G' provides an optimal partition of V_F into S and \bar{S} . Therefore, the min interval-range cut problem can be solved by enumerating all possible feasible pairs of I_1 and I_2 and for each finding the min s, t -cut partition as described above. Among all enumerated possibilities, the partition with the lowest objective value is the optimal solution to the min interval-range cut problem.

Let $T(m, n)$ be the complexity of a minimum s, t -cut procedure, on a graph with m arcs and n nodes. A straight-forward implementation of the algorithm would require $O(n^2)$ calls to the minimum s, t -cut procedure, one for each pair of intervals' selection, for a total complexity of $O(n^2 \cdot T(m, n))$.

It is shown next that this complexity can be reduced by a factor of $O(n)$ using a *parametric cut* procedure: The parametric flow, or parametric cut, problem is defined on a *parametric graph* where the source adjacent capacities and the sink adjacent capacities are functions of a parameter; the source adjacent capacities are monotone nondecreasing in the parameter; and the sink adjacent capacities are monotone nonincreasing in the parameter. All other arcs in the graph have fixed capacities. The parametric flow problem, and the parametric cut problem which is the focus here, is to solve the maximum flow problem for a list of q parameter values, or for all values of the parameter within an interval of length U . The parametric cut problem for a list of q values was shown to be solved in the complexity of a single cut plus $O(qn)$ for the Push-relabel and Hochbaum's PseudoFlow (HPF) algorithms (4, 11). Both these max-flow min-cut algorithms have complexity $T(m, n) = O\left(mn \log \frac{n^2}{m}\right)$. (There are other implementations of these algorithms with different complexities as well, e.g., $O(n^3)$ (6, 15).) Therefore the complexity of solving the parametric cut for a sequence of q parameter values is $O(T[m, n] + qn)$, where the term qn accounts for the updates of the source and sink adjacent capacities.

5.1 | Comment

The complexity of solving the parametric cut (or flow) problem for all parameter values in an interval of length U was shown in (4, 11) to be $O(T(m, n) + n \log U/\varepsilon)$ for values determined with ε accuracy (within a \mathbb{L}_∞ distance of ε from the optimal solution). Although the text in [4] claims the complexity of the algorithm to be $O(T(m, n))$, the actual complexity is as stated here.¹

Algorithm 1 Min Range Cut Algorithm

INPUT: An ordered array $\{\alpha_1, \dots, \alpha_n\}$

OUTPUT: (S^*, \bar{S}^*) , and the corresponding objective value z^*

begin:

$\min(I_1) := \alpha_1; z^* := \infty$

$\max(I_2) := \alpha_n$ %checking all possible $\max(I_1)$ and $\min(I_2)$

for $i = 1, \dots, n - 1$ **do**

$\max(I_1) := \alpha_i$.

Create graph $G(i) = (V(i), E(i))$: set node 1 as source node s , shrink nodes $i + 1, \dots, n$ into

sink t ; arc capacities for $(p, q) \in E(i)$ are $w_{pq}^{(1)}$.

for $j = 2, \dots, i - 1$ **do**

$\min(I_2) := \alpha_j$;

%Shrink node j into s

$$w_{pq}^{(j)} = \begin{cases} w_{pq}^{(j-1)} & \text{if } (p, q) \neq (s, j) \\ \infty & \text{if } (p, q) = (s, j). \end{cases}$$

end for

Call PARAMETRIC CUT($G(i), \{w_{sp}^{(j)}, w_{qt}^{(j)}\}_{p, q \in V(i) \setminus \{s, t\}, j = 2, \dots, i-1}$); return S_2, \dots, S_{i-1}

set $j^* = \arg \min_j z(j) = (\alpha_i - \alpha_1) + (\alpha_n - \alpha_j) + C(S_j, \bar{S}_j)$.

if $z(j) < z^*$ **then**

$(S^*, \bar{S}^*) := (S_{j^*}, \bar{S}_{j^*})$

$z^* := z(j)$

end if

end for

$\max(I_1) := \alpha_n$ % for $I_1 = [\alpha_1, \alpha_n]$ checking all possible $\max(I_2)$ and $\min(I_2)$

for $i = 2, \dots, n - 1$ **do**

$\max(I_2) := \alpha_i$.

Create graph $G(i) = (V(i), E(i))$: set node i as source node s , shrink nodes $1, i + 1, \dots, n$ into

sink t ; arc capacities for $(p, q) \in E(i)$ are $w_{pq}^{(1)}$.

for $j = 3, \dots, i$ **do**

$\min(I_2) := \alpha_j$.

%Shrink node $j - 1$ into s

$$w_{pq}^{(j)} = \begin{cases} w_{pq}^{(j-1)} & \text{if } (p, q) \neq (s, j - 1) \\ \infty & \text{if } (p, q) = (s, j - 1). \end{cases}$$

end for

Call PARAMETRIC CUT($G(i), \{w_{sp}^{(j)}, w_{qt}^{(j)}\}_{p, q \in V(i) \setminus \{s, t\}, j = 3, \dots, i}$); return S_2, \dots, S_{i-1}

set $j^* = \arg \min_j z(j) = (\alpha_n - \alpha_1) + (\alpha_i - \alpha_j) + C(S_j, \bar{S}_j)$.

if $z(j) < z^*$ **then**

$(S^*, \bar{S}^*) := (S, \bar{S})$

$z^* := z(j)$

end if

end for

end

¹The term $n \log U$ cannot be removed from the complexity as proved in (8, 10, 11). Therefore the complexity must depend on the term $\log U$ and hence cannot be strongly polynomial.

Consider the parameter list $\{1, 2, \dots, q\}$, and an s, t graph G with arc (or edge) capacities $\{w_{ij}^\ell\}$ that are functions of a parameter ℓ . An s, t graph G is a parametric graph if the arc capacities are functions of a parameter, and satisfy, for $i, j \neq s, t$:

$$\begin{aligned} w_{si}^\ell &\leq w_{si}^{\ell+1} \\ w_{jt}^\ell &\geq w_{jt}^{\ell+1} \\ w_{ij}^\ell &= w_{ij}^{\ell+1}. \end{aligned}$$

Let the parametric graph G with arc capacities $\{w_{ij}^\ell\}$ for a given value of ℓ be denoted by G_ℓ . The parametric graph procedure takes as input, the graph G with arc weights w_{ij} and the parametric sets of source and sink adjacent arcs' capacities: parametric cut($G, \{w_{si}^\ell, w_{jt}^\ell\}_{i,j \in V \setminus \{s,t\}}, \ell = 1, \dots, q$) The procedure outputs S_ℓ , for $\ell = 1, \dots, q$, so that (S_ℓ, \bar{S}_ℓ) is a minimum cut in G_ℓ .²

Let the graph G_j be generated from graph G_{j-1} by shrinking one node v in G_{j-1} with the source. Since the shrinking of node v with the source is equivalent to adding an arc (s, v) of infinite capacity, then this process increases the capacity of the arc (s, v) adjacent to source from a finite value to infinity, while all other capacities remain constant. Hence the sequence of graphs generated by shrinking with the source, one node at a time, form a parametric graph. Also, the sequence of graphs generated by shrinking one node at a time with the sink node t , done by adding an infinite capacity arc between the node and t , form a parametric graph with sink adjacent capacities monotone nondecreasing and source adjacent capacities that are constant, and thus nonincreasing. The complexity of solving the min-cut for such a sequence of q graphs, $G_\ell, \dots, G_{\ell+q}$, with parametric cut is therefore $O(T(m, n))$ since $q < n$, and the $O(n^2)$ complexity is dominated by $T(m, n)$.

In the following theorem we show that Algorithm 1, solves the min range cut problem in $O(nT(m, n))$ steps.

Theorem 1. *Algorithm 1 solves the min range cut problem in $O(nT(m, n))$ steps.*

Proof. The correctness of the algorithm follows from its enumeration of all possible endpoints of intervals I_1 and I_2 and for each finding the min-cut partition of the elements in the overlap of the two intervals. ■

The enumeration of all endpoints is done in two parts: lines 4-18 deal with pairs of feasible intervals $I_1 = [\alpha_p, \alpha_n]$ and $I_2 = [\alpha_1, \alpha_q]$, and lines 20-34 deal with pairs of feasible intervals $I_1 = [\alpha_1, \alpha_n]$ and $I_2 = [\alpha_p, \alpha_q]$ where $I_1 \cap I_2 = I_2$.

The complexity of the Min Range Cut Algorithm is dominated by the two **for** loops, in steps 6 and 21, each of which consists of calling at most $n - 2$ times for the parametric min-cut procedure. Each call for parametric cut has complexity of $O(T(m, n))$. Therefore the complexity of the entire algorithm is $O(nT(m, n))$.

6 | MIN NORMALIZED RANGE CUT PROBLEM

Unlike min range sum, min normalized range sum and min range cut, the min normalized range cut problem is not polynomial time solvable. We demonstrate here the NP-hardness of the problem. In this section we slightly abuse terminology by referring to optimization problems as NP-complete meaning that their decision version is NP-complete (the correct term for optimization problems is NP-hard). All problems we address are clearly in NP and we will omit explicitly showing so.

Theorem 2. *The min normalized range cut problem, $\min_{\emptyset \subset S \subset V} \frac{\text{range}(S)}{|S|} + \frac{\text{range}(\bar{S})}{|\bar{S}|} + C(S, \bar{S})$, is NP-complete.*

Proof. To prove NP-completeness, we use Karp reductions in two steps to show that min normalized range cut α_P min inverse set size cut α_P balanced cut, where $Q_1 \alpha_P Q_2$ means that problem Q_1 is at least as hard as problem Q_2 , and equivalently, that problem Q_2 is polynomial time reducible to problem Q_1 . We first introduce these problems and then provide the NP-completeness proof. The balanced cut problem is a known NP-complete problem of finding the minimum 2-cut where the two sets in the bipartition are of equal size [5]. For a graph $G = (V, E)$ on an even number of nodes $n = |V|$, the balanced cut problem is formulated as:

$$\min_{S \subset V, |S| = \frac{n}{2}} C(S, \bar{S}).$$

■

²The code for HPF parametric cut is available at <http://riot.ieor.berkeley.edu/Applications/Pseudoflow/parametric.html>.

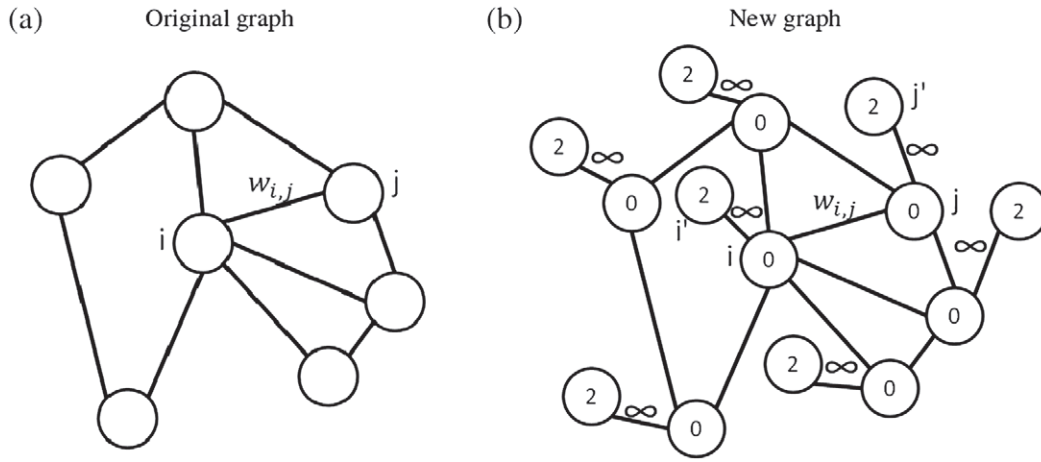


FIGURE 3 The original graph and the new graph in part 2's reduction. Numbers inside the nodes of the new graph are the values of the respective elements

The min inverse set size cut problem is used here as an intermediary problem. This problem is defined as:

$$\min_{\emptyset \subset S \subset V} \frac{1}{|S|} + \frac{1}{|\bar{S}|} + C(S, \bar{S}).$$

Part 1: We first show that balanced cut is reducible to min inverse set size cut: Given an instance of balanced cut defined on $G = (V, E)$ with edge weights w_{ij} . Define a new, scaled graph, in which the edge weights are $w'_{ij} = \frac{w_{ij}}{M}$, for some large number M . A suitable choice of M is $M = w_{\max} n^4$, where $w_{\max} = \max_{[i,j] \in E} w_{ij}$. We note that the minimum cut partition in a scaled graph is the same as the minimum cut partition as in the original graph, and the capacity of the scaled cut is $\frac{1}{M}$ times the capacity of the cut in the original graph. The min inverse set size cut on this scaled graph is:

$$\min_{\emptyset \subset S \subset V} \frac{1}{|S|} + \frac{1}{|\bar{S}|} + \frac{C(S, \bar{S})}{M}.$$

Since there are at most $O(n^2)$ arcs in a cut, then for our choice of the value of M , $\frac{C(S, \bar{S})}{M}$ is at most $O\left(\frac{1}{n^2}\right)$. Consequently the first two terms dominate the cut value term, $\frac{1}{|S|} + \frac{1}{|\bar{S}|} \gg \frac{C(S, \bar{S})}{M}$ by a factor of $O(n)$. Thus the optimal solution will necessarily minimize:

$$\min_{\emptyset \subset S \subset V} \frac{1}{|S|} + \frac{1}{|\bar{S}|},$$

for which the minimum is attained for $|S| = |\bar{S}| = \frac{n}{2}$. Among the solutions that minimize the first two terms, the min inverse set size cut problem minimizes the term $\frac{C(S, \bar{S})}{M}$, resulting in a minimum balanced cut. Thus, min inverse set size cut problem is NP-complete.

Part 2: We now demonstrate that min inverse set size cut is reducible to min normalized range cut: For a given problem instance of the min inverse set size cut problem with n nodes and m edges, we construct a new graph with $2n$ nodes and $m + n$ edges: Each original node v of V is connected to a new node v' with an edge of capacity ∞ . All original nodes are assigned a value of 0 and all new nodes are assigned a value of 2. See Figure 3. The presence of edges of infinite capacity guarantees that the range of both S and \bar{S} for any finite cut partition is exactly two, as otherwise the cut would have to have an infinite value. Also, for any finite cut partition in the original graph, the corresponding partition in the new graph has double the cardinality of S and \bar{S} and the same cut capacity. Thus, the solution to this new problem using min normalized range cut is exactly the solution to the original instance of the min inverse set size cut problem.

7 | RANGE SEGMENTATION IN k -PARTITIONS

In this section the bipartition results are extended to k -partitions. Specifically, we devise polynomial time algorithms for min k -range sum, min max k range, min k -normalized range sum and show that min k -normalized range cut is NP-complete. For any fixed value of k , the min k -range cut is shown here to be polynomial time solvable, as is the case for $k = 2$. But for arbitrary value of k we prove here that the min k -range cut is NP-hard. This follows since the problem generalizes the min k -cut problem which is NP-hard for arbitrary k [7].

7.1 | Min k -range sum problem

The min k -range sum problem is to find a partition (S_1, \dots, S_k) of V that minimizes $\sum_{i=1}^k \text{range}(S_i)$. This problem can be solved in linear time with Algorithm 2 as shown next.

Algorithm 2 Min k -Range Sum Algorithm

INPUT: An ordered array $\{\alpha_1, \dots, \alpha_n\}$

OUTPUT: $(S_1^*, S_2^*, \dots, S_k^*)$, such that $\cup_{i=1}^k S_i^* = V$ and the objective value z^* .

begin:

for $i = 1, \dots, n - 1$ **do**

$g_i := \alpha_{i+1} - \alpha_i$

end for

Let i_1, \dots, i_{k-1} be the indices of the largest $k - 1$ elements of $\{g_i\}_{i=1}^n$.

$(S_1^*, S_2^*, \dots, S_k^*) := (\{\alpha_1, \dots, \alpha_{i_1}\}, \{\alpha_{i_1+1}, \dots, \alpha_{i_2}\}, \dots, \{\alpha_{i_{k-1}+1}, \dots, \alpha_n\})$.

$z^* = \sum_{j=1}^k \text{range}(S_j^*) = \sum_{j=1}^k (\alpha_{i_j} - \alpha_{i_{j-1}+1})$, for $\alpha_{i_0+1} = \alpha_1$.

end

Proposition 2. Algorithm 2 solves the min k -range sum problem in $O(n)$ steps.

Proof. In an optimal solution $v_1 \in S_1$ and $v_n \in S_k$, and the sets (S_1, \dots, S_k) are nonoverlapping. These facts follow from the same arguments used in Lemmas 1 and 2 and are omitted for brevity. It remains to determine the boundaries between the nonoverlapping segments on the real line. This is equivalent to identifying the $k - 1$ largest gaps between consecutive values of α . This can be done by finding first the $(k - 1)$ st median in the set of $n - 1$ gaps, in linear time $O(n)$, using the algorithm of Blum et al. [2]. Once this median is found, the set of gaps is scanned once to mark all the gaps that have value greater or equal to that of the $(k - 1)$ st median. This produces the $k - 1$ largest gaps in linear time as shown in Algorithm 2. These gaps separate the sets in the partition with the smallest sum of ranges. ■

7.2 | Min max k range

The min max k range problem for $k > 2$ is solved differently and with significantly higher complexity than the case of $k = 2$ (in section 3). Recall that the min max k range problem is to find a partition (S_1, \dots, S_k) of V , that minimizes

$$\min_{(S_1, \dots, S_k)} (\max_{i \in \{1, \dots, k\}} \text{range}(S_i)).$$

As before, it is established, with the same arguments as in Lemma 1 and Lemma 2, that there is an optimal solution in which $v_1 \in S_1$ and $v_n \in S_k$, and that the intervals containing the sets (S_1, \dots, S_k) are nonoverlapping. Specifically, each set S_j in an optimal partition contains consecutive elements and is of the form $\{i_{j-1} + 1, \dots, i_j\}$ with range $\alpha_{i_j} - \alpha_{i_{j-1}+1}$.

Proposition 3. Algorithm 3 solves the min max k range problem in $O(n \log^3 n)$ steps.

Proof. The algorithm works by guessing one value for the max range at a time, and then conducting a *feasibility check* to verify whether there is a feasible solution for that value. The smallest value of a guessed range for which there is a feasible solution is the optimal range value. A natural way of implementing such an algorithm is by using binary search on the $\frac{n}{2}$ possible range values. Each possible range value is of the form $\alpha_q - \alpha_p$ corresponding to a pair p, q such that $1 \leq p < q \leq n$. Let the set of all possible range values be $\mathcal{D} = \{\alpha_q - \alpha_p \mid q > p, p = 1, \dots, n - 1, q = p + 1, \dots, n\}$. Note that we disregard the trivial case where the optimal solution is 0. The trivial case happens when the number of distinct scalar values is at most k . Since there are up to $\frac{n}{2}$ possible range values, these can be sorted, in $O(n^2 \log n)$ time (note that $\log \binom{n}{2}$ is $O(\log n)$). Let the sorted values in \mathcal{D} be $d_1 \geq d_2 \geq \dots \geq d_{\binom{n}{2}}$.

Consider the feasibility check for a given guessed value for the max range, z . To verify feasibility we first scan the values of $\alpha_1, \alpha_2, \dots, \alpha_n$, for the largest index j_1 so that $\alpha_{j_1} - \alpha_1 \leq z$. If $\alpha_2 - \alpha_1 > z$ then $j_1 = 1$. The interval $I_1 = [1, \alpha_{j_1}]$ is then the first of up to k intervals representing the k -partition of the set of values $\alpha_1, \alpha_2, \dots, \alpha_n$. Next we scan the values of $\alpha_{j_1+1}, \alpha_{j_1+2}, \dots, \alpha_n$ for the largest index j_2 such that $\alpha_{j_2} - \alpha_{j_1+1} \leq z$. Again, if $\alpha_{j_1+1} - \alpha_{j_1} > z$, then $j_2 = j_1$, corresponding to an interval that contains only one element, which is of range 0. The interval I_2 is then equal to $[\alpha_{j_1+1}, \alpha_{j_2}]$. This is repeated up to k times or until the last value α_n is reached. If after k repetitions $j_k < n$ then the guessed value z is not feasible and

therefore the optimal range value must be greater. Otherwise the guessed value is feasible and the optimal range value can only be smaller than z . This feasibility check runs in $O(n)$ steps as it scans the values of α at most once each.

We comment that there is an alternative feasibility check on the guessed value z that runs in $O(k \log n)$ steps. For α_{j_i} , the start of the i th interval, we search, using binary search on the set of α values, for the next value of α , $\alpha_{j_{i+1}}$, which is the largest while the difference from the current value of α is still less than z . Each such search requires $O(\log n)$ steps, and since there are up to k such intervals the total complexity is $O(k \log n)$. This complexity is faster than $O(n)$ if $k = o\left(\frac{n}{\log n}\right)$, but even then it does not improve the overall complexity since the other steps dominate it, as discussed next.

The optimal solution to the min max k range problem can then be found using binary search on the sorted sequence \mathcal{D} . This requires $O(\log n)$ calls for feasibility check for a total complexity of $O(n \log n)$ which is dominated by the complexity of the sorting of \mathcal{D} , $O(n^2 \log n)$. Next we show that the need to sort the distances in \mathcal{D} can be avoided, resulting in substantial speed-up from $O(n^2 \log n)$ down to $O(n \log^3 n)$.

Megiddo et al. [16] devised an efficient algorithm, called here the M-algorithm, for finding the i th longest path among the set of all simple paths in a tree with edge weights. For a tree on n nodes the complexity of the M-algorithm is $O(n \log^2 n)$. Note that the number of different simple paths in a tree is $O(n^2)$, since each simple path is uniquely characterized by its pair of endpoints. Consider a path graph on nodes $\{1, 2, \dots, n\}$, where all edges are of the form $[i, i+1]$ with weights $\alpha_{i+1} - \alpha_i$ for $i = 1, \dots, n-1$. A path graph is obviously a tree and thus the M-algorithm is applicable to this path graph. The distance between node p and node q , for $p < q$ is then $\alpha_q - \alpha_p$. Now, instead of sorting the distances in \mathcal{D} , we use the M-algorithm to identify the i th longest of the potentially feasible ranges which is the value z to be checked for feasibility. Since each call and feasibility check reduces the number of potentially feasible ranges by a factor of 2, the total number of calls is $\log \binom{n}{2}$, which is $O(\log n)$.

Initially the interval of integer indices $\left[1, \binom{n}{2}\right]$ contains the list of the index positions of all the potentially feasible ranges. At each iteration we find the median value in this interval of indices, without having the sorting available, by calling the M-algorithm. If this median range value is feasible, then we conclude that the min max feasible range can be only smaller and thus resides in the list of indices smaller than or equal to the median. Otherwise it resides in the list of larger indices. Initially the endpoints of the interval of indices are $a_{\min} = 1$ and $a_{\max} = \binom{n}{2}$. At each iteration the M-algorithm finds the median range value z in the interval, which is the $\left\lfloor \frac{1}{2}(a_{\min} + a_{\max}) \right\rfloor$ th longest in the original list. If z is feasible then a_{\max} is updated to be equal to this median index, otherwise a_{\min} is updated to be equal to this median index plus 1. The length of the interval is hence reduced by a factor of 2 at each iteration, thus requiring at most $\log \binom{n}{2}$ iterations of the procedure.

At each of the $O(\log n)$ calls for the guessed value of the range there is one call for finding the $\left\lfloor \frac{1}{2}(a_{\min} + a_{\max}) \right\rfloor$ th longest value z and one call for feasibility check of z . The first requires $O(n \log^2 n)$ steps and the second requires $O(n)$ steps. The total complexity is then $O(\log n \cdot (n \log^2 n + n))$. Thus, the optimal solution to the min max k range problem is computed in time $O(n \log^3 n)$. ■

The pseudocode of the algorithm solving min max k range is given as Algorithm 3.

7.3 | Min k -range cut

Recall that the min k -range cut problem is $\min_{(S_1, \dots, S_k)} \sum_{i=1}^k \text{range}(S_i) + \sum_{i=1}^{k-1} \sum_{j=i+1}^k C(S_i, S_j)$. For the range-cut (2-range cut) problem we reduced the problem to calls, for each configuration of interval partitioning, to a min cut procedure (section 5). The idea here is analogous, exploiting the use of the polynomial time minimum k -cut algorithm for k fixed of [7]. First, we note that for arbitrary k the NP-hardness of the minimum k -cut problem [7] implies that min k -range cut is NP-hard. This is easy to show by selecting the range of the values to be very small, by a factor of n at least, than the smallest weights of the graph. And then the value of any solution is dominated by the value of the k -cut partition.

The k -cut problem was shown to be solved in polynomial time for fixed k with the algorithm of Goldschmidt and Hochbaum [7] (GH-algorithm). The GH-algorithm involves $O(n^{O(k^2)})$ calls to a min s, t -cut procedure. For k not fixed the problem was shown to be NP-hard. The way the algorithm works is by guessing a set of “seeds” that must belong to one set in the partition, and a set of “seeds” that belong to the other $k-1$ sets. The algorithm calls for a respective min s, t -cut for the seeds in the set shrunk into the source set s , and the seeds for the other sets shrunk into the sink node t . The resulting source set is then considered to be one set in the k -partition and the process then continues, recursively, on the subgraph induced by the sink set,

Algorithm 3 Min Max k Range Algorithm

INPUT: An ordered array $\alpha = \{\alpha_1, \dots, \alpha_n\}$, in increasing order.
OUTPUT: $(S_1^*, S_2^*, \dots, S_k^*)$, and the corresponding objective value z^* .
begin:
 $a_{\min} := 1; a_{\max} := \binom{n}{2};$
while $a_{\min} < a_{\max}$ **do**
 $I_i := \emptyset, \forall i = 1, \dots, k;$
 $m := \lfloor \frac{1}{2}(a_{\min} + a_{\max}) \rfloor;$
 $z := m$ th largest value in the set $\mathcal{D} = \{\alpha_q - \alpha_p | q > p, p = 1, \dots, n-1, q = p, \dots, n\};$
(determined by the M-algorithm);
%Checking whether a feasible solution exists with value $\leq z$:
 $\min(I_1) := \alpha_1;$
 $\max(I_1) :=$ the largest $\alpha_j, j \geq 1$, such that $\alpha_j - \alpha_1 \leq z;$
for $i = 2, \dots, k$ **until** $\max I_{i-1} = \alpha_n$ **do**
 $\min(I_i) :=$ the smallest α_j , such that $\alpha_j > \max(I_{i-1});$
 $\max(I_i) :=$ the largest α_j , such that $\alpha_j \leq \min(I_i) + z;$
end for
if $\alpha_n \in \cup_{i=1}^k \{I_i\}$ **then**
 $a_{\max} := m;$
else
 $a_{\min} := m + 1;$
end if
end while
%Computing the output:
 $z^* := z;$
for $i = 1, \dots, k$ **do**
 $S_i^* := \cup_{v_j \in I_i} \{v_j\};$
end for
end

$k-1$ additional times, resulting in a k -partition. It was shown in [7] that it is sufficient to select seed sets that contain at most $O(k)$ seeds, and thus enumerating all of them takes time that is polynomial for fixed k .

Our algorithm for min k -range cut is a generalization of the min range cut algorithm, for $k=2$, and similarly works in two steps. In the first step select all possible feasible collections of k intervals, each corresponding to one set S_i , determined by the endpoints $[\min(S_i), \max(S_i)]$ (possibly $\min(S_i) = \max(S_i)$). Since the sets S_i form a partition, the endpoints of the respective intervals are distinct. Hence there are up to $2k-2$ distinct endpoints, and respectively up to $O(kn^{2k-2})$ interval configurations. A k -intervals configuration is feasible, if all values of α are contained in the union of the intervals. For each interval configuration that is feasible we let the seeds for the i th set in the partition S_i include $\min(S_i)$ and $\max(S_i)$ and all the nodes that correspond to values in the interval $(\min(S_i), \max(S_i))$ that are not in any of the other $k-1$ intervals. The resulting set of seeds is then augmented, if necessary, by the GH-algorithm, which is otherwise employed without change.

The total complexity involves then $O(kn^{2k-2})$ calls to the k -cut algorithm, each of complexity $O(n^{O(k^2)}T(m, n))$. Since $kn^{2k-2} \cdot n^{O(k^2)}$ is $O(n^{O(k^2)})$ the total complexity of solving the min k -range cut problem is $O(n^{O(k^2)}T(m, n))$.

7.4 | Min k -normalized range sum

The min k -normalized range sum problem is to find a k -partition, (S_1, \dots, S_k) , to achieve the following objective:

$$Q(n, k) = \min_{(S_1, \dots, S_k)} \sum_{i=1}^k \frac{\text{range}(S_i)}{f(|S_i|)}$$

We next present a polynomial time dynamic programming algorithm for the problem:

Proposition 4. The min k -normalized range sum problem is solvable in polynomial time $O(n^2k)$.

Proof. As before, it can be proved that there exists an optimal solution in which $v_1 \in S_1$ and $v_n \in S_k$, and that the sets (S_1, \dots, S_k) are nonoverlapping. The proof follows the same argument as in Lemmas 3 and 4 and is omitted.

The value of the objective function $Q(n, k)$ is the minimum *cost* required to partition n elements into k sets. For the ordered input elements according to $\{\alpha_1 < \dots < \alpha_n\}$, let $Q(p, j)$, be the minimum cost for a partition of the first p elements

of the input array into j sets, for $p \in \{1, \dots, n\}$ and $j \in \{1, \dots, k\}$. We construct a dynamic programming recursion with the boundary conditions: $Q(j, j) = 0 \forall j$, $Q(p, 1) = \frac{\alpha_p - \alpha_1}{f(p)} \forall p$, and $Q(i, j) = \infty \forall i < j$, with the latter being infeasible and therefore set to ∞ . The following recursion is used to calculate $Q(p, j)$ for $p, j > 1$, once the values of $Q(p', j')$ have been evaluated for all $p' < p$ and $j' < j$:

$$Q(p, j) = \min_{\ell \in \{j-1, \dots, p-1\}} \left(Q(\ell, j-1) + \frac{\alpha_p - \alpha_{\ell+1}}{f(p-\ell)} \right)$$

The rationale for the recursion is that optimal partitioning of p elements into j sets consists, for some value ℓ , of an optimal partitioning of elements $1, \dots, \ell$ into $j-1$ sets and allocating elements $\ell+1, \dots, p$ into the j th set. Since each recursion evaluation is accomplished in at most $O(p)$ steps, and there are $O(kn)$ function evaluations it follows that the optimal solution $Q(n, k)$ is determined by this dynamic programming procedure with complexity $O(n^2k)$. ■

7.5 | Min k -normalized range cut

In section 6, we proved this problem to be NP-complete in a bipartition setting. The following theorem follows by simply noticing that the bipartition problem is a special case of min k -normalized range cut problem, with $k=2$.

Theorem 3. *The min k -normalized range cut problem is NP-complete.*

8 | AN EXPERIMENT

In testing the applicability of range models we used one of the range models presented here on the source image in Figure 4A, from the Berkeley Segmentation Dataset and Benchmark (<http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/>). Our goal is to segment the white snow areas. This task is challenging for known image segmentation methods as they tend to select a single contiguous area. In particular, all cut related methods would not fully segment the snow areas as these areas are not contiguous and there are lengthy boundaries between them and the nonsnow areas. Algorithms that involve any form of cut, or attempt to achieve smooth and short boundaries between the segmented section and its complement, will not be able to attain a good segmentation of the snowy areas. This is because these boundaries contribute to increase the value of the cut separating the snow areas from the nonsnow areas. Therefore any min-cut based algorithm will tend to segment just one area of the snow and leave the others in the background. Applying the min range sum objective, however, results in a close to perfect segmentation that segments the snow together with very light surfaces and thus captures all the snow areas, as seen in Figure 4B.

We comment about the characteristics of each of the five range algorithms as per the prospective success in segmenting the image in Figure 4A. The segmentation sought is a 2-partition into the foreground set S and the background set \bar{S} .

1. min range sum: With the min range sum objective the entire area of the snow is captured since the range of the foreground set and its complement are the dominant factors in determining the optimal solution. This image has a distinct separation between light color areas and dark color areas. Therefore the largest gap is separating these two types of areas and delivers the desired output segmentation.

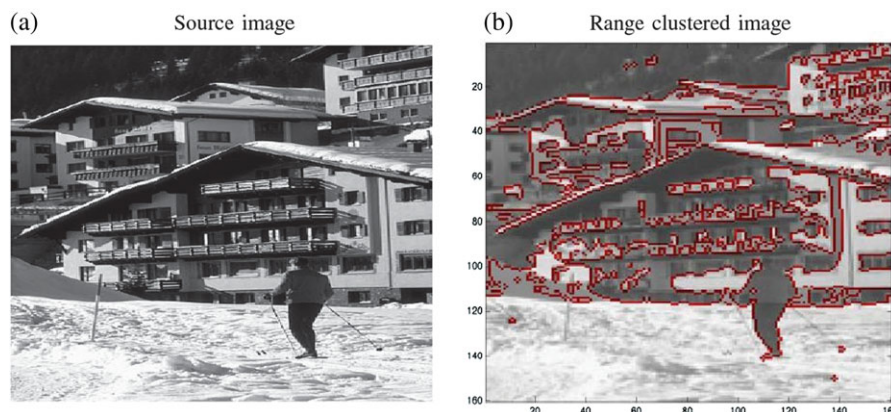


FIGURE 4 (a) Source image. (b) Range clustered image [Colour figure can be viewed at wileyonlinelibrary.com]

2. min max range: Because the focus here is just on the small and limited range of the foreground snow, the goal of min max range, which drives towards balancing the range values between S and \bar{S} , is not relevant.
3. min normalized range sum: The objective function here has the ranges of sets S and \bar{S} divided by the respective number of pixels it contains. This drives the more range concentrated set (the snow) to be smaller, and the larger range set (the background) to be significantly larger. Since here we do not want to assume any specific proportion on the partition into S and \bar{S} , this objective is unsuitable. Still, since the specific partition in Figure 4B is fairly balanced in terms of the number of pixels that fall in S and \bar{S} it is likely that for this image min normalized range sum will produce the same or similar output.
4. min range cut: Here, the addition of the cut value to the sum of ranges mitigates the effect of the ranges' sum in favor of driving towards a small cut value. Since most of the separation (the edges) between the snow and the background is quite distinct in this particular image, the cost of the edges in the cut is fairly small. Even so, the minimum cut part of the objective favors more contiguous areas since their boundary is shorter than the sum of boundaries of multiple separate areas. It is possible for an optimal solution to have \bar{S} assuming the full range in which case the goal of having S with very small range and very small cut, can be achieved when S is a singleton, having very few edges of the cut adjacent to it (at most 4 in a 4-neighbor set up). This would be an undesirable segmentation.
5. min normalized range cut: Here the output will be driven towards the balancing of the sizes of S and \bar{S} , as well as achieving a single contiguous area. Neither one of these characteristics is desired in the ideal segmentation. Moreover, this model is NP-hard and therefore not a practical candidate.

As a result of this evaluation we conclude that the min range sum is the most suitable range objective to be used.

9 | CONCLUSIONS

We introduce here a novel criterion in clustering that seeks clusters with limited *range* of values that characterize each cluster's elements. We present a family of *range-based* clustering objective functions based on commonly considered goals in clustering and demonstrate that, in general, the range-based optimization problems are easier to solve (complexity-wise) than the corresponding total similarity problems. The proposed objectives could therefore be a viable alternative to existing clustering criteria, that are NP-hard, offering the advantage of efficient algorithms. Moreover, the range-based problems are meaningful in clustering applications, such as image segmentation, where the diameter, or range of values associated with objects in each cluster, should be small.

ACKNOWLEDGMENTS

The author is grateful to John E. Baumler for inspiring discussions that initiated the idea of range-clustering, and to Marjan Baghaie for her contributions to the writing of this paper and pointing out reference 16. This is to thank the referee for detailed and perceptive comments as well as providing reference 17. The referee's numerous suggestions contributed significantly to improving the presentation and are much appreciated. This author's research was supported in part by NSF award no. CMMI-1760102 (D.S.H.).

ORCID

Dorit S. Hochbaum  <http://orcid.org/0000-0002-2498-0512>

REFERENCES

- [1] L. Blum, M. Shub and S. Smale, *On a theory of computation over the real numbers; NP-completeness, recursive functions and universal machines*, Bull. Amer. Math. Soc. (N.S.) **21**(1) (1989), 1–46.
- [2] M. Blum, R. Floyd, V. Pratt, R. Rivest and R. Tarjan, *Time bounds for selection*, J. Comput. Systems Sci. **7** (1973), 448–461.
- [3] T. Feo, O. Goldschmidt and M. Khellaf, *One-half approximation algorithms for the k-partition problem*, Oper. Res. **40**(1) (1992), 170–173.
- [4] G. Gallo, M.D. Grigoriadis and R.E. Tarjan, *A fast parametric maximum flow algorithm and applications*, SIAM J. Comput. **18**(30) (1989), 30–55.
- [5] M.R. Garey, D.S. Johnson and L. Stockmeyer, *Some simplified NP-complete graph problems*, Theor. Comput. Sci. **1** (1976), 237–267.
- [6] A.V. Goldberg and R.E. Tarjan, *A new approach to the maximum flow problem*, J. ACM **35** (1988), 921–940.
- [7] O. Goldschmidt and D.S. Hochbaum, *A polynomial algorithm for the k-cut problem for fixed k*, Math. Oper. Res. **19**(1) (1994), 24–37.
- [8] D.S. Hochbaum, *Lower and upper bounds for the allocation problem and other nonlinear optimization problems*, Math. Oper. Res. **19**(2) (1994), 390–409.
- [9] D.S. Hochbaum, *An efficient algorithm for image segmentation, Markov random fields and related problems*, J. ACM **48**(4) (2001), 686–701.
- [10] D.S. Hochbaum, *Complexity and algorithms for nonlinear optimization problems*, Ann. Oper. Res. **153** (2007), 257–296.

- [11] D.S. Hochbaum, *The pseudoflow algorithm: a new algorithm for the maximum-flow problem*, Oper. Res. **56**(4) (2008), 992–1009.
- [12] D.S. Hochbaum, *Polynomial time algorithms for ratio regions and a variant of normalized cut*, IEEE Trans. Pattern Anal. Mach. Intell. **32**(5) (2010), 889–898.
- [13] D.S. Hochbaum, *A polynomial time algorithm for Rayleigh ratio on discrete variables: replacing spectral techniques for expander ratio, normalized cut and Cheeger constant*, Oper. Res. **61**(1) (2013), 184–198.
- [14] D.S. Hochbaum, B. Fishbain, J.E. Baumler and G.E. Gold, *Automated and Semi-Automated Knee Cartilage Segmentation Using Graph-Cuts*. Manuscript, UC Berkeley, Berkeley, CA, 2009.
- [15] D.S. Hochbaum and J.B. Orlin, *Simplifications and speedups of the Pseudoflow algorithm*, Networks **61**(1) (2013), 40–57.
- [16] N. Megiddo, A. Tamir, E. Zemel and R. Chandrasekaran, *An $O(n \log^2 n)$ algorithm for the k th longest path in a tree with applications to location problems*, SIAM J. Comput. **10**(2) (1981), 328–337.
- [17] T.N. Phan, M. Jäger, S. Nadschläger and J. Küng, *Range-based clustering supporting similarity search in big data*. 26th International Workshop on Database and Expert Systems Applications (DEXA), 2015, pp. 120–124.
- [18] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell. **22**(8) (2000), 888–905.
- [19] Z. Wu and R. Leahy, *An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation*, IEEE Trans. Pattern Anal. Machine Intell. **15**(11) (1993), 1101–1113.

How to cite this article: Hochbaum D. Algorithms and complexity of range clustering. *Networks*. 2018;1–17.
<https://doi.org/10.1002/net.21854>