Algorithmica
DOI 10.1007/s00453-017-0374-5

Clustering with Lower-Bounded Sizes

A General Graph-Theoretic Framework

Faisal N. Abu-Khzam[1] · Cristina Bazgan[2,3] · Katrin Casel[4] · Henning Fernau[4]

Received: 7 April 2017 / Accepted: 8 September 2017

Abstract Classical clustering problems search for a partition of objects into a fixed number of clusters. In many scenarios, however, the number of clusters is not known or necessarily fixed. Further, clusters are sometimes only considered to be of significance if they have a certain size. We discuss clustering into sets of minimum cardinality k without a fixed number of sets and present a general model for these types of problems. This general framework allows the comparison of different measures to assess the quality of a clustering. We specifically consider nine quality-measures and classify the complexity of the resulting problems with respect to k. Further, we derive some polynomial-time solvable cases for $k = 2$ with connections to matching-type problems which, among other graph problems, then are used to compute approximations for larger values of k.

✉ Henning Fernau
fernau@uni-trier.de

Faisal N. Abu-Khzam
faisal.abukhzam@lau.edu.lb

Cristina Bazgan
bazgan@lamsade.dauphine.fr

Katrin Casel
casel@uni-trier.de

1 Lebanese American University, Chouran, Beirut 1102 2801, Lebanon

2 CNRS, UMR 7243, LAMSADE, Université Paris-Dauphine, PSL Research University, 75016 Paris, France

3 Institut Universitaire de France, Paris, France

4 Universität Trier, Fachber. 4 - Abteilung Informatikwissenschaften, 54286 Trier, Germany

Published online: 19 September 2017

Keywords Clustering · Computational complexity · Approximation algorithms · Anonymisation

1 Introduction

Clustering problems arise in different areas in very diverse forms with the only common objective of finding a partition of a given set of objects into, by some measure, similar parts. Most models consider variants of the classical k-MEANS or k-MEDIAN problem in the sense that k is a fixed given integer which determines the number of clusters one searches for. In some applications however it is not necessary to compute a partition with exactly k parts, sometimes it is not even clear how to reasonably choose a number for k. We want to discuss a clustering model which does not fix the number of clusters but instead requires that each cluster contains at least k objects. This constraint can be seen as searching for a clustering into parts of a specified minimum significance. For general classification or compression tasks, one might consider small clusters as disposable outliers.

One concrete scenario for this type of partitioning is LOAD BALANCED FACILITY LOCATION [13], a variant of the facility location problem where one is only interested in building facilities which are profitable. In this scenario, a facility is not measured by the initial cost of building it but by its profitability once it is opened. Consequently, it is only reasonable to build a facility if there are enough (but maybe not too many) customers who use it but aside from this constraint we can build as many facilities as we want. The considered cardinality-constraint also models the basic principle of "hiding in a crowd" introduced by the concept of *k-anonymity* [17] which introduces formal problems such as r-GATHER [2], k-MEMBER CLUSTERING [6] and MICROAGGREGATION [8]. A cluster in this context is a collection of personal records which has to have a certain minimum cardinality in order to be considered anonymous.

We want to consider the general task of computing a clustering into sets of minimum cardinality $k \in \mathbb{N}$ with the objective to introduce an abstract framework to model such types of problems. For this purpose, we define the generic problem $(\|\cdot\|, f)$-k-CLUSTER and specifically discuss nine variants of it, characterised via three different choices for the local measure f and the global measure $\|\cdot\|$; a detailed description of these variants follows in Sect. 2. Our main contributions are the abstract model and the complexity- and approximation-results which become more apparent due to this model, as they are derived mostly via reductions to/from other graph problems. Section 3 compares the nine problem variants with respect to structural differences. In Sects. 4 and 5, we classify the complexity for small values of k by identifying polynomial-time solvable cases with connections to matching-type problems and deriving (also improving known) NP-hardness results for the remaining cases. Section 6 uses a large variety of connections to other graph problems, including the results from Sect. 4, to develop approximation-algorithms. A more detailed description of the results as well as the comparison to results from related work follows in the respective sections and is summarised in Tables 1 and 2 in the conclusions.

An extended abstract of this paper was published with ISAAC 2016; see [1].

2 General Abstract Model

In the following, we consider the general task of partitioning a set of n given objects into sets of cardinality at least k. Our model represents the objects as vertices of an undirected graph $G = (V, E)$. A feasible solution is any partition $P_1, \ldots, P_s$ of V such that $|P_i| \geq k$ for all $i \in \{1, \ldots, s\}$. In the following we will refer to such a partition as *k-cluster*. Recall that in contrast to the classical clustering problems like s-MEANS or s-MEDIAN, the number of clusters s is not necessarily part of the input. Of course, one does not search for just any k-cluster but for a partition which preferably only combines objects which are in some sense "close". This similarity can be very hard to capture and the appropriate way to measure it highly depends on the clustering-task and the structure of the input. We therefore consider an arbitrary given distance function $d\colon V \times V \to \mathbb{Q}_+$ which for any two objects $u, v \in V$ represents the distortion which is caused by combining u and v. This general view allows to simultaneously study many different measures for dissimilarity.

In our model, the distance d is defined via a given edge-weight function $w_E\colon E \to \mathbb{Q}_+$. For two vertices $u, v \in V$ we define $d(u, v) := w_E(\{u, v\})$ if $\{u, v\} \in E$, and if $\{u, v\} \notin E$, the distance $d(u, v)$ is defined by the shortest path from u to v in G. We will say that d satisfies the *triangle inequality* (and hence is a metric) if $d(u, v) \leq d(u, w) + d(w, v)$ for all $u, v, w \in V$. Observe that our definition allows for distances d which do not satisfy this property, a simple example is the complete graph over $V = \{u, v, w\}$ with $w_E(\{u, v\}) = w_E(\{u, w\}) = 1$ and $w_E(\{v, w\}) = 3$. Violations of the triangle inequality are only possible for distances defined by an edge. Edges hence do not necessarily imply similarity but can reflect a difference greater than the shortest path between two objects and make it more unattractive to cluster them together; very different from the multiedges introduced in the hypergraph-model for k-anonymous clustering from [20], where hyperedges reflect similar groups.

The overall cost of a partition $P_1, \ldots, P_s$ is always in some sense proportional to the dissimilarities within each set or *cluster* P. On an abstract level, the *global cost* induced by a partition $P_1, \ldots, P_s$ is calculated by first computing the *local cost* of each cluster and second by combining all this individual information. In this paper, we discuss three different measures for the local cost caused by a cluster P:

Radius	$\mathrm{rad}(P) := \min\{\max\{d(x, y)\colon y \in P\}\colon x \in P\}$.
Diameter	$\mathrm{diam}(P) := \max\{\max\{d(x, y)\colon y \in P\}\colon x \in P\}$.
Average Distortion	$\mathrm{avg}(P) := \lvert P\rvert^{-1} \cdot \min\{\sum_{y \in P} d(x, y)\colon x \in P\}$.

Here and throughout the paper d denotes the distance induced *on the whole graph*; hence we consider for $u, v \in P$ with $\{u, v\} \notin E$ as distance $d(u, v)$ the shortest path from u to v in G even if this path contains vertices which are not in P. For the local measures average distortion or radius we will sometimes call a vertex $x \in P$ *a central vertex* for cluster P, if $\mathrm{avg}(P) = \frac{1}{|P|} \sum_{y \in P} d(x, y)$ or $\mathrm{rad}(P) = \max\{d(x, y)\colon y \in P\}$, respectively. Observe that central vertices with respect to average distortion and radius may be different; in the cluster $P = \{x, y, x_1, y_1, y_2\}$ with $w_E(\{y, x\}) = w_E(\{y, y_1\}) = w_E(\{y, y_2\}) = 1$ and $w_E(\{x, x_1\}) = 2$, the vertex x is the only central vertex with respect to radius and y is the only central vertex with respect to average distortion.

The overall cost of a k-cluster $P_1, \ldots, P_s$ is given by a combination of the local costs $f(P_1), \ldots, f(P_s)$ with $f \in \{\text{rad}, \text{diam}, \text{avg}\}$. In order to model the most common problem-versions we consider the following three possibilities:

Worst Local Cost: Maximum cost among all clusters, formally computed by $\max\{f(P_i) : 1 \leq i \leq s\}$, denoted by $\|\cdot\|_\infty$.
Worst Weighted Local Cost: Maximum cost among all clusters, weighted by their sizes computed by $\max\{|P_i| f(P_i) : 1 \leq i \leq s\}$, denoted by $\|\cdot\|_\infty^w$.
Accumulated Weighted Local Cost: The sum of the local costs of all clusters, weighted by their sizes, computed by $\sum_{i=1}^{s} |P_i| f(P_i)$, denoted by $\|\cdot\|_1^w$.

Any combination of $f \in \{\text{rad}, \text{diam}, \text{avg}\}$ with $\|\cdot\| \in \{\|\cdot\|_1^w, \|\cdot\|_\infty^w, \|\cdot\|_\infty\}$ yields a different problem. Structural properties discussed in Sect. 3 will explain why we do not consider unweighted 1-norm. For a fixed $k \in \mathbb{N}$, the general optimisation problem is given by:

$(\|\cdot\|, f)$-k-CLUSTER

Input: Graph $G = (V, E)$ with edge-weight function $w_E : E \to \mathbb{Q}_+$, $k \in \mathbb{N}$.
Output: A k-cluster $P_1, \ldots, P_s$ of V for some $s \in \mathbb{N}$, which minimises $\|(f(P_1), \ldots, f(P_s))\|$.

We will use the name $(\|\cdot\|, f)$-k-CLUSTER to also refer to the natural corresponding decision problem, i.e., given a graph G with edge-weights, an integer k and a bound $D \in \mathbb{Q}_+$, does there exist a k-cluster $P_1, \ldots, P_s$ of V for some $s \in \mathbb{N}$ such that $\|(f(P_1), \ldots, f(P_s))\| \leq D$.

Some of the variants of $(\|\cdot\|, f)$-k-CLUSTER are known under different names. $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER is equivalent to k-MEMBER CLUSTERING [6] and with d chosen as the Euclidean distance, $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER is the problem r-GATHER [2] (with $r = k$). The variant $(\|\cdot\|_1^w, \text{avg})$-$k$-CLUSTER models LOAD BALANCED FACILITY LOCATION [13] with unit demands and without facility costs. Further, again with d being the Euclidean distance, $(\|\cdot\|_1^w, \text{avg})$-$k$-CLUSTER is equivalent to MICROAGGREGATION [8].

Choosing between the cluster-measures and norms allows adjustment for specific types of objects and different forms of output representation. The norm decides if the desired output has preferably uniformly structured clusters with or without uniform cardinalities (∞-norms) or builds clusters of object-specific irregular structure (1-norm). For cohesive clustering, the diameter-measure is more suitable for the choice of f. Average distortion is best used when the output chooses one representative of each cluster and projects all other objects in this cluster to it; a scenario which for example occurs for facility-location type problems. If the output does not project to one representative but considers clusters as circular areas, the radius measure is the most reasonable choice for f. Optimal k-clusters may differ for different choices of $\|\cdot\|$ and/or f as we will discuss in the next section. Still, we will see that there are also very useful similarities.

3 Structural Properties of Optimal Partitions

The diverse behaviour for different choices of f and $\|\cdot\|$ is nicely displayed in the cluster-cardinalities of optimal solutions. For the example $G = (V, E)$ with $V = \{c, v_1, v_2, \ldots, v_n\}$ and $E = \{\{v_i, c\}\colon 1 \leq i \leq n\}$ with $w_E(\{c, v_i\}) = 1$ for all i, we find that for radius and average distortion, the single cluster V is *the* optimal solution with $\|\cdot\|_\infty$ or $\|\cdot\|_1^w$. If $w_E(\{v_i, v_j\}) = D$ for some large value D, any k-cluster with more than one set is arbitrarily worse. For the diameter-measure however we know that in general $\operatorname{diam}(S) \leq \operatorname{diam}(P)$ for all sets $S \subseteq P$, which immediately yields:

Proposition 1 *From any given solution $\mathfrak{P}$ for an instance of $(\|\cdot\|, \operatorname{diam})$-$k$-CLUSTER it is possible to compute in polynomial time a solution $\mathfrak{P}'$ of the same global cost for which $|P| < 2k$ for all $P \in \mathfrak{P}'$, for all choices of $\|\cdot\| \in \{\|\cdot\|_1^w, \|\cdot\|_\infty^w, \|\cdot\|_\infty\}$ and $k \in \mathbb{N}$.*

For radius we only have the weaker property that $\operatorname{rad}(S) \leq \operatorname{rad}(P)$ for all sets $S \subseteq P$ such that a central vertex for P with respect to radius is contained in S. Average distortion lacks such monotone behaviour entirely. Observe that a large cardinality of a cluster can somehow "smooth over" some larger distances, for example for three vertices u, v, w with $w_E(\{u, v\}) = 3$ and $w_E(\{u, w\}) = 1$, adding w to the cluster $\{u, v\}$ decreases the average distortion from $\frac{3}{2}$ to $\frac{4}{3}$. Examples like these show that, even with triangle inequality for d, we cannot in general restrict the maximum cluster-cardinality for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER , which is a bit undesirable, given that most applications also prefer to have some natural upper bound on the cardinality (not too many customers). In a realistic scenario, we encounter sets of cardinality $2k$ or larger in optimal solutions for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER , if they contain an object (often called outlier) which has a large distance from all objects. Deleting such outliers before computing clusters is generally a reasonable pre-processing step, which makes large clusters in $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER unlikely.

In general, we would like the computation of global cost to somehow favour finer partitions in order to exploit the difference to clustering models which bound the number of sets. This is the reason why we do not consider the unweighted 1-norm, i.e., $\|\,(f(P_1), \ldots, f(P_s))\,\|_1 := \sum_{i=1}^{s} f(P_i)$. For the example $V = \{v_i^1, v_i^2\colon 1 \leq i \leq n\}$ with $w_E(\{v_i^1, v_i^2\}) = 1$ for $i \in \{1, \ldots, n\}$ and $w_E(\{v_i^h, v_j^k\}) = n - 1$ for $i, j \in \{1, \ldots, n\}$ with $i \neq j$ and $h, k \in \{1, 2\}$, the best 2-cluster w.r.t. $\|\cdot\|_1$ with any choice for f is V itself, while the most reasonable 2-cluster for most applications one can think of for this graph is obviously $\{\{v_i^1, v_i^2\}\colon 1 \leq i \leq n\}$. This makes $\|\cdot\|_1$ very unattractive for our clustering purposes. Observe that triangle inequality does not improve this behaviour, since the distance d for this example satisfies it.

Triangle inequality however has the strong advantage that we can restrict (for most variants of $(\|\cdot\|, f)$-k-CLUSTER without loss of generality) the set of solutions to only contain clusters of a maximum cardinality of $2k - 1$.

Theorem 1 *For any $k \in \mathbb{N}$ and any graph G with edge-weights for which the induced distance d satisfies the triangle inequality, it is possible to compute in polynomial time from any given k-cluster $\mathfrak{P}$ for G, a k-cluster $\mathfrak{P}'$ for which $|P| < 2k$ for all $P \in \mathfrak{P}'$ and such that:*

– $\mathfrak{P}'$ *has the same global cost as* $\mathfrak{P}$ *with respect to* $\|\cdot\|_\infty^w$ *and* rad *or* avg.
– $\mathfrak{P}'$ *has at most twice the global cost of* $\mathfrak{P}$ *with respect to* $\|\cdot\|_1^w$ *and* rad *or* avg, *and also with respect to* $\|\cdot\|_\infty$ *and* rad.

Proof Consider a k-cluster $\mathfrak{P}$ containing a cluster P of cardinality $s = tk+r$ for some $t \geq 2$ and $k > r \geq 0$ with some central vertex $c \in P$ with respect to the considered local measure $f \in \{\text{rad}, \text{avg}\}$. Construct successively for $i \in \{1, \ldots, t-1\}$ the sets V_i containing k vertices from $P_i \backslash \{c\}$, where $P_i := P \backslash (V_1 \cup \cdots \cup V_{i-1})$, including $v_i := \text{argmin}\{d(p, c) \colon p \in P_i \backslash \{c\}\}$. We consider the increase of global cost for replacing P by $V_1, \ldots, V_{t-1}, P_{t-1}$ in $\mathfrak{P}$:

For the local measure radius, we see that $\text{rad}(P_i) \leq \text{rad}(P)$ for all i and hence especially for $i = t-1$. The radius of the sets V_i can be bounded by:

$$\text{rad}(V_i) \leq \max\{d(v_i, p) \colon p \in V_i\} \leq d(v_i, c) + \max\{d(c, p) \colon p \in V_i\} \leq 2 \cdot \text{rad}(P).$$

The global cost for $(\|\cdot\|_1^w, \text{rad})$-$k$-CLUSTER and $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER only increases by a factor of at most two. For the weighted ∞-norm, these inequalities yield:

$$|V_i| \cdot \text{rad}(V_i) = k \cdot \text{rad}(V_i) \leq 2k \cdot \text{rad}(P) \leq |P| \cdot \text{rad}(P).$$

The global cost for $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER consequently does not increase.

For the local measure average distortion, the weighted average for each P_i with $i \in \{1, \ldots, t-1\}$ is bounded by:

$$|P_i| \cdot \text{avg}(P_i) \leq \sum_{p \in P_i} d(c, p) \leq |P| \cdot \text{avg}(P).$$

The local cost for V_i with $i \in \{1, \ldots, t-1\}$ is bounded by:

$$|V_i| \cdot \text{avg}(V_i) \leq \sum_{p \in V_i} d(v_i, p) \leq k \cdot d(v_i, c) + \sum_{p \in V_i} d(c, p).$$

By the choice of the vertices v_i we can bound $k \cdot d(v_i, c) \leq \sum_{p \in P_i} d(c, p)$ and conclude that:

$$|V_i| \cdot \text{avg}(V_i) \leq \sum_{p \in P_i} d(c, p) + \sum_{p \in V_i} d(c, p) = \sum_{p \in P_{i-1}} d(c, p) \leq |P| \cdot \text{avg}(P).$$

The global cost with respect to the weighted ∞-norm $\|\cdot\|_\infty^w$ consequently does not increase by replacing P by $V_1, \ldots, V_{t-1}, P_{t-1}$. For $(\|\cdot\|_1^w, \text{avg})$-$k$-CLUSTER the partition $V_1, \ldots, V_{t-1}, P_{t-1}$ adds each distance $d(c, p)$ with $p \in P$ at most twice compared to partitioning into P, which also means that the global cost is at most doubled. □

We will look at the particular case of $k = 2$ in the next section and therefore also show:

Proposition 2 *For any instance of* $(\|\cdot\|_1^w$, avg)-2-CLUSTER *for which the induced distance d satisfies the triangle inequality, it is possible to compute in polynomial time from any optimal solution* $\mathfrak{P}$*, an optimal solution* $\mathfrak{P}'$ *for which* $|P| \in \{2, 3\}$ *for all* $P \in \mathfrak{P}'$.

Proof For a cluster $P = \{x_1, x_2 \ldots, x_r\}$ with $r > 3$, let x_r be a central vertex with respect to average distortion. A further partitioning of P into $\{x_{2i}, x_{2i+1}\}$ for $i \in \{1, \ldots, z-1\}$ with $z = \lceil \frac{r}{2} \rceil$ and $\{x_1, x_{2z}, x_r\}$ does not increase the global cost for $(\|\cdot\|_1^w$, avg)-2-CLUSTER , since:

$$
\begin{aligned}
|P| \cdot \operatorname{avg}(P) &= \sum_{i=1}^{r} d(x_i, x_r) \\
&= d(x_{2z}, x_r) + d(x_r, x_1) + \sum_{i=1}^{z-1} d(x_{2i}, x_r) + d(x_{2i+1}, x_r) \\
&\geq |\{x_1, x_{2z}, x_r\}| \cdot \operatorname{avg}(\{x_1, x_{2z}, x_r\}) + \sum_{i=1}^{z-1} d(x_{2i}, x_{2i+1}) \\
&= |\{x_1, x_{2z}, x_r\}| \cdot \operatorname{avg}(\{x_1, x_{2z}, x_r\}) + \sum_{i=1}^{z-1} 2 \cdot \operatorname{avg}(\{x_{2i}, x_{2i+1}\}) \\
&= \| \operatorname{avg}(\{x_1, x_{2z}, x_r\}), \operatorname{avg}(\{x_2, x_3\}), \ldots, \operatorname{avg}(\{x_{2z-2}, x_{2z-1}\}) \|_1^w
\end{aligned}
$$

□

4 Connections to Matching Problems

The graph-representation we chose to define $(\|\cdot\|, f)$-k-CLUSTER reveals relations to other well studied graph problems, in case of $k = 2$ not to classical clustering but to matching problems. Some variants can be reduced to finding a minimum-weight edge cover in a graph, that is, a problem which can be reduced to the problem of finding a minimum weight perfect matching (a simple reduction is described, e.g., in the first volume of Schrijver's monograph [[18], Section 19.3]). As a consequence, a minimum-weight edge cover can be found in $O(n^3)$ time by the results of Edmonds and Johnson [10].

Theorem 2 $(\|\cdot\|_1^w$, avg)-2-CLUSTER *can be solved in* $\mathcal{O}(n^3)$ *time.*

Proof $(\|\cdot\|_1^w$, avg)-2-CLUSTER searches for a 2-cluster $P_1, \ldots, P_s$ minimising:

$$
\sum_{i=1}^{s} \min \left\{ \sum_{y \in P_i} d(x, y) \colon x \in P_i \right\}.
$$

In other words, for any graph $G = (V, E)$, the global cost is the weight of the cheapest edge-set $E' \subseteq V \times V$ for which the graph $G' := (V, E')$ has s connected components

$P_1, \ldots, P_s$ with at least 2 vertices such that the induced subgraph of each P_i is a star-graph. This property is equivalent to E' being a minimum-weight edge cover for the complete graph on V with edge-weights equal to the distance d; observe that the graph (V, E') is a forest without isolates and without paths of length 3 for every minimum-weight edge cover E' which means that its connected components are star-graphs. □

Proposition 3 $(\|\cdot\|_\infty, \text{rad})$-2-CLUSTER *can be solved in* $\mathcal{O}(n^3)$ *time.*

Proof For a graph $G = (V, E)$, first check all vertices in V and find the smallest value $c > 0$ such that each vertex v has distance at most c from at least one other vertex. This c is obviously a general lower bound on the global cost, since each vertex needs at least one 'partner'.

For $k = 2$, this c is also the optimal value. To see this, let $\bar{E}$ be any minimum edge cover for the graph $G' := (V, E')$ with $E' := \{\{u, v\} : 0 < d(u, v) \leq c\}$. Such a cover exists, as there are no isolated vertices in G' by the choice of c. Let $C_1, \ldots, C_s$ be the connected components of the graph induced by the edges in $\bar{E}$. Each such component C_i is a star graph by the minimality of the edge cover and contains at least two vertices, hence the partition $\{V[C_i] : 1 \leq i \leq s\}$ is a 2-cluster for G with radius at most c for each cluster. An optimal solution for $(\|\cdot\|_\infty, \text{rad})$-2-CLUSTER can hence be obtained by computing a minimum edge cover for G'. □

With respect to diameter, this edge cover strategy is not applicable for clusters of cardinality larger than two. Even for $k = 2$ there are cases for which clusters of cardinality 3 are required in every optimal solution. It seems difficult to compute the diameter of a cluster by summing up certain edge-weights. We therefore consider the following matching problem which is more involved but still solvable in $\mathcal{O}(n^3 m^2 \log n)$ [3] (this kind of generalised matching can also be used for anonymisation by deletion, see [5]):

SIMPLEX MATCHING

Input: Hypergraph $H = (V, F)$ with $F \subseteq (V^2 \cup V^3)$ and cost-function $c : F \to \mathbb{Q}$ satisfying:

(a) $\{\{u, v\}, \{v, w\}, \{u, w\}\} \subseteq F$ for all $\{u, v, w\} \in F$. (*subset condition*)

(b) $c(\{u, v\}) + c(\{v, w\}) + c(\{u, w\}) \leq 2c(\{u, v, w\})$ for all $\{u, v, w\} \in F$. (*simplex condition*)

Output: A perfect matching of H (that is a set $S \subseteq F$ such that every vertex in V appears in exactly one hyperedge of S) of minimal cost.

Proposition 4 $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER *can be solved in* $\mathcal{O}(n^9 \log n)$ *time.*

Proof We model our problem as a particular instance of SIMPLEX MATCHING. Let $G = (V, E)$ be an input graph for $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER . The corresponding input for SIMPLEX MATCHING is the hypergraph $H = (V, V^2 \cup V^3)$ which obviously satisfies the subset condition. By Proposition 1, there exists an optimal solution for $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER among the perfect matchings for H. According to the original problem, the cost-function c for any $u, v, w \in V$ is defined as:

– $c(\{u, v\}) := 2d(u, v)$ and
– $c(\{u, v, w\}) := 3 \cdot \max\{d(u, v), d(v, w), d(u, w)\}$

and hence satisfies the simplex condition. Since this complete hypergraph has $\mathcal{O}(n^3)$ hyperedges, the overall running-time is in $\mathcal{O}(n^9 \log n)$. □

Diameter combined with the ∞-norms could be solved using Proposition 4 by fixing some maximum diameter D and multiplying all hyperedge-costs which exceed D with a large value C, say $C = n \cdot \max\{d(u, v) : u, v \in V\}$. This does not violate the simplex condition for the cost-function and there exists a solution for $(\|\cdot\|_\infty, \text{diam})$-2-CLUSTER of value D for the input graph if and only if the hypergraph with adjusted costs has a SIMPLEX MATCHING solution of value less than C.

To improve upon the running-time from Proposition 4 for the ∞-norms, we will use following problem from [22].[1]

SIMPLEX COVER	
Input:	Hypergraph $H = (V, F)$ with $F \subseteq (V^2 \cup V^3)$ satisfying the subset condition, i.e., $\{\{u, v\}, \{v, w\}, \{u, w\}\} \subseteq F$ for all $\{u, v, w\} \in F$.
Output:	A perfect matching of H.

Proposition 5 $(\|\cdot\|_\infty, \text{diam})$- *and* $(\|\cdot\|_\infty^w, \text{diam})$-2-CLUSTER *and can be solved in* $\mathcal{O}(n^6 \log n)$ *time. On instances for which d satisfies the triangle inequality,* $(\|\cdot\|_\infty^w, \text{avg})$-2-CLUSTER *can also be solved in* $\mathcal{O}(n^6 \log n)$ *time.*

Proof We will reduce solving each of the 2-CLUSTER problem variants to solving an instance of SIMPLEX COVER. Let $G = (V, E)$ be the input graph for the clustering problem. By Proposition 1 and Theorem 1 we can find optimal solutions for each considered problem variant among the set of perfect matchings for the hypergraph $H = (V, F)$ with $F = V^2 \cup V^3$. For a fixed value D, we build a subset $F' \subseteq F$ by removing from F all $e \in F$ depending on the problem variant by the following rule:

– Remove e if $\text{diam}(e) > D$ for $(\|\cdot\|_\infty, \text{diam})$-2-CLUSTER .
– Remove e if $|e| \cdot \text{diam}(e) > D$ for $(\|\cdot\|_\infty^w, \text{diam})$-2-CLUSTER .
– Remove e if $|e| \cdot \text{avg}(e) > D$ for $(\|\cdot\|_\infty^w, \text{avg})$-2-CLUSTER .

We claim that in all three cases, this deletion yields a subset of $V^2 \cup V^3$ that satisfies the subset condition:

– $\{u, v, w\} \in F'$ for $(\|\cdot\|_\infty, \text{diam})$-2-CLUSTER implies $\text{diam}(\{u, v, w\}) \leq D$ and hence $\text{diam}(\{u, v\}), \text{diam}(\{u, w\}), \text{diam}(\{v, w\}) \leq \text{diam}(\{u, v, w\}) \leq D$, so $\{\{u, v\}, \{v, w\}, \{u, w\}\} \subseteq F'$.
– If $\{u, v, w\} \in F'$ for $(\|\cdot\|_\infty^w, \text{diam})$-2-CLUSTER , we have $3 \cdot \text{diam}(\{u, v, w\}) \leq D$, hence $2 \cdot \text{diam}(\{u, v\}) \leq D$, $2 \cdot \text{diam}(\{u, w\}) \leq D$ and $2 \cdot \text{diam}(\{v, w\}) \leq D$, so $\{\{u, v\}, \{v, w\}, \{u, w\}\} \subseteq F'$.

[1] This covering problem is sometimes also called UNWEIGHTED SIMPLEX MATCHING and is equivalent to $\{K_2, K_3\}$-PACKING, an old, well studied generalisation of the classical matching problem [7].

– If $\{u, v, w\} \in F'$ for $(\|\cdot\|_\infty^w, \text{avg})$-2-CLUSTER, we have $3{\cdot}\text{avg}(\{u, v, w\}) \leq D$. Let u be central for $\{u, v, w\}$, so $d(u, v) + d(u, w) = 3 \cdot \text{avg}(\{u, v, w\})$. It follows that $2{\cdot}\text{avg}(\{u, v\}) = d(u, v) \leq D, 2{\cdot}\text{avg}(\{u, w\}) = d(u, w) \leq D$. For the edge $\{v, w\}$ we require that d satisfies the triangle inequality, in which case $2 \cdot \text{avg}(\{v, w\}) = d(v, w) \leq d(u, v) + d(u, w) \leq D$, so $\{\{u, v\}, \{v, w\}, \{u, w\}\} \subseteq F'$.

In all three cases, any subset of F' which exactly covers V, i.e., a simplex cover for $H' := (V, F')$, yields a feasible 2-cluster with global cost at most D. The augmenting-path strategy from [19] solves SIMPLEX COVER in time $\mathcal{O}(m^2)$, where m is the number of hyperedges of the input graph, here at most $\mathcal{O}(n^3)$. Possible values for D are the $\mathcal{O}(n^2)$ possible different distances $d(u, v)$ for all $u, v \in V$, which, including a binary search among all possible values for D, yields an overall running-time in $\mathcal{O}(n^6 \log n)$ to solve each of the 2-CLUSTER variants. □

Remark 1 We would like to point out that SIMPLEX MATCHING is also an interesting way to solve a sort of geometric version of $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER, originally introduced as MICROAGGREGATION in [8], which considers clustering a set of vectors in $\mathbb{R}^d$ and measures local cost for a cluster $\{x_1, \ldots, x_t\}$ by $\sum_{i=1}^t ||x_i - x||_2^2$ where x is the centroid $\frac{1}{t}(x_1 + \cdots + x_t)$ and $\|\cdot\|_2^2$ is the squared Euclidean norm. With the hypergraph $(V, V^2 \cup V^3)$ with $V = \{v_1, \ldots, v_n\}$ representing $\{x_1, \ldots, x_n\}$ and the cost-function c defined by: $c(\{v_i, v_j, v_k\}) := \sum_{h \in \{i,j,k\}} ||x_h - \frac{1}{3}(x_i + x_j + x_k)||_2^2$ for all $1 \leq i < j < k \leq n$ and $c(\{v_i, v_j\}) := \frac{1}{2}||x_i - x_j||_2^2$ for all $1 \leq i < j \leq n$, the simplex condition holds, since:

$$2 \cdot c(\{v_i, v_j, v_k\}) = \tfrac{4}{3}(c(\{v_i, v_j\}) + c(\{v_j, v_k\}) + c(\{v_i, v_k\})).$$

This construction gives a polynomial-time algorithm to solve 2- MICROAGGREGATION which improves on the 2-approximation from [9].

Observe that a similar construction for $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER does not work, since the cluster-cardinality is not bounded by three. Also, even if d satisfies the triangle inequality, the corresponding cost-function c would not satisfy the simplex condition, since for the small example of three vertices u, v, w with $d(u, v) = d(u, w) = 1$ and $d(v, w) = 2$, the cost with respect to radius would give $1 = c(\{u, v, w\}) < \frac{1}{2}(c(\{u, v\}) + c(\{u, w\}) + c(\{v, w\})) = 2$. Similar problems arise for the other so far unresolved variants of $(\|\cdot\|, f)$-2-CLUSTER.

At last, we would like to point out that the running-times presented in this section all assume the worst-case in which there are $\mathcal{O}(n^2)$ pairs of vertices with small distance to each other; a property that might be avoided for certain specific clustering tasks. We further believe that an augmenting path strategy which is specifically tailored to the above problems can also yield significant improvement on the worst-case running-time.

5 Complexity Results

The problem variant $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER with the specific choice of d being the Euclidean distance was discussed in [2] under the name r-GATHER (where r

takes the role of k) and was there shown to be NP-complete for $k \geq 7$. In [4] this result was strengthened by a reduction from EXACT- t- COVER to $k \geq 3$, however for a type of problem where the cluster-center exists as an input vertex but is assigned to a different cluster (i.e., with the radius of a cluster P calculated by: $\min\{\max\{d(x, y) : y \in P\} : x \in V\}$) which does not comply with our definition. We establish different reductions which show NP-hardness for all variants of $(\|\cdot\|, f)$-k-CLUSTER with $k \geq 3$. We also reduce from the problem EXACT- t- COVER, formally given by:

EXACT- t- COVER	
Input:	A universe $X = \{x_1, \ldots, x_n\}$ and a collection $C = \{S_1, \ldots, S_r\}$ of subsets of X, such that each S_i, $i \in \{1, \ldots, r\}$, has cardinality t.
Question:	Does there exist a subset $C' \subseteq C$ (*exact cover*) that is a partition of X?

EXACT-t- COVER is known to be NP-hard for all $t \geq 3$ [11].

Theorem 3 *The problem* $(\|\cdot\|, \text{rad})$-k- CLUSTER *is NP-hard for each* $k \geq 3$ *and all choices of* $\|\cdot\| \in \{\|\cdot\|_\infty, \|\cdot\|_\infty^w, \|\cdot\|_1^w\}$, *even with the restriction to distances* d *which satisfy the triangle inequality.*

Proof We reduce from EXACT-k-COVER. Let $S_1, \ldots, S_r$ be subsets of the universe $\{x_1, \ldots, x_n\}$, with $|S_i| = k$, an instance of EXACT-k-COVER and let $\ell := r - \frac{n}{k}$ (exactly the number of sets not included in an exact cover). We construct a graph $G = (V, E)$ for $(\|\cdot\|, \text{rad})$-k-CLUSTER with a vertex set V built from the following three types of vertices (for an illustration of this construction see Fig. 1):

- $u_1, \ldots, u_n$ representing $x_1, \ldots, x_n$,
- $w_1, \ldots, w_r$ representing $S_1, \ldots, S_r$ and
- y_i^j for $i \in \{1, \ldots, \ell\}$ and $j \in \{1, \ldots, k-1\}$, vertices which will be clustered with the w-vertices corresponding to sets which are not in the exact cover.

The set E contains the following edges, all of weight 1:

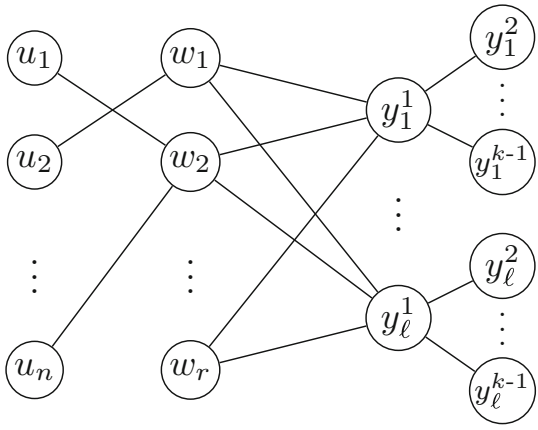


Fig. 1 Illustration of the reduction for Theorem 3

– $\{u_i, w_j\}$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, r\}$ with $x_i \in S_j$,
– $\{y_i^1, w_j\}$ for each $i \in \{1, \ldots, \ell\}$ and $j \in \{1, \ldots, r\}$ and
– $\{y_i^1, y_i^h\}$ for each $i \in \{1, \ldots, \ell\}$ and $h \in \{2, \ldots, k-1\}$.

We claim that there exists a k-cluster for G which only contains clusters of radius 1 if and only if there exists an exact cover for $S_1, \ldots, S_r$.

Let $\mathfrak{P}$ be a k-cluster for G which only contains clusters of radius 1, and let d be the distance on $V \times V$ induced by the edges of G. For each $i \in \{1, \ldots, \ell\}$, let P_i denote the cluster in $\mathfrak{P}$ containing y_i^2, as $k \geq 3$, a vertex y_i^j with index $j = 2$ is always included in G. Since y_i^1 is the only vertex at distance 1 from y_i^2, it follows that y_i^1 is included as the unique central vertex in P_i which means that $P_i \subseteq \{v \in V : d(v, y_i^1) \leq 1\}$. As $\{v \in V : d(v, y_i^1) = 1\} = \{y_i^1, \ldots, y_i^{k-1}\} \cup \{w_1, \ldots, w_r\}$ and $|P_i| \geq k$, it follows that at least ℓ of the vertices $w_1, \ldots, w_r$ are included in the clusters $P_1, \ldots, P_\ell$, none of which contain a vertex from $\{u_1, \ldots, u_n\}$. Since $d(u_i, u_j) \geq 2$ for all $i \neq j$, a cluster in $\mathfrak{P}$ which contains two vertices from $\{u_1, \ldots, u_n\}$ has to contain at least one of the vertices w_z as central vertex. Such a cluster then has to be a subset of $\{w_z\} \cup \{u_i : x_i \in S_z\}$. There are only $\frac{n}{k}$ vertices from $\{w_1, \ldots, w_r\}$ which lie in such a cluster, so $\mathfrak{P}$ has to contain exactly the clusters $\{w_z\} \cup \{u_i : x_i \in S_z\}$ for all $w_z \notin P_1 \cup \cdots \cup P_\ell$ in order to include all vertices u_i in a cluster of radius 1. This means that the sets S_z with $\{w_z\} \cup \{u_i : x_i \in S_z\} \in \mathfrak{P}$ build an exact cover for $\{x_1, \ldots, x_n\}$. It also follows that all clusters in a k-cluster of maximum radius 1 contain at most $k+1$ vertices.

Conversely, for any exact cover $S \subseteq \{S_1, \ldots, S_r\}$ the union of the sets $\{w_z\} \cup \{u_i : x_i \in S_z\}$ for all z with $S_z \in S$ and $\{y_i^1, \ldots, y_i^{k-1}\} \cup \{w_{j_i}\}$ for all $i \in \{1, \ldots, \ell\}$ where $\{S_1, \ldots, S_r\} \backslash S = \{S_{j_1}, \ldots, S_{j_\ell}\}$ yields a k-cluster of radius 1 for G.

If $\mathrm{rad}(P) > 1$ for some cluster P in a k-cluster $\mathfrak{P}$ for G, it follows that $\mathrm{rad}(P) \geq 2$; observe that since G only has edges of weight 1, all shortest paths have integer length. This means that the global cost of $\mathfrak{P}$ with respect to radius and $\|\cdot\|_\infty^w$ is at least $2k$, so strictly larger than the global cost of a k-cluster of maximum radius 1 for this norm, which is $k+1$ by the above stated property of k-cluster of maximum radius 1 for G. Also, the global cost of $\mathfrak{P}$ with respect to $\|\cdot\|_1^w$ is at least $kr + \frac{n}{k} + k$ (at least k vertices produce a cost of 2), while a k-cluster of maximum radius 1 with respect to this norm yields a global cost of $kr + \frac{n}{k}$ (each vertex produces a cost of 1). In summary, there exists an exact cover for $S_1, \ldots, S_r$ if and only if there exists a solution for $(\|\cdot\|, \mathrm{rad})$-$k$-CLUSTER of global cost 1, $k+1$ and $kr + \frac{n}{k}$ for norm $\|\cdot\|_\infty$, $\|\cdot\|_\infty^w$ and $\|\cdot\|_1^w$, respectively. □

In the above proof of Theorem 3 there is a gap of 2 for the maximum radius between "yes"- and "no"-instance for EXACT- k- COVER, which implies:

Corollary 1 *There is no* $(2-\varepsilon)$*-approximation for* $(\|\cdot\|_\infty, \mathrm{rad})$-$k$-CLUSTER *in polynomial time for any* $k \geq 3$ *and any* $\varepsilon > 0$*, unless* $P = NP$*, even if* d *satisfies the triangle inequality.*

If we alter the reduction used for Theorem 3 for $k \geq 4$ to reduce to EXACT-$(k-1)$-COVER, we can conclude that in case of a "yes"-instance for EXACT-$(k-1)$-COVER all clusters in a k-cluster of maximum radius 1 for the corresponding graph G contain

exactly k vertices. This yields a gap of 2 also for the maximum weighted radius between "yes"- and "no"-instance for EXACT-$(k-1)$-COVER, which implies:

Corollary 2 *There is no* $(2-\varepsilon)$*-approximation for* $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER *in polynomial time for any* $k \geq 4$ *and any* $\varepsilon > 0$*, unless* $P = NP$*, even if* d *satisfies the triangle inequality.*

For diameter, we need a different construction, since for this measure, the vertices $u_1, \ldots, u_n$ have to also be at distance 1 to enable some of them to be in the same cluster. With such distances, we need a different structure which makes sure that a solution of diameter 1 does not build clusters only containing vertices from $u_1, \ldots, u_n$.

Theorem 4 *The problem* $(\|\cdot\|, \text{diam})$-k-CLUSTER *is NP-hard for each* $k \geq 3$ *and all choices for* $\|\cdot\| \in \{\|\cdot\|_\infty, \|\cdot\|_\infty^w, \|\cdot\|_1^w\}$*, even with the restriction to distances* d *which satisfy the triangle inequality.*

Proof We reduce from EXACT- t- COVER with $t = (k-1)^2$. Let $S_1, \ldots, S_r$ be subsets of $\{x_1, \ldots, x_n\}$, with $|S_i| = t$, an instance of EXACT- t- COVER and let $\ell := r - \frac{n}{t}$. We construct a graph G for $(\|\cdot\|, \text{diam})$-k-CLUSTER with the following three types of vertices (for an illustration of this construction see Fig. 2):

- $u_1, \ldots, u_n$ representing $x_1, \ldots, x_n$,
- $w_i^1, \ldots, w_i^{k-1}$ representing S_i for $i \in \{1, \ldots, r\}$ and
- $v_1, \ldots, v_\ell$ which will be used to select the ℓ sets which are not in the cover.

The graph G contains the following edges, all of of weight 1:

- edges such that the set $\{u_1, \ldots, u_n\}$ is a clique,
- edges such that the set $\{w_i^1, \ldots, w_i^{k-1}\}$ is a clique for each $i \in \{1, \ldots, r\}$,

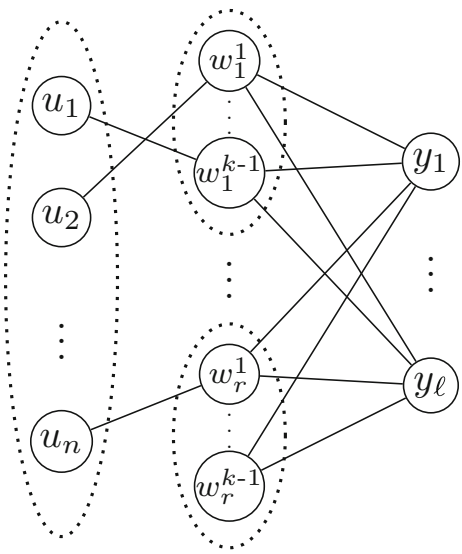


Fig. 2 Illustration of the reduction for Theorem 4. Dotted ellipses surround cliques

– each v_h, $h \in \{1, \ldots, \ell\}$ is connected to all w_i^z with $i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k-1\}$ and
– to model the sets, edges connect u_j to one of the vertices $w_i^1, \ldots, w_i^{k-1}$ if $x_j \in S_i$, more precisely, for every set S_i pick and fix an arbitrary partition $S_i = S_i^1 \cup \cdots \cup S_i^{k-1}$ into disjoint subsets of cardinality $k-1$ and connect u_j with w_i^z if $u_j \in S_i^z$.

We claim that there exists an exact cover for $S_1, \ldots, S_r$ if and only if a there exists a k-cluster of maximum diameter 1 for G.

Let $\mathfrak{P}$ be a k-cluster for G which only contains clusters of diameter 1, and let d be the distance on $V \times V$ induced by the edges of G. Since $d(w_i^y, w_j^z) = 2$ for $i \neq j$ and any $y, z \in \{1, \ldots, k-1\}$ and $d(v_q, v_p) = 2$ for $q \neq p$, each vertex v_h can only be in a cluster of cardinality at least k and diameter 1, if v_h is contained in the cluster $N_i^h := \{v_h, w_i^1, \ldots, w_i^{k-1}\}$ for some $i \in \{1, \ldots, r\}$. The only possibilities for a a cluster of cardinality at least k and diameter 1 which contains a vertex w_i^z is either exactly the cluster $C_i^z := \{w_i^z\} \cup \{u_j : x_j \in S_i^z\}$ or the cluster N_i^h for some $h \in \{1, \ldots, \ell\}$. As $|N_i^h| = |C_i^z| = k$ and $N_i^h \cap C_i^z = \{w_i^z\}$ for all $i \in \{1, \ldots, r\}$, $h \in \{1, \ldots, \ell\}$ and $z \in \{1, \ldots, k-1\}$, it follows that for each $i \in \{1, \ldots, r\}$ either $N_i^h \in \mathfrak{P}$ for some $h \in \{1, \ldots, \ell\}$ or $C_i^z \in \mathfrak{P}$ for all $z \in \{1, \ldots, k-1\}$. As there are exactly $\ell = r - \frac{n}{t}$ vertices v_h, which have to be included in some cluster N_i^h, $\mathfrak{P}$ contains exactly $\frac{n}{t}$ cluster-sets $C_i^1, \ldots, C_i^{k-1}$ which is possible if and only if $\{S_i : C_i^1 \in \mathfrak{P}\}$ is an exact cover; observe that all sets in $\mathfrak{P}$ are disjoint, so the $(k-1)\frac{n}{t}$ sets of type C_i^z in $\mathfrak{P}$ contain exactly $(k-1)(k-1)\frac{n}{t} = n$ vertices from $\{u_1, \ldots, u_n\}$.

Conversely, for every exact cover $S \subseteq \{S_1, \ldots, S_r\}$, the union of the set $\{C_i^1, \ldots, C_i^{k-1} : S_i \in S\}$ and $\{N_{j_h}^h : 1 \leq h \leq \ell\}$ where $\{S_1, \ldots, S_r\} \backslash S =: \{S_{j_1}, \ldots, S_{j_h}\}$ is a k-cluster of diameter 1 for G.

Specific to the norm, it follows that there exists a k-cluster of global cost 1 for $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER if and only if $S_1, \ldots, S_r$ is a "yes"-instance for Exact-t-Cover. Further, each cluster that has the possibility of being of diameter 1 contains exactly k vertices, so $S_1, \ldots, S_r$ is a "yes"-instance for Exact-t-Cover if and only if there exists a solution of global cost k for $(\|\cdot\|_\infty^w, \text{diam})$-$k$-CLUSTER. At last, a solution of global cost $n + r(k-1) + \ell$ for $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER is possible if and only if each cluster has diameter 1, hence if and only if $S_1, \ldots, S_r$ is a "yes"-instance for Exact-t-Cover. □

The reduction shown in the above proof of Theorem 4 is also a gap-reduction with a gap of 2 for the maximum diameter between "yes"- and "no"-instance for EXACT-t- COVER. The maximum cardinality of a cluster in an optimal solution in case of a "yes"-instance for EXACT- t- COVER is k, so the reduction also gives a gap of 2 for the maximum weighted diameter and hence implies:

Corollary 3 *There is no* $(2-\varepsilon)$*-approximation for* $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER *and* $(\|\cdot\|_\infty^w, \text{diam})$-$k$-CLUSTER *in polynomial time for any* $k \geq 3$ *and any* $\varepsilon > 0$*, unless* $P = NP$*, even if* d *satisfies the triangle inequality.*

The construction in the proof of Theorem 3 almost also shows the same hardness result for average distortion. The only problem is that an optimal solution requires clusters of cardinality $k+1$ which means that with respect to $\|\cdot\|_\infty^w$, we have a global cost of

k, which is also achieved by a cluster of cardinality k in which 1 vertex has distance 2 from the central vertex. We will therefore use a third reduction for average distortion which represents each set by $k-1$ vertices as in the construction for diameter and combines this with the idea to use stars with $k-1$ vertices to disable $r - \frac{n}{t}$ sets from being used to "cover" $u_1, \ldots, u_n$, as used for radius.

Theorem 5 *The problem* $(\|\cdot\|, \text{avg})$-k-CLUSTER *is NP-hard for each* $k \geq 3$ *and all choices for* $\|\cdot\| \in \{\|\cdot\|_\infty, \|\cdot\|_\infty^w, \|\cdot\|_1^w\}$*, even with the restriction to distances* d *which satisfy the triangle inequality.*

Proof We reduce from EXACT- t- COVER with $t = (k-1)^2$. Let $S_1, \ldots, S_r$ be subsets of $\{x_1, \ldots, x_n\}$, with $|S_i| = t$, an instance of EXACT- t- COVER. We construct a graph G for $(\|\cdot\|, \text{avg})$-k-CLUSTER with the following vertices (for an illustration of this construction see Fig. 3):

- $u_1, \ldots, u_n$ representing $x_1, \ldots, x_n$,
- $w_i^1, \ldots, w_i^{k-1}$ representing S_i for $i \in \{1, \ldots, r\}$,
- $\bar{w}_i^z$ for all $i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k-1\}$,
- a set of $k-2$ vertices W_i^z for all $i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k-1\}$,
- $v_i, v_i^1, \ldots, v_i^{k-1}$ for all $i \in \{1, \ldots, r\}$ and
- y_i^j for $i \in \{1, \ldots, \frac{n}{t}\}$ and $j \in \{1, \ldots, k-1\}$.

The graph G contains the following edges, all of weight 1:

- like for diameter, pick and fix for every set S_i an arbitrary partition $S_i = S_i^1 \cup \cdots \cup S_i^{k-1}$ into disjoint subsets of cardinality $k-1$ and connect u_j with w_i^z if $u_j \in S_i^z$,

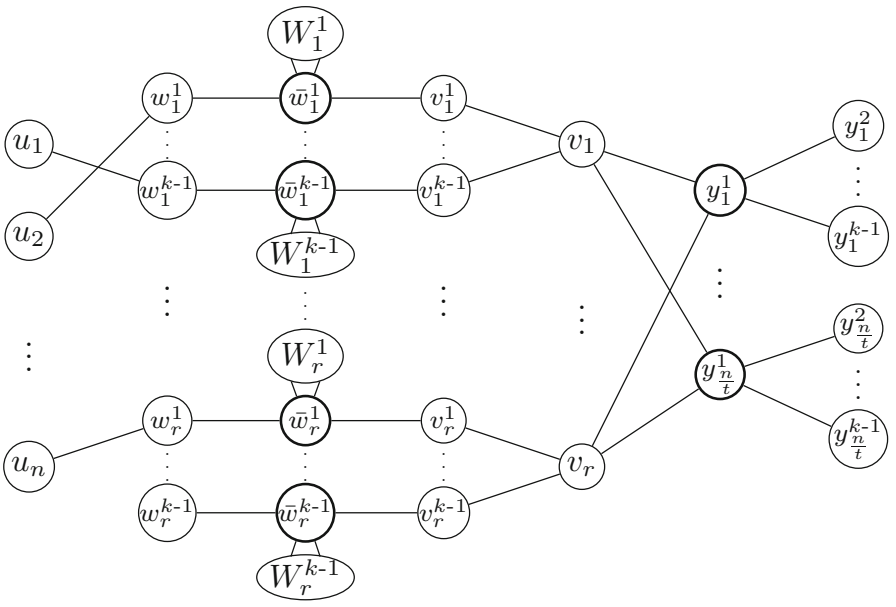


Fig. 3 Illustration of the reduction for Theorem 5. Thick vertices have to be central in a k-cluster of maximum radius 1

- $\{w, \bar{w}_i^z\}$ for all $w \in W_i^z, i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k-1\}$ (the graph induced by the vertices $W_i^z \cup \{\bar{w}_i^z\}$ is a star graph with center $\bar{w}_i^z$),
- $\{v_i, v_i^h\}$ for all $i \in \{1, \ldots, r\}$ and $h \in \{1, \ldots, k-1\}$,
- $\{w_i^z, \bar{w}_i^z\}$ and $\{\bar{w}_i^z, v_i^z\}$ for all $i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, r\}$,
- $\{v_i, y_j^1\}$ for all $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, \frac{n}{t}\}$,
- $\{y_i^1, y_i^h\}$ for each $i \in \{1, \ldots, \frac{n}{t}\}$ and $h \in \{2, \ldots, k-1\}$.

We claim that there exists an exact cover for $S_1, \ldots, S_r$ if and only if a there exists a k-cluster for G such that each cluster has average distortion $\frac{k-1}{k}$.

Let $\mathfrak{P}$ be a k-cluster for G such that each cluster has average distortion $\frac{k-1}{k}$, and let d be the distance on $V \times V$ induced by the edges of G. First of all, observe that any cluster of cardinality at least k has average distortion $\frac{k-1}{k}$ if and only if it hast radius 1 and cardinality k. Similar to the proof of Theorem 3, denote for each $i \in \{1, \ldots, \frac{n}{t}\}$ by P_i the cluster in $\mathfrak{P}$ which contains y_i^2. With the property of P_i having radius 1 and cardinality k for each $i \in \{1, \ldots, \frac{n}{t}\}$, it follows that exactly $\frac{n}{t}$ of the vertices $v_1, \ldots, v_r$ are included in some cluster P_i, which otherwise only contains the vertices $y_i^1, \ldots, y_i^{k-1}$. A similar argument applies for a cluster P which contains a vertex from W_i^r, as these vertices also only have one vertex ($\bar{w}_i^z$) at distance 1, which then has to be central for P; this cluster then always contains the whole set W_i^r. So, denote by P_i^z the cluster containing the set W_i^r and $\bar{w}_i^z$. For each $i \in \{1, \ldots, r\}$ and $r \in \{1, \ldots, k-1\}$, the set P_i^z contains either v_i^z or w_i^z, as these are the only other vertices at distance 1 from $\bar{w}_i^z$. For each of the exactly $r - \frac{n}{t}$ vertices v_i which are not contained in any of the clusters $P_1, \ldots, P_{\frac{n}{t}}$, the only option for a cluster of cardinality k and radius 1 is the cluster $V_i := \{v_i, v_i^1, \ldots, v_i^{k-1}\}$; observe that a vertex v_i^h with $h \in \{1, \ldots, k-1\}$ cannot be central for a cluster of cardinality $k \geq 3$ as the only vertices at distance 1 from v_i^h are v_i and $\bar{w}_i^z$ and the latter one already has to be the central vertex for P_i^z. Also, the only vertices at distance 1 from v_i which are not in some cluster P_j are $v_i^1, \ldots, v_i^{k-1}$. Hence there are exactly $r - \frac{n}{t}$ indices i in $\{1, \ldots, r\}$ such that V_i is a cluster in $\mathfrak{P}$. For all $i \in \{1, \ldots, r\}$ for which V_i is a cluster in $\mathfrak{P}$, the cluster P_i^z contains w_i^z for all $z \in \{1, \ldots, k-1\}$ since v_i^z is not available as k-th vertex in P_i^z. Again similar to the proof of Theorem 3, there are exactly enough vertices w_i^z not included in a set of the form P_i^z in $\mathfrak{P}$ to build clusters of radius 1 for the vertices $\{u_1, \ldots, u_n\}$ if and only if the sets S_i with indices $i \in \{1, \ldots, r\}$ for which V_i is not a cluster in $\mathfrak{P}$ are an exact cover.

Conversely, for every exact cover $S \subseteq \{S_1, \ldots, S_r\}$, a k-cluster of average distortion $\frac{k-1}{k}$ for G can be built with the following sets:

- $\{w_i^z\} \cup \{u_j : x_j \in S_i^z\}$ and $P_i^z \cup \{v_i^z\}$ for all i with $S_i \in S, z \in \{1, \ldots, k-1\}$,
- V_i and $P_i^z \cup \{w_i^z\}$ for all i with $S_i \notin S, z \in \{1, \ldots, k-1\}$ and
- $\{v_{j_i}, y_i^1, \ldots, y_i^{k-1}\}$ for all $i \in \{1, \ldots, \frac{n}{t}\}$ with $S = \{S_{j_1}, \ldots, S_{j_{\frac{n}{t}}}\}$.

So, there exists an exact cover for $S_1, \ldots, S_r$ if and only if there exists a solution for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER of global cost $\frac{k-1}{k}$. Further, for any k-cluster for G, a global cost of $k-1$ with respect to average distortion and $\|\cdot\|_\infty^w$ is only possible if each cluster has radius 1 and cardinality k; a cluster with $k' > k$ vertices gives a global cost of at least $k' - 1 > k - 1$ and a cluster of radius larger than 1 contains at least one vertex at

distance 2 from the central vertex which gives a global cost of at least k. So, there exists an exact cover for $S_1, \ldots, S_r$ if and only if there exists a solution for $(\|\cdot\|_\infty^w, \text{avg})$-$k$-CLUSTER of global cost $k-1$. At last, for any k-cluster for G, a global cost of $n+r(k-1)k$ with respect to average distortion and $\|\cdot\|_\infty^1$ is only possible if each vertex contributes exactly the minimum cost of $\frac{k-1}{k}$ to the global cost. Hence there exists an exact cover for $S_1, \ldots, S_r$ if and only if there exists a solution for $(\|\cdot\|_1^w, \text{avg})$-$k$-CLUSTER of global cost $n + r(k-1)k$. □

In the above reduction used to prove Theorem 5, a "yes"-instance for EXACT- t-COVER corresponds to a graph for which there exists a k-cluster of maximum weighted average $k-1$ while a "no"-instance for EXACT- t- COVER corresponds to graph for which the maximum weighted average of any k-cluster is at least k. This gives the following result.

Corollary 4 *There is no* $(\frac{k}{k-1} - \varepsilon)$*-approximation for* $(\|\cdot\|_\infty^w, \text{avg})$-$k$-CLUSTER *in polynomial time for any* $k \geq 3$ *and any* $\varepsilon > 0$*, unless* $P = NP$*, even if* d *satisfies the triangle inequality.*

If we consider instances of $(\|\cdot\|, f)$-k-CLUSTER for which the induced distance d can violate the triangle inequality, additional edges of a large weight w in the constructions for Theorems 3 and 4 can be used to amplify the gap between a "yes"- and a "no"-instance of EXACT- t- COVER strictly monotonically with w which gives:

Proposition 6 *If* d *violates the triangle inequality, there is no polynomial constant-factor approximation for* $(\|\cdot\|, f)$-k-CLUSTER*, for all choices of* $f \in \{rad, diam, avg\}$ *and* $\|\cdot\| \in \{\|\cdot\|_1^w, \|\cdot\|_\infty^w, \|\cdot\|_\infty\}$*, unless* $P = NP$.

Proof Let $G = (V, E)$ be the graph constructed in the proof of Theorem 3 for a given instance I of EXACT- k- COVER, so there exists a k-cluster of maximum radius 1 for G if and only if I is a "yes"-instance. Further, every k-cluster of maximum radius 1 for G only contains sets of maximum cardinality $k+1$. If I is a "no"-instance, any k-cluster for G contains at least one set S of radius larger than 1, so for every choice of $v \in S$ there exists at least one vertex $v' \in S \backslash \{v\}$ such that $\{v, v'\} \notin E$. If we now turn the graph G into a complete graph with additional edges of weight w, it follows that the radius of such a cluster S is w. This also means that the average distortion for such a cluster S is larger than $\frac{w}{n}$, while the minimum average distortion of a k-cluster for G is $\frac{k}{k+1}$ in case I is a "yes"-instance. For every norm, the global cost grows strictly monotonically with the local cost. This means that the gap between I being a "yes" or "no"-instance for the optimum value of a k-cluster for G with respect to radius or average distortion with any norm grows strictly monotonically with w. As this is true for every value of w, a constant-factor approximation in polynomial time for $(\|\cdot\|, f)$-k-CLUSTER with $f \in$ {rad,avg} with any norm would solve EXACT- k-COVER which however is NP-hard for any $k \geq 3$.

For diameter, we use the same idea and turn the graph G constructed in the proof of Theorem 4 into a complete graph by adding edges of weight w. Similarly, it follows that there exists a k-cluster of maximum diameter 1 if the corresponding instance I of EXACT- t- COVER is a "yes"-instance, while the maximum diameter is w if I is a "no"-instance. So, a constant-factor approximation in polynomial time for $(\|\cdot\|, \text{diam})$-k-CLUSTER with any norm would solve the NP-hard problem EXACT- t- COVER. □

The previous section only provided polynomial-time solvability for roughly half of the variants of $(\|\cdot\|, f)$-2-CLUSTER . We will now complete the complexity picture for $k = 2$. For $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER we give a reduction that does not only show NP-hardness for the associated decision problem but also proves APX-hardness for the optimisation problem. We reduce from the restriction of the vertex cover problem to cubic graphs, formally defined by:

CUBIC VERTEX COVER	
Input:	Graph $G = (V, E)$ such that all vertices $v \in V$ have degree 3.
Output:	A set $C \subseteq V$ (*vertex cover*) of minimum cardinality such that $e \cap C \neq \emptyset$ for all $e \in E$.

CUBIC VERTEX COVER is APX-hard by [16].

Theorem 6 $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER *is APX-hard, even with the restriction to distances d which satisfy the triangle inequality.*

Proof Let $G = (V, E)$ with $V = \{v_1, \ldots, v_n\}$ and $m := |E|$ be the input for CUBIC VERTEX COVER. We construct a graph $G' = (V', E')$ for $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER with vertex set $V' := \{v_i^1, v_i^2 \colon 1 \leq i \leq n\} \cup \{v_e \colon e \in E\}$ and edge set $E' := \{\{v_i^1, v_i^2\} \colon 1 \leq i \leq n\} \cup \{\{v_i^1, v_e\} \colon v_i \in e\}$ with weights $w_E(\{v_i^1, v_i^2\}) = 1$ and $w_E(\{v_i^1, v_e\}) = 2$. We claim that G has a vertex cover of cardinality ℓ if and only if there exists a solution for $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER with global cost $2n + 2\ell + 2m$.

For any vertex cover C of G, we construct a 2-cluster for G' by first building clusters $\{v_i^1, v_i^2\}$ for all $i \in \{1, \ldots, n\}$. We then pick (arbitrarily, if there is a choice) for every edge $e = \{u, w\} \in E$ an index $i \in \{1, \ldots, n\}$ such that $v_i \in C$ and $v_i \in \{u, w\}$ and add the vertex v_e to the cluster $\{v_i^1, v_i^2\}$. As C is a vertex cover for G, we can assign each vertex v_e in such a way and arrive at a 2-cluster $\mathfrak{P}$ for G' which contains only the following two types of clusters:

- $\{v_i^1, v_i^2\} \in \mathfrak{P}$ for all $i \in \{1, \ldots, n\}$ with $v_i \notin C$,
- for all $i \in \{1, \ldots, n\}$ with $v_i \in C$, $\mathfrak{P}$ contains a cluster P_i with $\{v_i^1, v_i^2\} \subseteq P_i$ and $P_i \setminus \{v_i^1, v_i^2\} \subseteq \{v_e \colon \exists 1 \leq j \leq n \colon e = \{v_i, v_j\}\}$. With v_i^1 considered as central vertex, P_i has radius at most 2, as all vertices v_e with $e = \{v_i, v_j\}$ for some $j \in \{1, \ldots, n\}$ have distance 2 from v_i^1.

Considering, w.l.o.g., a vertex numbering such that $C = \{v_1, \ldots, v_\ell\}$, the global cost of $\mathfrak{P}$ with respect to radius and weighted 1-norm is hence at most:

$$\sum_{i=\ell+1}^{n} 2 \cdot \text{rad}(\{v_i^1, v_i^2\}) + \sum_{i=1}^{\ell} |P_i| \cdot \text{rad}(P_i) \;\leq\; 2(n - \ell) + 2 \cdot \sum_{i=1}^{\ell} |P_i|.$$

As the union of all the clusters P_i with $i \in \{1, \ldots, \ell\}$ contains exactly all vertices v_e, $e \in E$ and all vertices v_i^1, v_i^2 with $i \in \{1, \ldots, \ell\}$, it follows that $\sum_{i=1}^{\ell} |P_i| = |E| + 2\ell$. The global cost of $\mathfrak{P}$ as solution for $(\|\cdot\|_1^w, \text{rad})$-2-CLUSTER is hence at most $2(n - \ell) + 2(m + 2\ell) \;=\; 2n + 2\ell + 2m$.

Conversely, let $\mathfrak{P}$ be a 2-cluster for G' such that the global cost with respect to radius and 1-norm is at most $2n+2\ell+2m$. We define for this solution a cost-function c on V' by $c(v) := \mathrm{rad}(P)$ for all $v \in P$ and $P \in \mathfrak{P}$. The global cost of $\mathfrak{P}$ with respect to $(\|\cdot\|_1^w, \mathrm{rad})$ can hence be computed by $\sum_{v\in V'} c(v)$. Observe first that from the structure of the graph G' it immediately follows that $c(v) \geq 1$ and $c(v) \in \mathbb{N}$ for all $v \in V'$. We consider the possible costs $c(v)$ for all types of vertices:

- For all $i \in \{1,\ldots,n\}$ and $h \in \{1,2\}$, $c(v_i^h) = 1$ if and only if $\{v_i^1, v_i^2\} \in \mathfrak{P}$.
- For all $e \in E$, we know that $c(v_e) \geq 2$.
- For any $e = \{v_i, v_j\} \in E$, $c(v_e) = 2$ is only possible if $\{v_i^1, v_j^1\} \cap P \neq \emptyset$ for the set $P \in \mathfrak{P}$ with $v_e \in P$.
- For any $e = \{v_i, v_j\} \in E$, $d(v_e, w) = 3$ is only possible if $\{v_i^2, v_j^2\} \cap P \neq \emptyset$ for the set $P \in \mathfrak{P}$ with $v_e \in P$.

Assume that $C := \{v_i : c(v_i^1) \geq 2\}$ is not a vertex-cover of size ℓ for G. If $|C| > \ell$, we see that, since $c(v_i^1) \geq 2$ if and only if $c(v_i^2) \geq 2$, the global cost of $\mathfrak{P}$ exceeds the assumed value, as:

$$\begin{aligned}\sum_{v\in V'} c(v) \;&\geq \sum_{i=1}^{n} (c(v_i^1) + c(v_i^2)) + 2m \geq 2\cdot 2|C| \\ &\quad + 2(n-|C|) + 2m > 2n + 2\ell + 2m.\end{aligned}$$

If there is some edge $e = \{v_i, v_j\}$ which is not covered by C, the sets $\{v_i^1, v_i^2\}$ and $\{v_j^1, v_j^2\}$ are both in $\mathfrak{P}$ by the definition of C, hence $c(v_e) \geq 4$. So let $\bar{E} \subseteq E$ be the set of edges which are not covered by C. It follows that:

$$\begin{aligned}2n+2\ell+2m &\geq \sum_{v_i\in C}(c(v_i^1)+c(v_i^2)) + \sum_{v_i\notin C}(c(v_i^1)+c(v_i^2)) + \sum_{e\in\bar{E}} c(v_e) + \sum_{e\in E\setminus\bar{E}} c(v_e) \\ &\geq 4\cdot|C| + 2(n-|C|) + 4\cdot|\bar{E}| + 2(m-|\bar{E}|) \\ &= 2n + 2m + 2(|C| + |\bar{E}|)\end{aligned}$$

This means that $|C| \leq \ell - |\bar{E}|$, so if C is not already a vertex cover for G, we can greedily chose for each edge in $\bar{E}$ an arbitrary adjacent vertex to cover it and arrive at a vertex cover for G of cardinality at most ℓ.

At last, since $m = 3n/2$ and $\ell \geq n/2$ for a cubic graph, we have $2n+2\ell+2m \leq 12k$ which makes this reduction an L-reduction and hence translates the APX-hardness from CUBIC VERTEX COVER to $(\|\cdot\|_1^w, \mathrm{rad})$-2-CLUSTER . □

The reduction above cannot be adapted for the cases of $(\|\cdot\|, f)$-2-CLUSTER which were not shown to be polynomial-time solvable so far. We therefore consider the following variation of SATISFIABILITY for the remaining cases:

(3, 3)-SATISFIABILITY (or (3, 3)-SAT)

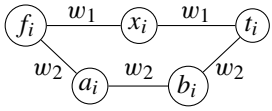
Input: Boolean formula F in conjunctive normal form such that each clause contains at most 3 literals and each variable occurs both positively and negatively in F and overall at most 3 times.

Question: Does there exist a satisfying assignment for F?

(3, 3)-SAT is NP-hard by [21].

Theorem 7 *$(\|\cdot\|_\infty^w, \text{avg})$-, $(\|\cdot\|_\infty, \text{avg})$ - and $(\|\cdot\|_\infty^w, \text{rad})$-2-CLUSTER are NP-hard, for the latter two even with the restriction to distances d which satisfy the triangle inequality.*

Proof Let $v_1, \ldots, v_n$ be the variables and $c_1, \ldots, c_m$ be the clauses of a (3, 3)-SAT formula F. We construct a graph $G = (V, E)$ by introducing five vertices t_i, f_i, x_i, a_i, b_i for each v_i and edges $\{x_i, f_i\}, \{x_i, t_i\}$ of weight w_1 and $\{a_i, f_i\}, \{b_i, t_i\}, \{a_i, b_i\}$ of weight w_2 as in the picture below.



Also, for each clause c_j, introduce a vertex y_j and edges of weight w_2 from y_j to each literal in c_j, i.e., to f_i if $\bar{v}_i$ is a literal in c_j and to t_i if v_i is a literal in c_j. We will assign values for w_1 and w_2 differently for each problem variant such that a 2-cluster for G has global cost (and hence maximum (weighted) cost of each cluster) at most 1 if and only if the following *assignment properties* hold:

- Each x_i has to be in a cluster of cardinality 2 with either t_i or f_i (this reflects the assignment for v_i to be the vertex not clustered with x_i).
- Each y_j is in a cluster with 1 adjacent vertex, so t_i (or f_i) for some i with v_i ($\bar{v}_i$) being a literal in c_j (this literal satisfies the clause).
- For all $i \in \{1, \ldots, n\}$, the vertices a_i and b_i lie in the same cluster which otherwise can only possibly contain either t_i or f_i (in case we do not need the variable v_i to satisfy any clause).

Assuming $w_1 \geq w_2$, the induced distance d on G satisfies:

- $d(x_i, v) \geq w_1 + w_2$ for all $v \in V \backslash \{x_i, t_i, f_i\}$,
- $d(t_i, y_j) \geq 3w_2$ for all i, j such that v_i is no literal in c_j,
- $d(f_i, y_j) \geq 3w_2$ for all i, j such that $\bar{v}_i$ is no literal in c_j,
- $d(y_j, v) \geq 2w_2$ for all $v \in V \backslash \{t_i, f_i : 1 \leq i \leq n\}$.

These distances imply that a 2-cluster which does not satisfy the assignment properties contains at least one cluster of either a cardinality at least 3 and radius at least w_1 (some vertex x_i not properly clustered), or a radius of at least $2w_2$ (some vertex y_j not in a cluster with adjacent vertex). We now consider each problem variant and define respective weights w_1 and w_2.

For $(\|\cdot\|_\infty^w, \text{rad})$-2-CLUSTER we choose $w_1 = \frac{1}{2}$ and $w_2 = \frac{1}{3}$. With these weights, a cluster P in a 2-cluster for G of weighted radius at most 1 can have radius w_1 only if it

has cardinality 2 and otherwise has radius w_2 and cardinality at most 3. As $2w_2 > w_1$, a solution for $(\|\cdot\|_\infty^w$, rad)-2-CLUSTER of global cost at most 1 fulfils the assignment properties.

For $(\|\cdot\|_\infty$, avg)-2-CLUSTER we choose $w_1 = 2$ and $w_2 = \frac{3}{2}$. With these weights, all pairs of distinct vertices in G have a distance at least $\frac{3}{2}$ and hence the average distortion of every cluster is at least $\frac{3}{2}(|P| - 1)/|P|$, which means that the maximum cardinality of a cluster of average distortion 1 is 3. Also, a cluster of cardinality 3 has average distortion at most 1 only if it has radius $\frac{3}{2} = w_2$. A cluster of cardinality 2 has average distortion at most 1 only if it has radius at most $2 = w_1$. As again $2w_2 > w_1$, this means that a solution for $(\|\cdot\|_\infty$, avg)-2-CLUSTER of global cost at most 1 fulfils the assignment properties.

For $(\|\cdot\|_\infty^w$, avg)-2-CLUSTER we choose $w_1 = 1$ and $w_2 = \frac{1}{2}$ but also have to add some more edges; observe that so far, the induced distance d satisfies the triangle inequality, so by Proposition 5, a 2-cluster could be computed in polynomial time, hence our construction cannot be complete. With the current definition we have $2w_2 = w_1$ which yields that clusters of the form $\{y_i, y_j\}$ or $\{a_i, y_j\}$ could also have a weighted average distortion of 1 as there could be a shortest path from y_j to y_i or a_i via two edges of weight $\frac{1}{2}$. If we add edges $\{y_i, y_j\}$ for all $i \neq j$ and $\{a_i, y_j\}, \{b_i, y_j\}$ for all $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$ each of weight 2, these types of clusters have a weighted average distortion of 2. Other clusters in a 2-cluster which yield a violation of the assignment property have either cardinality at least 3 and radius at least w_1, which yields a weighted average distortion of at least $\frac{3}{2}$, or radius at least $\min\{w_1 + w_2, 3w_2\} = \frac{3}{2}$, so weighted average distortion at least $\frac{3}{2}$. A solution for $(\|\cdot\|_\infty^w$, avg)-2-CLUSTER of global cost at most 1 fulfils the assignment properties.

Finally, there exists a 2-cluster with assignment properties for G (for any choice of the weights w_1, w_2) if and only if the formula F is satisfiable:

Given a 2-cluster $\mathfrak{P}$ for G with assignment property, the vertices of G corresponding to the clauses are clustered with their satisfying literal and for each variable v_i either $\{t_i, x_i\} \in \mathfrak{P}$ or $\{f_i, x_i\} \in \mathfrak{P}$, so the assignment $v_i = \mathit{true}$ if and only if $\{f_i, x_i\} \in \mathfrak{P}$ is a satisfying assignment for F.

Conversely, given a satisfying assignment ϕ for F, build a partition from the union of the sets $\{\{x_i, t_i\}, \{f_i\} \colon \phi(v_i) = \mathit{false}\}, \{\{x_i, f_i\}, \{t_i\} \colon \phi(v_i) = \mathit{true}\}$ and $\{\{a_i, b_i\} \colon 1 \leq i \leq n\}$, and put for each $j \in \{1, \ldots, m\}$ the vertex y_j into the cluster which contains the assignment of the literal (an arbitrary literal if there is a choice) which satisfies c_j, i.e., if v_i ($\bar{v}_i$) is a literal in c_j and $\phi(v_i) = \mathit{true}$ ($\phi(v_i) = \mathit{false}$) put y_j in the cluster containing t_i (f_i). As F is an instance of (3,3)-SAT, at most 2 clause-vertices are assigned to the same cluster. If there is some i such that $\{t_i\}$ or $\{f_i\}$ remains a cluster of cardinality 1, merge this cluster with $\{a_i, b_i\}$. The resulting partition is a 2-cluster with assignment properties for G. □

6 Approximation Results

We will now discuss polynomial time approximations for $(\|\cdot\|, f)$-k-CLUSTER but only consider the case where d satisfies the triangle inequality in this section. This restriction is not just reasonable in most scenarios but in some sense necessary to

achieve any kind of approximation as Proposition 6 indicates. In the following, we denote the global cost of an optimal solution for $(\|\cdot\|, f)$-k-CLUSTER on G with distance d by $\text{opt}(G, d, \|\cdot\|, f, k)$.

Known approximation results for clustering with size constraints include a 9-approximation from [4] for LOAD BALANCED FACILITY LOCATION without facility cost, which is related to $(\|\cdot\|_1^w, \text{avg})$-$k$-CLUSTER here, but with the additional constraint that at each customer should be assigned to the nearest open facility. The techniques used for this result highly rely on the additional constraint, which unfortunately means that they can not be applied here. Other approximations for this problem instead relax the constraint that each cluster needs to contain at least k vertices; Guha et al. [13] for example presents a $2k$-approximation which constructs clusters of cardinality at least $k/3$. We will see that for our problem such an approximation factor can be achieved without relaxing the cardinality constraints. In general, our results however do not extend to LOAD BALANCED FACILITY LOCATION, since the addition of facility-costs yields a very different type of problem; we especially lose the upper bound of $2k - 1$ on the cardinality of clusters in an optimal solution from Theorem 1.

Other known approximation results however also apply here and can be altered for other problem variants. The problem that we call $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER is discussed under the name r-GATHER in [2], where r takes the role of k. The concept for the 2-approximation presented there can be altered and also used to compute a 2-approximation for $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER .

Theorem 8 $(\|\cdot\|_\infty, \text{rad})$- *and* $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER *are 2-approximable in polynomial time for all* $k \geq 2$*, if* d *satisfies the triangle inequality.*

Proof Let $G = (V, E)$ be the input graph with induced distance d. By a binary search among all values in $\{d(v, v') : v, v' \in V\}$, we search for the smallest value D such that the procedure described below to build a k-cluster of maximum radius D is successful.

For a fixed D, we first build a partition of V in the following way: Beginning with $i = 1$ and $V_1 := V$ we iteratively, until $V_i = \emptyset$, choose arbitrarily some $c_i \in V_i$ and build clusters $P(c_i) := \{v \in V_i : d(c_i, v) \leq D\}$ and set $V_{i+1} = V_i \backslash P(c_i)$. This yields a partition of V into a finite number of clusters $P(c_i)$. Let z be the number of clusters created by this strategy. If each cluster $P(c_i)$ contains at least k vertices, we have found a k-cluster of maximum radius D.

Some of the clusters $P(c_i)$ however might have a cardinality of less than k. In this case, we try to reassign some vertices to adjust the cardinalities. Observe that by the strategy used to build the clusters, possible vertices outside $P(c_j)$ at distance at most D from c_j can only be in clusters $P(c_i)$ with $i < j$. Hence, we define the sets $S(i, j) := \{v \in P(c_j) \backslash \{c_j\} : d(v, c_i) \leq D\}$ for all $1 \leq j < i$ to collect all vertices which can be moved from cluster $P(c_j)$ to cluster $P(c_i)$ without increasing the radius of $P(c_j)$ to be more than D. If $\sum_{i=1}^{j-1} |S(i, j)| < k - |P(c_j)|$ for some $j \in \{1, \ldots, z\}$ there are not enough vertices to move to cluster $P(c_j)$ and we delete this clustering-attempt and try again for a larger value for D. Otherwise, we try to move some vertices in $S(i, j)$ from $P(c_j)$ to $P(c_i)$, $1 \leq j < i \leq z$, in order to arrive at a partition which is a k-cluster. Moving some vertices from $S(i, j)$ into $P(c_i)$ to increase the cardinality of $P(c_i)$ might mean that the cardinality of $P(c_j)$ decreases below k and hence requires moving some vertices from $S(j, \ell)$ into $P(c_j)$ for some

$\ell < j$. This kind of ripple effect is the reason why we solve this problem of moving vertices in $S(i, j)$ to create a k-cluster by modelling it as a max-flow problem with the following network:

- The network has a source s and target t.
- For each $i \in \{1, \ldots, z\}$ we create a vertex p_i representing $P(c_i)$ in the network. If $|P(c_i)| > k$ we add an arc from s to p_i of capacity $|P(c_i)| - k$. If $|P(c_i)| < k$ we add an arc from p_i to t of capacity $k - |P(c_i)|$ to t.
- For each $v \in \bigcup_{i=1}^{z-1} \bigcup_{j=1}^{i-1} S(i, j)$, create a vertex v' in the network with an arc of capacity 1 from v' to p_i for all i with $v \in S(i, j)$ for some j and an arc of capacity 1 from p_j to v' for all j with $v \in S(i, j)$ for some i.

There exists a maximum flow of $\sum_{i=2}^{z} \max\{0, k - |P(c_i)|\}$ from s to t in this network if and only if we can find a reassignment of the vertices in the sets $S(i, j)$ to turn $P(c_1), \ldots, P(c_z)$ into a k-cluster: Moving a vertex $v \in S(i, j)$ from $P(c_i)$ to $P(c_j)$ corresponds to a flow of 1 in the network from p_i to v' and then to p_j. If $|P(c_i)| > k$, at most $|P(c_i)| - k$ vertices are allowed to be moved out of $P(c_i)$ which corresponds to the capacity of the arc from s to p_i. If $|P(c_i)| < k$, exactly $k - |P(c_i)|$ vertices have to be moved into $P(c_i)$, saturating the capacities of the arc from p_i to t. MAX FLOW can be solved in time $O(m \cdot n)$ [14,15] on a directed graph with n vertices and m edges. If we find a flow of size $\sum_{i=2}^{z} \max\{0, k - |P(c_i)|\}$, we can build a k-cluster for V with maximum radius D and maximum diameter $2D$, otherwise we abort and try a larger value for D.

We claim that the procedure described above is successful for $D = 2r^*$ with $r^* = \text{opt}(G, d, \|\cdot\|_\infty, \text{rad}, k)$. The vertices c_i chosen while computing a solution for $D = 2r^*$ belong to different clusters in an optimal solution, since vertices in the same cluster have a distance of at most $2r^*$ (observe that this is false if d violates the triangle inequality). Since at most one vertex from each optimal cluster was chosen to be some c_j, there are enough vertices at distance at most $2r^*$ from each such vertex to distribute them among the sets $P(c_j)$ such that each has a cardinality of at least k. A similar reasoning proves that the greedy procedure is successful for $D = d^*$ with $d^* = \text{opt}(G, d, \|\cdot\|_\infty, \text{diam}, k)$. In case of diameter, the vertices c_j can not belong to the same cluster in the optimal solution as soon as their distance is larger than d^*. □

Remark 2 A natural greedy procedure for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER could build up the sets $P(c_i)$ by successively adding $\text{argmin}\{d(v, c_i) : v \in V_i \setminus P(c_i)\}$ until $\text{avg}(P(c_i))$ exceeds D, but moving vertices from $S(i, j)$ to $P(c_i)$ could unfortunately increase the average distortion of $P(c_j)$.

The general class of constraint forest problems introduced in [12] also has a close relation to clustering with lower bounds. We will now use the following problem from the class of constraint forest optimisation problems:

LOWER CAPACITATED TREE PARTITIONING	
Input:	Graph $G = (V, E)$, edge-weights $w_E \colon E \to \mathbb{Q}_+$, capacity $k \in \mathbb{N}$.
Output:	A set $E' \subseteq E$ minimising $\sum_{e \in E'} w_E(e)$ such that each $v \in V$ occurs in at least one $e \in E'$ and each component in the graph induced by E' is a tree with at least k vertices.

LOWER CAPACITATED TREE PARTITIONING is 2-approximation in polynomial time by an application of the approximation presented in [12].

Proposition 7 *($\|\cdot\|_1^w$, avg)-k-CLUSTER is $2k$-approximable in polynomial time for all $k \geq 2$, if d satisfies the triangle inequality.*

Proof Let $G = (V, E)$ be an instance of ($\|\cdot\|_1^w$, avg)-k-CLUSTER with induced distances d. We consider solving LOWER CAPACITATED TREE PARTITIONING with capacity k on $G' = (V, V \times V)$ with edge-weights computed via d. Any solution $P_1, \ldots, P_s$ for ($\|\cdot\|_1^w$, avg)-k-CLUSTER of global cost L on $G = (V, E)$ can be interpreted as a solution of cost L for LOWER CAPACITATED TREE PARTITIONING on G'; a spanning forest for G' with connected components $P_1, \ldots, P_s$ and cost $\|(\text{avg}(P_1), \ldots, \text{avg}(P_s))\|_1^w$ is given by the edge-set:

$$\bigcup_{i=1}^{s} \{\{c_i, v_i\} \colon v_i \in P_i\} \text{ with } c_i = \operatorname{argmin}\left\{ \sum_{v \in P_i} d(v, c) \colon c \in P_i \right\}.$$

Conversely, any minimal solution $\bar{E}$ for LOWER CAPACITATED TREE PARTITIONING for G' of cost L can be interpreted as a solution for ($\|\cdot\|_1^w$, avg)-k-CLUSTER with global cost at most $k \cdot L$. Let $\mathfrak{C}$ be the set of connected components of the graph induced by $\bar{E}$. Any component $C \in \mathfrak{C}$ which contains a path with more than $2k - 1$ vertices can be split into two connected components, each of cardinality at least k by deleting an edge (i.e., reducing the cost of the partition). We can hence assume that for all components $C \in \mathfrak{C}$ there is at least one $c \in C$ such that every $v \in C$ can be reached from c travelling via at most k edges in C, for example, a vertex in the middle of a longest path in C. This implies:

$$\sum_{C \in \mathfrak{C}} |C| \cdot \text{avg}(C) \leq \sum_{C \in \mathfrak{C}} \sum_{v \in C} d(v, c) \leq \sum_{C \in \mathfrak{C}} \sum_{\{u,v\} \in E(C)} k \cdot d(u, v) = k \cdot L.$$

Since LOWER CAPACITATED TREE PARTITIONING can be 2-approximated, this yields a $2k$-approximation for ($\|\cdot\|_1^w$, avg)-k-CLUSTER . □

Remark 3 Theorem 2 showed that ($\|\cdot\|_1^w$, avg)-2-CLUSTER can be solved in polynomial time which also translates to LOWER CAPACITATED TREE PARTITIONING with capacity $k = 2$; tree partitioning with capacity 2 is equivalent to weighted edge cover.

Essential for the result above is excluding paths of length $2k$ in all components C, but this property does not prevent C from containing arbitrarily many vertices. For

$(\|\cdot\|_1^w, \text{diam})$- or $(\|\cdot\|_1^w, \text{rad})$-$k$-CLUSTER we need an upper bound on the cardinality to prove an approximation ratio. We therefore consider LOWER CAPACITATED PATH PARTITIONING, the restriction of LOWER CAPACITATED TREE PARTITIONING to paths as connected components. On weighted graphs for which the weights obey the triangle inequality, Goemans and Williamson [12] provides a 4-approximation for LOWER CAPACITATED PATH PARTITIONING.

Proposition 8 $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER *is* $8(k-1)$*-approximable in polynomial time for all* $k \geq 2$*, if* d *satisfies the triangle inequality.*

Proof Consider for any input $G = (V, E)$ with induced distances d for the problem $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER , the complete graph $G' = (V, V \times V)$ with d as input for path partitioning. Let $P_1, \ldots, P_s$ be an optimal solution for $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER with $|P_i| \leq 2k-1$ (transformed with Proposition 1). For each $i \in \{1, \ldots, s\}$, a cheapest spanning path for P_i has a cost of at most $(|P_i|-1) \cdot \text{diam}(P_i)$. Building a cheapest spanning path for each set P_i hence gives a solution of LOWER CAPACITATED PATH PARTITIONING on G' of cost at most

$$\begin{aligned}\sum_{i=1}^{s}(|P_i|-1)\cdot\text{diam}(P_i) &= \sum_{i=1}^{s}\frac{|P_i|-1}{|P_i|}\cdot|P_i|\cdot\text{diam}(P_i)\\ &\leq \tfrac{2k-2}{2k-1}\cdot\sum_{i=1}^{s}|P_i|\cdot\text{diam}(P_i)\\ &= \tfrac{2k-2}{2k-1}\cdot\text{opt}(G,d,\|\cdot\|_1^w,\text{diam},k).\end{aligned}$$

This especially implies that the cost T^* of an optimal path partitioning for G' is at most $\frac{2k-2}{2k-1}\cdot\text{opt}(G,d,\|\cdot\|_1^w,\text{diam},k)$.

Let $\tilde{E} \subseteq V \times V$ be a solution for LOWER CAPACITATED PATH PARTITIONING for G' of cost T. Let $P'_1, \ldots, P'_s$ be the vertex sets corresponding to the connected components of the graph induced by $\tilde{E}$. The partition $P'_1, \ldots, P'_s$ yields a solution for $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER of global cost at most $(2k-1)T$; observe that any set P'_i contains at most $2k-1$ vertices as a path containing more than $2k-1$ vertices can be split into 2 paths by deleting an edge from $\tilde{E}$. Considering $\bar{E}$ to be a 4-approximation for LOWER CAPACITATED PATH PARTITIONING on G' computed with [12], the partition $P'_1, \ldots, P'_s$ gives a solution for $(\|\cdot\|_1^w, \text{diam})$-$k$-CLUSTER of global cost at most:

$$\begin{aligned}(2k-1)T &\leq (2k-1)4T^*\\ &\leq (2k-1)4\cdot\tfrac{2k-2}{2k-1}\cdot\text{opt}(G,d,\|\cdot\|_1^w,\text{diam},k)\\ &\leq 8(k-1)\cdot\text{opt}(G,d,\|\cdot\|_1^w,\text{diam},k).\end{aligned}$$

□

One advantage of the unified model for $(\|\cdot\|, f)$-k-CLUSTER is that if d satisfies the triangle inequality, the different measures relate in the following way:

$$\operatorname{avg}(P_i) \leq \operatorname{rad}(P_i) \leq \operatorname{diam}(P_i) \leq 2 \cdot \operatorname{rad}(P_i) \tag{1}$$

With this relation, Proposition 8 immediately yields:

Corollary 5 $(\|\cdot\|_1^w, \text{rad})$-$k$-CLUSTER *is* $16(k-1)$*-approximable in polynomial time for all* $k \geq 2$*, if* d *satisfies the triangle inequality.*

By definition, the two ∞-norms also relate optimal values in the following way for every choice of $f \in \{\text{rad},\text{diam},\text{avg}\}$:

$$\operatorname{opt}(G, d, f, \|\cdot\|_\infty^w, k) \geq k \cdot \operatorname{opt}(G, d, f, \|\cdot\|_\infty, k) \tag{2}$$

This equation is helpful to derive approximations for the weighted ∞-norm:

Proposition 9 $(\|\cdot\|_\infty^w, \text{diam})$ *and* $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER *are* 4*-approximable in polynomial time for all* $k \geq 2$*, if* d *satisfies the triangle inequality.*

Proof Let for a given graph G with induced distances d the sets $P_1, \ldots, P_s$ be the 2-approximation for $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER from Theorem 8. By Proposition 1, we can assume that $|P_i| \leq 2k-1$. This yields:

$$\begin{aligned}
&\max\{|P_i| \cdot \operatorname{diam}(P_i) \colon 1 \leq i \leq s\} \\
&\quad \leq (2k-1) \cdot \max\{\operatorname{diam}(P_i) \colon 1 \leq i \leq s\} \\
&\quad \leq 2(2k-1) \cdot \operatorname{opt}(G, d, \text{diam}, \|\cdot\|_\infty, k)
\end{aligned}$$

By Eq. (2) this implies

$$\max\{|P_i| \cdot \operatorname{diam}(P_i) \colon 1 \leq i \leq s\} \leq (4k-2) \cdot \tfrac{1}{k} \cdot \operatorname{opt}(G, d, \text{diam}, \|\cdot\|_\infty^w, k)$$

which makes $P_1, \ldots, P_s$ a 4-approximation for $(\|\cdot\|_\infty^w, \text{diam})$-$k$-CLUSTER .

A similar reasoning can be used with a 2-approximation for $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER in order to compute a 4-approximation for $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER . If a cluster P in the approximate solution for $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER contains more than $2k-1$ vertices, we remove exactly k vertices from it (keeping at least one of its central vertices with respect to radius) and build a new cluster $\bar{P}$ with them. By triangle-inequality this cluster has a radius of at most $2 \cdot \operatorname{rad}(P)$. We repeat this cluster-splitting until all clusters have at most $2k-1$ vertices. Let $P'_1, \ldots, P'_s$ be the clusters created from the approximation $P_1, \ldots, P_s$ for $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER by removing vertices and let $\bar{P}_1, \ldots, \bar{P}_r$ be all newly created clusters of cardinality k. Since at least one central vertex of P_i remains in P'_i, we know that $\operatorname{rad}(P'_i) \leq \operatorname{rad}(P_i)$. This partition yields a solution for $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER of size:

$$\begin{aligned}
&\max\{\max\{|P'_i| \cdot \operatorname{rad}(P_i) \colon 1 \leq i \leq s\}, \max\{|\bar{P}_j| \cdot \operatorname{rad}(\bar{P}_j) \colon 1 \leq j \leq r\}\} \\
&\quad \leq \max\{\max\{(2k-1) \cdot \operatorname{rad}(P'_i) \colon 1 \leq i \leq s\}, \max\{k \cdot \operatorname{rad}(\bar{P}_j) \colon 1 \leq j \leq r\}\} \\
&\quad \leq \max\{\max\{(2k-1) \cdot \operatorname{rad}(P_i) \colon 1 \leq i \leq s\}, \max\{k \cdot (2 \cdot \operatorname{rad}(P_i)) \colon 1 \leq i \leq s\}\} \\
&\quad \leq 2k \cdot \max\{\operatorname{rad}(P_i) \colon 1 \leq i \leq s\} \\
&\quad \leq 4k \cdot \operatorname{opt}(G, d, \text{rad}, \|\cdot\|_\infty, k)
\end{aligned}$$

By Eq. (2), this means that $P'_1, \ldots, P'_s, \bar{P}_1, \ldots, \bar{P}_r$ is a 4-approximation for $(\|\cdot\|_\infty^w, \text{rad})$-$k$-CLUSTER . □

For $(\|\cdot\|_\infty^w, \text{avg})$-$k$-CLUSTER we do not have a result for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER to transfer. Interestingly, a variant with different norm and measure can be used instead:

Proposition 10 $(\|\cdot\|_\infty^w, \text{avg})$-$k$-CLUSTER *is* $(4k-2)$-*approximable in polynomial time for all* $k \geq 2$*, if* d *satisfies the triangle inequality.*

Proof We first show $\text{opt}(G, d, \text{avg}, \|\cdot\|_\infty^w, k) \geq \text{opt}(G, d, \text{diam}, \|\cdot\|_\infty, k)$. Consider any set P in an optimal solution for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER . Triangle inequality yields:

$$\begin{aligned} |P| \cdot \text{avg}(P) &= \min\left\{ \sum_{p \in P} d(c, p) \colon c \in P \right\} \\ &\geq \min\{\max\{d(u, c) + d(v, c) \colon u, v \in P\} \colon c \in P\} \\ &\geq \max\{d(u, v) \colon u, v \in P\} = \text{diam}(P) \end{aligned}$$

Theorem 8 and Proposition 1 produce a 2-approximation for $(\|\cdot\|_\infty, \text{diam})$-$k$-CLUSTER for which each set contains at least $2k-1$ vertices. The global cost of this partition with respect to the weighted ∞-norm and average distortion is at most $2(2k-1) \cdot \text{opt}(G, d, \text{diam}, \|\cdot\|_\infty, k)$, and hence yields a $(4k-2)$-approximation for $(\|\cdot\|_\infty^w, \text{avg})$-$k$-CLUSTER . □

At last, we want to present an approximation which exploits the unified model by combining the solutions for $k = 2$ derived in Sect. 4 for two different problem variants to compute an approximate solution for $k = 4$. Explicitly, we will combine the SIMPLEX MATCHING approach for $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER and the EDGE COVER approach for $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER . For this result, we need the following connection between $(\|\cdot\|_1^w, \text{diam})$-4-CLUSTER and $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER.

Lemma 1 *Let* $P_1, \ldots, P_s$ *with* $|P_i| \leq 3$ *for all* $i \in \{1, \ldots, s\}$ *be an optimal solution for* $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER *on a graph* G *with distance* d*. Let* $G' = (P, P \times P)$ *be the graph with* $P := \{p_1, \ldots, p_s\}$ *and edge-weights* w *defined by* $w_{i,j} := w(\{p_i, p_j\}) := \min\{d(u, v) \colon u \in P_i, v \in P_j\}$*, then:*

$$\text{opt}(G, d, \text{diam}, \|\cdot\|_1^w, 4) \geq 3 \cdot \text{opt}(G', w, \text{avg}, \|\cdot\|_1^w, 2).$$

Proof Let $S_1, \ldots, S_r$ be an optimal solution for $(\|\cdot\|_1^w, \text{diam})$-4-CLUSTER on G and define $c(v) := \text{diam}(S_i)$ for all $v \in S_i$, $i \in \{1, \ldots, r\}$. This yields:

$$D^* := \text{opt}(G, d, \text{diam}, \|\cdot\|_1^w, 4) = \sum_{v \in V} c(v).$$

Let $\tilde{G} = (P, \tilde{E})$ be the restriction of G' (edge-weights inherited) to the edges:

$$\tilde{E} := \bigcup_{k=1}^{r} \{\{p_i, p_j\} \colon (i \neq j) \wedge (P_i \cap S_k \neq \emptyset) \wedge (P_j \cap S_k) \neq \emptyset\}.$$

Observe that $|P_i| \leq 3$ for all $i \in \{1, \ldots, s\}$ implies that the minimum degree in $\tilde{G}$ is 1; each $v \in P_i$ lies in some set S_j, $j \in \{1, \ldots, r\}$, with $|S_j| \geq 4$, so there exists a vertex $v' \in S_j$ with $v' \in P_{i'}$ and $i' \neq i$ which yields $\{p_i, p_{i'}\} \in \tilde{E}$. By the definition of $\tilde{G}$, we know that, for any $v \in P_i$

$$c(v) \geq \min\{w_{i,j} \colon 1 \leq j \leq s, \{p_i, p_j\} \in \tilde{E}\}. \tag{3}$$

Let $C \subseteq \tilde{E}$ be a minimum-weight edge cover for $\tilde{G}$. We claim that $3 \cdot w(C) \leq D^*$ and consider three cases for edges C based on the cardinality of the neighbourhoods of vertices p_i in C, formally defined by $N_C(i) := \{r \colon \{p_i, p_r\} \in C\}$. First observe that if $|N_C(i)| > 1$, minimality of C yields:

$$w_{i,j} \leq \min\{w_{h,j} \colon 1 \leq h \leq s, \{p_h, p_j\} \in \tilde{E}\} \text{ for all } j \in N_C(i). \tag{4}$$

Case 1: $|N_C(i)| = |N_C(j)| = 1$ for some $j \in \{1, \ldots, s\}$ with $\{p_i, p_j\} \in C$. As $\{p_i, p_j\} \in \tilde{E}$, there exists some $k \in \{1, \ldots, r\}$ such that $P_i \cap S_k \neq \emptyset$ and $P_j \cap S_k \neq \emptyset$, so let $u_1^i, u_1^j \in S_k$ be two vertices with $u_1^i \in P_i$ and $u_1^j \in P_j$. By definition of the functions w and c, it follows that:

$$c(u_1^i) = c(u_1^j) = \operatorname{diam}(S_k) \geq d(u_1^i, u_1^j) \geq w_{i,j}.$$

By minimality of C, we know that $w_{i,j} \leq w_{i,z_i} + w_{j,z_j}$ for any choice of $z_i, z_j \in \{1, \ldots, s\}$ with $\{p_i, p_{z_i}\}, \{p_j, p_{z_j}\} \in \tilde{E}$, so especially for z_h such that $w_{h,z_h} = \min\{w_{h,x} \colon 1 \leq x \leq s, \{p_h, p_x\} \in \tilde{E}\}$, $h \in \{i, j\}$. By Eq. (3) this means that for any two vertices $v_h \in P_h$, $h \in \{i, j\}$:

$$w_{i,j} \leq w_{i,z_i} + w_{j,z_j} \leq c(v_i) + c(v_j)$$

As $|P_h| \geq 2$, let $v_h \in P_h \backslash \{u_1^h\}$ for $h \in \{1, 2\}$, which gives:

$$c(P_i \cup P_j) := \sum_{v \in P_i \cup P_j} c(v) \geq c(u_1^i) + c(u_1^j) + c(v_i) + c(v_j) \geq 3 \cdot w_{i,j}.$$

Case 2: If $|N_C(i)| = 2$ let $N_C(i) = \{j, k\}$. Equation (4) yields $w_{h,z} \geq w_{h,i}$ for $h \in \{j, k\}$ and all $z \in \{1, \ldots, s\}$ with $\{p_h, p_z\} \in \tilde{E}$. Equation (3) hence yields $c(v) \geq w_{i,h}$ for all $v \in P_h$, $h \in \{j, k\}$. Let the edge $\{p_i, p_h\}$ be in $\tilde{E}$ because of u_h^i, u_h^1 for $h \in \{j, k\}$, i.e., $u_h^i \in P_i$ and $u_h^1 \in P_h$ and there exist $y_h \in \{1, \ldots, r\}$

such that $u_h^i, u_h^1 \in S_{y_h}$. By this definition, it follows that $c(u_h^i) \geq \text{diam}(S_{y_h}) \geq d(u_h^i, u_h^1) = w_{i,h}$. If $u_j^i \neq u_k^i$, it follows that

$$c(P_i \cup P_j \cup P_k) \geq c(u_j^i) + c(u_k^i) + |P_j| \cdot w_{ij} + |P_k| \cdot w_{i,k} \geq 3 \cdot (w_{i,j} + w_{i,k}).$$

If $u_j^i = u_k^i$, it follows that $y_j = y_k =: y$ and hence $u_j^1, u_k^1 \in S_y$, which means that the edge $\{p_j, p_k\}$ is in $\tilde{E}$ with weight at most $d(u_j^1, u_k^1) \leq \text{diam}(S_y) = c(u_j^i)$. Minimality of C yields that $w_{j,k} + \min\{w_{i,x} : 1 \leq x \leq s\} \geq w_{i,j} + w_{i,k}$, hence Eq. (3) yields that $w_{j,k} + c(|P_i \backslash \{u_j^i\}|) \geq w_{i,j} + w_{i,k}$, so:

$$c(P_i) \geq c(u_j^i) + c(|P_i \backslash \{u_j^i\}|) \geq w_{j,k} + (w_{i,j} + w_{i,k} - w_{j,k}) \geq w_{i,j} + w_{i,k},$$

which overall gives $c(P_i \cup P_j \cup P_k) \geq 3 \cdot (w_{i,j} + w_{i,k})$.

Case 3: If $|N_C(i)| \geq 3$, let $N_C(i) = \{i_1, \ldots, i_t\}$. Equations (4) and (3) yield:

$$c(v_{i_j}) \geq w_{i,i_j} \text{ for all } v_{i_j} \in P_{i_j},\ j \in \{1, \ldots, t\}. \tag{5}$$

Let for each $j \in \{1, \ldots, t\}$, $u_j \in P_i$ and $v_j \in P_{i_j}$ be the vertices defining the edge $\{p_i, p_{i_j}\}$, i.e., there exists $x_j \in \{1, \ldots, r\}$ such that $u_j, v_j \in S_{x_j}$. By this definition, it follows that:

$$c(u_j) = \text{diam}(S_{x_j}) \geq d(u_j, v_j) \geq w_{i,i_j} \text{ for all } j \in \{1, \ldots, t\} \tag{6}$$

If $u_j = u_{j'}$ for some $j \neq j'$, it follows that $x_j = x_{j'}$ and consequently the edge $\{p_{i_j}, p_{i_{j'}}\}$ is in $\bar{E}$ and has a cost of at most $d(v_j, v_{j'})$. Minimality of C implies that $d(v_j, v_{j'}) \geq w_{i,i_j} + w_{i,i_{j'}}$. On the other hand, we have $c(v) = \text{diam}(S_{x_j}) \geq d(v_j, v_{j'})$, for all $v \in S_{x_j}$, so especially for $v \in \{v_j, v_{j'}\}$. With Eq. (5), this gives:

$$\begin{aligned} c(P_{i_j} \cup P_{i_{j'}}) &\geq c(v_j) + c(v_{j'}) + c(P_{i_j} \backslash \{v_j\}) + c(P_{i_{j'}} \backslash \{v_{j'}\}) \\ &\geq 2(w_{i,i_j} + w_{i,i_{j'}}) + w_{i,i_j} + w_{i,i_{j'}} \\ &= 3(w_{i,i_j} + w_{i,i_{j'}}) \end{aligned} \tag{7}$$

Let M be a maximum matching for the graph $H = (\{1, \ldots, t\}, \hat{E})$ with $\hat{E} = \{\{j, j'\} : u_j = u_{j'}\}$. By the definition of the edges, maximality of M yields that for the unmatched indices $N := \{j : \{j, j'\} \notin M\ \forall\ 1 \leq j' \leq t\}$, we have $|\{u_j : j \in N\}| = |N|$. With Eqs. (4), (6) and (7) this yields:

$$
\begin{aligned}
c(P_i) + \sum_{j=1}^{t} c(P_{i_j}) &\geq \sum_{\{j,j'\}\in M} c(P_{i_j} \cup P_{i_{j'}}) + c(P_i) + \sum_{j\in N} c(P_{i_j}) \\
&\geq \sum_{\{j,j'\}\in M} 3(w_{i,i_j} + w_{i,i_{j'}}) + \sum_{j\in N} c(P_{i_j} \cup \{u_j\}) \\
&\geq \sum_{\{j,j'\}\in M} 3(w_{i,i_j} + w_{i,i_{j'}}) + \sum_{j\in N} w_{i,i_j} |P_{i_j} \cup \{u_j\}| \\
&\geq 3 \cdot \sum_{j=1}^{t} w_{i,i_j}.
\end{aligned}
$$

Let $C_1, \ldots, C_x$ be the connected components (stars) of the graph induced by the edges in C, and let p_{i_j} be the center of C_{i_j} for each $j \in \{1, \ldots, x\}$, then:

$$
D^* = \sum_{v\in V} c(v) = \sum_{j=1}^{r} c(P_j) = \sum_{t=1}^{x} c\left(\bigcup_{p_j\in C_t} P_j \right) \geq \sum_{t=1}^{x} 3 \cdot \sum_{j\in N_C(i_t)} w_{i_t,j} = 3 \cdot w(C).
$$

At last, since $\tilde{G}$ is a restriction of G', $w(C)$ is at least the cost of a minimum-weight edge cover for G' and by the proof of Theorem 2 any minimal edge cover for G' yields a solution for $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER . □

With the help of this Lemma, we can show that:

Theorem 9 *The problem* $(\|\cdot\|_1^w, \text{diam})$-4-CLUSTER *can be approximated in polynomial time within a factor of* $\frac{35}{6}$*, if d satisfies the triangle inequality.*

Proof Let $G = (V, E)$ be the input graph with induced distances d. First, compute an optimal solution $P_1, \ldots, P_s$ for $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER with Proposition 4. This solution satisfies $|P_i| \leq 3$ for all $i \in \{1, \ldots, s\}$. Let D^* be the global cost of $P_1, \ldots, P_s$. It follows that

$$
D^* \leq \text{opt}(G, d, \text{diam}, \|\cdot\|_1^w, 4), \tag{8}
$$

simply because any 4-cluster is also a 2-cluster.

Then, consider the complete graph $G' = (P, P \times P)$ with vertices $P = \{p_1, \ldots, p_s\}$ (p_i represents the set P_i) and edge-weights w defined by $w_{i,j} := w(\{p_i, p_j\}) := \min\{d(u, v) \colon u \in P_i, v \in P_j\}$. Compute an optimal solution $S_1, \ldots, S_r$ for $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER on G' with Theorem 2 such that with $|S_i| \leq 3$ for all $i \in \{1, \ldots, s\}$ by Proposition 2. Lemma1 then yields:

$$
D^* \geq 3 \cdot \text{opt}(G', w, \text{avg}, \|\cdot\|_1^w, 2) = 3 \cdot \sum_{j=1}^{s} |S_j| \cdot \text{avg}(S_j). \tag{9}
$$

We interpret this partition $S_1, \ldots, S_r$ as a partition $\mathfrak{S} = \{S'_1, \ldots, S'_r\}$ on the graph G, i.e., $S'_j := \bigcup_{p_i\in S_j} P_i$ for all $j \in \{1, \ldots, r\}$. As $|P_i|, |S_j| \geq 2$ for all $i \in \{1, \ldots, r\}$

and $j \in \{1, \ldots, s\}$ it follows that $|S'_j| \geq 4$ for all $j \in \{1, \ldots, s\}$, so $S'_1, \ldots, S'_r$ is a 4-cluster for G.

If $S_q = \{p_i, p_j, p_k\}$ with central vertex p_i for some $i, j, k \in \{1, \ldots, s\}$ with $|P_j| = 3$, we replace the cluster S'_q in $\mathfrak{S}$ by the two clusters $P' := P_j \cup \{u_i\}$ and $P'' := S'_q \backslash P'$ with $u_i \in P_i$ such that

$$w_{i,j} = \min\{d(u_i, v) \colon v \in P_j\}.$$

These new clusters satisfy:

$$|P'| \cdot \operatorname{diam}(P') \leq 4 \cdot (\operatorname{diam}(P_j) + w_{i,j}) < 2 \cdot |P_j| \cdot \operatorname{diam}(P_j) + 4 \cdot w_{i,j}$$

and

$$\begin{aligned} |P''| \cdot \operatorname{diam}(P'') &\leq 5 \cdot (\operatorname{diam}(P_i) + \operatorname{diam}(P_k) + w_{i,k}) \\ &\leq \tfrac{5}{2} \cdot |P_i| \cdot \operatorname{diam}(P_i) + \tfrac{5}{2} \cdot |P_k| \cdot \operatorname{diam}(P_k) + 5 \cdot w_{i,k}. \end{aligned}$$

Consider any set $R \in \mathfrak{S}$ which is not the result of splitting up a cluster S'_q.

- If $R = P_i \cup P_j$, we know that $\operatorname{diam}(R) \leq \operatorname{diam}(P_i) + \operatorname{diam}(P_j) + w_{i,j}$ and $|R| \leq 6$, hence:

$$|R| \cdot \operatorname{diam}(R) \leq 3 \cdot |P_i| \cdot \operatorname{diam}(P_i) + 3 \cdot |P_j| \cdot \operatorname{diam}(P_j) + 6 \cdot w_{i,j} \tag{10}$$

- If $R = P_i \cup P_j \cup P_k$, with p_i as central vertex for $S_q = \{p_i, p_j, p_k\}$; we know that $|R| \leq 7$ (as P_j and P_k have cardinality 2) and

$$\operatorname{diam}(R) \leq \operatorname{diam}(P_i) + \operatorname{diam}(P_j) + \operatorname{diam}(P_k) + w_{i,j} + w_{i,k},$$

hence $|R| \cdot \operatorname{diam}(R)$ is bounded by:

$$\begin{aligned} |R| \cdot \operatorname{diam}(R) &\leq 7 \cdot (\operatorname{diam}(P_i) + \operatorname{diam}(P_j) + \operatorname{diam}(P_k) + w_{i,j} + w_{i,k}) \\ &\leq \tfrac{7}{2} \sum_{h \in \{i,j,k\}} |P_h| \cdot \operatorname{diam}(P_h) \; + \; 7(w_{i,j} + w_{i,k}) \end{aligned} \tag{11}$$

Equations (9), (10) and (11) yield:

$$\begin{aligned} \sum_{R \in \mathfrak{S}} |R| \cdot \operatorname{diam}(R) &\leq \tfrac{7}{2} \cdot \sum_{i=1}^{r} |P_i| \cdot \operatorname{diam}(P_i) + 6 \cdot \sum_{R \subseteq P_i \cup P_j} w_{i,j} + 7 \cdot \sum_{R = P_i \cup P_j \cup P_k} (w_{i,j} + w_{i,k}) \\ &\leq \tfrac{7}{2} \parallel (\operatorname{diam}((P_1), \ldots, \operatorname{diam}(P_s)) \parallel_1^w + 7 \cdot \sum_{i=1}^{q} |S_i| \cdot \operatorname{avg}(S_i) \\ &\leq \tfrac{7}{2} D^* + \tfrac{7}{3} D^* = \tfrac{35}{6} D^*. \end{aligned}$$

□

Table 1 Summary of the complexity of all problem variants for $k = 2$

$k = 2$	rad	diam	avg
$\|\cdot\|_\infty$	in P (*Edge Cover*) (Proposition 3)	in P (*Simplex Cover*) (Proposition 5)	NP-complete (Theorem 7)
$\|\cdot\|_\infty^w$	NP-complete (Theorem 7)	in P (*Simplex Cover*) (Proposition 5)	NP-complete (Theorem 7)
$\|\cdot\|_1^w$	APX-hard (Theorem 6)	in P (*Simplex Matching*) (Proposition 4)	in P (*Weighted Edge Cover*) (Theorem 2)

Table 2 Summary of the approximation ratios for all problem variants

	rad	diam	avg
$\|\cdot\|_\infty$	2 (Theorem 8)	2 (Theorem 8)	?
$\|\cdot\|_\infty^w$	4 (Proposition 9)	4 (Proposition 9)	$4k - 2$ (Proposition 10)
$\|\cdot\|_\infty^w$	$16(k - 1)$ (Corollary 5)	$8(k - 1)$ (Proposition 8)	$2k$ (Proposition 7)

Remark 4 With Eq. 1, the above result yields a $\frac{35}{3}$-approximation for $(\|\cdot\|_1^w, \text{rad})$- 4-CLUSTER. Since the approximation-ratios from Theorem 9 are significantly better than the path-partitioning approximation from Proposition 8 (factor 24 and 48, respectively), it would be interesting to nest this construction further and extend it for larger values of k.

7 Conclusions

We have introduced and discussed the general problem $(\|\cdot\|, f)$-k-CLUSTER to model clustering tasks which do not fix the number of clusters but require each cluster to contain at least k objects. The nine chosen problem variants in this paper generalise many previous models but, of course, do not capture every possible way to measure the quality of the clustering. We however tried to cover many previous models while maintaining a clear framework in which similarities turned out to be quite fruitful.

Our NP-hardness result for $k = 3$ for all variants of $(\|\cdot\|, f)$-k-CLUSTER generalises all known complexity-results for these types of problems. Further, we completely characterise the complexity with respect to k; see Table 1.

The restriction to distances d which satisfy the triangle inequality turns the generally NP-hard problem $(\|\cdot\|_\infty^w, \text{avg})$-2-CLUSTER into a problem that can be solved in polynomial time. We further showed that this restriction is necessary for polynomial time approximations and derived a number of approximation strategies, mostly based on different other graph problems. Our approximation ratios are summarised in Table 2.

An interesting open question is whether $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER can be approximated within some constant ratio or at least within some ratio in $\mathcal{O}(k)$. The lack

of monotonicity for average distortion makes this measure the most challenging for approximation.

Acknowledgements Katrin Casel and Henning Fernau were supported by the German Science Foundation Deutsche Forschungsgemeinschaft (FE 560/6-1). Faisal Abu-Khzam and Cristina Bazgan were partially supported by the bilateral research cooperation CEDRE between France and Lebanon (Grant Number 30885TM). We are grateful for the helpful comments of the anonymous reviewers.

References

1. Abu-Khzam, F.N., Bazgan, C., Casel, K., Fernau, H.: Building clusters with lower-bounded sizes. In: Hong, S. (ed.) 27th International Symposium on Algorithms and Computation, ISAAC, LIPIcs, vol. 64, pp. 4:1–4:13. Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2016)
2. Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., Zhu, A.: Achieving anonymity via clustering. ACM Trans. Algorithms **6**(3), 49 (2010)
3. Anshelevich, E., Karagiozova, A.: Terminal backup, 3D matching, and covering cubic graphs. SIAM J. Comput. **40**(3), 678–708 (2011)
4. Armon, A.: On min–max r-gatherings. Theor. Comput. Sci. **412**(7), 573–582 (2011)
5. Blocki, J., Williams, R.: Resolving the complexity of some data privacy problems. In: Abramsky, S., Gavoille, C., Kirchner, C., auf der Heide, F.M., Spirakis, P.G. (eds.) Proceedings of the 37th International Colloquium Conference on Automata, Languages and Programming, ICALP'10: Part II, LNCS, vol. 6199, pp. 393–404. Springer (2010)
6. Byun, J.W., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. In: Kotagiri, R., Krishna, P.R., Mohania, M., Nantajeewarawat, E. (eds.) Advances in Databases: Concepts, Systems and Applications, LNCS, vol. 4443, pp. 188–200. Springer, Berlin (2007)
7. Cornuéjols, G., Hartvigsen, D., Pulleyblank, W.: Packing subgraphs in a graph. Oper. Res. Lett. **1**(4), 139–143 (1982)
8. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)
9. Domingo-Ferrer, J., Sebé, F.: Optimal multivariate 2-microaggregation for microdata protection: a 2-approximation. In: Domingo-Ferrer, J., Franconi, L. (eds.) Privacy in Statistical Databases, PSD'06, LNCS, vol. 4302, pp. 129–138. Springer, Berlin (2006)
10. Edmonds, J., Johnson, E.L.: Matching, Euler tours and the Chinese postman. Math. Program. **5**, 88124 (1973)
11. Ergün, F., Kumar, R., Rubinfeld, R.: Fast approximate PCPs. In: Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, 1–4 May 1999, Atlanta, Georgia, USA, pp. 41–50 (1999)
12. Goemans, M., Williamson, D.: A general approximation technique for constrained forest problems. SIAM J. Comput. **24**(2), 296–317 (1995)
13. Guha, S., Meyerson, A., Munagala, K.: Hierarchical placement and network design problems. In: In Proceedings of the 41th Annual IEEE Symposium on Foundations of Computer Science, FOCS'00, pp. 603–612. IEEE Computer Society (2000)
14. King, V., Rao, S., Tarjan, R.: A faster deterministic maximum flow algorithm. J. Algorithms **17**(3), 447–474 (1994)
15. Orlin, J.B.: Max flows in $O(nm)$ time, or better. In: Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC, pp. 765–774. ACM (2013)
16. Papadimitriou, C.H., Yannakakis, M.: Optimization, approximation, and complexity classes. J. Comput. Syst. Sci. **43**, 425–440 (1991)
17. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
18. Schrijver, A.: Combinatorial Optimization. Springer, Berlin (2003)
19. Shalita, A., Zwick, U.: Efficient algorithms for the 2-gathering problem. ACM Trans. Algorithms **6**(2), 34 (2010)
20. Stokes, K.: On computational anonymity. In: Privacy in Statistical Databases—UNESCO Chair in Data Privacy, International Conference, PSD 2012, Palermo, Italy, 26–28 September 2012. Proceedings, pp. 336–347 (2012)

21. Tovey, C.: A simplified NP-complete satisfiability problem. Discrete Appl. Math. **8**(1), 85–89 (1984)
22. Xu, D., Anshelevich, E., Chiang, M.: On survivable access network design: complexity and algorithms. In: INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13–18 April 2008, Phoenix, AZ, USA, pp. 186–190 (2008)