Building Clusters with Lower-Bounded Sizes

Faisal Abu-Khzam[1], Cristina Bazgan[*2], Katrin Casel[†3], and Henning Fernau[‡4]

1 **Lebanese American University, Beirut, Lebanon**
faisal.abukhzam@lau.edu.lb
2 **Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, Paris, France**
bazgan@lamsade.dauphine.fr
3 **Fachbereich 4, Informatikwissenschaften, Universität Trier, Germany**
casel@uni-trier.de
4 **Fachbereich 4, Informatikwissenschaften, Universität Trier, Germany**
fernau@uni-trier.de

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.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity, G.1.2 Approximation, I.5.3 Clustering

Keywords and phrases Clustering, Approximation Algorithms, Complexity, Matching

Digital Object Identifier 10.4230/LIPIcs.ISAAC.2016.4

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 known which number for k would be a reasonable choice. 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

* Institut Universitaire de France.
† K. Casel gratefully acknowledges the support by the Deutsche Forschungsgemeinschaft, grant FE 560/6-1.
‡ H. Fernau gratefully acknowledges the support by the Deutsche Forschungsgemeinschaft, grant FE 560/6-1.


© Faisal Abu-Khzam, Cristina Bazgan, Katrin Casel, and Henning Fernau;
licensed under Creative Commons License CC-BY

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 [11], a variant of the facility location problem where one is only interested in building profitable facilities. 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 it is possible to build an unrestricted number of facilities. The considered cardinality-constraint also models the basic principle of "hiding in a crowd" introduced by the concept of k-*anonymity* [14] which introduces formal problems such as r-Gather [1] and k-Member Clustering [4]. A cluster in this scenario 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 each f and $\|\cdot\|$; a detailed description of these variants follows in Section 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 similarities to other graph problems. Section 3 compares the nine problem variants with respect to structural differences. In Section 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 Section 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 the conclusions.

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 n input-objects as vertices of an undirected graph $G = (V, E)$. A feasible solution is any partitioning $P_1, \dots, P_s$ of V such that $|P_i| \geq k$ for all $i \in \{1, \dots, s\}$, 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 partitioning 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 $d\colon V^2 \to \mathbb{R}_+$ 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 $w_E\colon E \to \mathbb{R}_+$. 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

$w_E(\{v,w\}) = 3$. Distances which are defined directly via an edge are the only possible 'non-metric' distances. 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 [17], where hyperedges reflect similar groups.

The overall cost of a partitioning $P_1, \ldots, P_s$ is always in some sense proportional to the dissimilarities within each set or *cluster* P_i. On an abstract level, the *global cost* induced by a partitioning $P_1, \ldots, P_s$ 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_i:

Radius: $\mathrm{rad}(P_i) := \min_{x \in P_i} \max_{y \in P_i} d(x,y)$.

Diameter: $\mathrm{diam}(P_i) := \max_{x \in P_i} \max_{y \in P_i} d(x,y)$.

Average Distortion: $\mathrm{avg}(P_i) := \frac{1}{|P_i|} \cdot \min_{x \in P_i} \sum_{y \in P_i} d(x,y)$.

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

Worst Local Cost: The maximum cost of an individual cluster: $\max_{1 \le i \le s} f(P_i)$. Because of its structure with respect to the values $f(P_1), \ldots, f(P_s)$, denoted by $\|\cdot\|_\infty$.

Worst Weighted Local Cost: The maximum cost of an individual cluster, weighted by its size: $\max_{1 \le i \le s} |P_i| f(P_i)$, denoted by $\|\cdot\|_\infty^w$.

Accumulated Local Cost: The sum of the distortion for each cluster, denoted by $\|\cdot\|_1^w$, with respect to the cost of the individual clusters computed by: $\sum_{i=1}^{s} |P_i| f(P_i)$.

Any combination of $f \in \{\mathrm{rad}, \mathrm{diam}, \mathrm{avg}\}$ with $\|\cdot\| \in \{\|\cdot\|_1^w, \|\cdot\|_\infty^w, \|\cdot\|_\infty\}$ yields a different problem. (Structural properties discussed in Section 3 will explain why we do not consider the 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 $w_E : E \to \mathbb{R}_+$, $k \in \mathbb{N}$.
Output: k-cluster $P_1, \ldots, P_s$ of V for some $s \in \mathbb{N}$, which minimises $\|(f(P_1), \ldots, f(P_s))\|$.

$(\|\cdot\|_\infty, \mathrm{rad})$-$k$-CLUSTER, for example, searches for a k-CLUSTER which minimises:

$$\max_{1 \le i \le s} \min_{x \in P_i} \max_{y \in P_i} d(x,y).$$

Some of the variants of $(\|\cdot\|, f)$-k-CLUSTER are already known under different names. The variant $(\|\cdot\|_1^w, \mathrm{diam})$-$k$-CLUSTER is also known as k-MEMBER CLUSTERING [4] and with d chosen as the Euclidean distance, $(\|\cdot\|_\infty, \mathrm{rad})$-$k$-CLUSTER is the so-called r-GATHER problem [1] (with $r = k$). Variant $(\|\cdot\|_1^w, \mathrm{avg})$-$k$-CLUSTER is LOAD BALANCED FACILITY LOCATION [11] with unit demands and without facility costs and, with Euclidean distance, also models MICROAGGREGATION [6].

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 $V := \{c, v_1, v_2, \dots, v_n\}$ 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.** *For any $k \in \mathbb{N}$ and any $\|\cdot\| \in \{\|\cdot\|_1^w, \|\cdot\|_\infty^w, \|\cdot\|_\infty\}$, optimal solutions $P_1, \dots, P_s$ for $(\|\cdot\|, diam)$-k-CLUSTER can be assumed to satisfy $|P_i| < 2k$ for all $1 \leq i \leq s$.*

For radius we only have the weaker property that $\operatorname{rad}(S) \leq \operatorname{rad}(P)$ for all sets $S \subseteq P$ such that the center of P is contained in S. Average distortion lacks such monotone behaviour entirely. Observe that a large cardinality of a cluster can sort of "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 this show that, even with triangle inequality for d, we can not in general restrict the maximum cluster-cardinality for $(\|\cdot\|_\infty, \text{avg})$-$k$-CLUSTER, which is a bit unsetteling, given that most applications also like 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, formally computed by $\|(f(P_1), \dots, f(P_s))\|_1 := \sum_{i=1}^s f(P_i)$. 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 $i \in \{1, \dots, n\}$ and $w_E(\{v_i^h, v_j^k\}) = n - 1$ for $i, j \in \{1, \dots, n\}$ with $i \neq j$ and $h, k \in \{1, 2\}$, the best 2-clustering w.r.t. $\|\cdot\|_1$ with any choice for f is V itself, while the most reasonable 2-clustering 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 makes a big difference for the worst-case example in the beginning of the section and allows to conclude:

▶ **Theorem 2.** *If d satisfies the triangle inequality, the restriction to partitions into sets of cardinality at most $2k - 1$ yields a 2-approximation for $(\|\cdot\|_\infty, rad)$-, $(\|\cdot\|_1^w, rad)$- and $(\|\cdot\|_1^w, avg)$-k-CLUSTER and is optimal for $(\|\cdot\|_\infty^w, avg)$- and $(\|\cdot\|_\infty^w, rad)$-$k$-CLUSTER.*

As we will look at the cases $k = 2$ and $k = 3$ in the next section, we further conclude:

▶ **Corollary 3.** *If d satisfies the triangle inequality, sets in partitions for $(\|\cdot\|_1, avg)$-2-CLUSTER can be assumed to have cardinality two or three.*

Proof. For a cluster $S := \{x_1, x_2 \dots, x_r\}$ with center x_1 and $r > 3$, a further partitioning into $\{x_{2i}, x_{2i+1}\}$ for $i \in \{1, \dots, z-1\}$ with $z = \lfloor \frac{r}{2} \rfloor$ and $\{x_1, x_{2z}, x_r\}$ does not increase the

global cost for $(\|\cdot\|_1, \text{avg})$-2-CLUSTER, since:

$$
\begin{aligned}
&|S|\text{avg}(S) \\
=&\sum_{i=1}^{r} d(x_i, x_r) \leq (r-2z)d(x_{2z}, x_r) + d(x_r, x_{2z-1}) + \sum_{i=1}^{z-1} d(x_{2i}, x_r) + d(x_{2i+1}, x_r) \\
\leq&|\{x_1, x_{2z}, x_r\}|\text{avg}(\{x_1, x_{2z}, x_r\}) + \sum_{i=1}^{z-1} 2\text{avg}(\{x_{2i}, x_{2i+1}\})\,.
\end{aligned}
$$

◀

4 Connections to Matching Problems

The graph-representation we chose to define $(\|\cdot\|, f)$-k-CLUSTER reveals relations to other well studied graph problems, interestingly 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, a problem which can be reduced to finding a minimum weight perfect matching (a simple reduction is described, e.g., in the first volume of Schrijver's monograph [[15], 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 [8].

▶ **Theorem 4.** $(\|\cdot\|_1^w, avg)$-2-CLUSTER *can be solved in* $\mathcal{O}(n^3)$ *time.*

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

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

In other words, for any graph $G = (V, E)$, the global cost is the weight of the cheapest edge-set $E' \subset V^2$ 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 three for every minimum weight edge cover E' which means that its connected components are star-graphs. ◀

▶ **Theorem 5.** $(\|\cdot\|_\infty, rad)$-2-CLUSTER *can be solved in* $\mathcal{O}(n^2)$ *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 since any minimal edge cover for the graph $G' := (V, E')$ with $E' := \{(u, v) \colon 0 < d(u, v) \leq c\}$ yields a 2-cluster for G with radius at most c for each cluster. ◀

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 three are required in every optimal solution. It seems difficult to define a correct way 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 $\mathcal{O}(n^3 m^2 \log n)$ [2]:

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

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

Output: A perfect matching of H (that is a collection S of hyperedges such that every vertex in V appears in exactly one hyperedge of S) of minimal cost.

▶ **Corollary 6.** $(\|\cdot\|_1^w, \mathit{diam})$-2-CLUSTER *can be solved in* $\mathcal{O}(n^9 \log n)$ *time.*

Proof. Let $G = (V, E)$ be an input graph for $(\|\cdot\|_1^w, \mathrm{diam})$-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 $(\|\cdot\|_1^w, \mathrm{diam})$-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 $c(\{u,v,w\}) := 3\max\{d(u,v), d(v,w), d(u,w)\}$ 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 can be solved using Corollary 6 by fixing some maximum diameter D and multiplying all hyperedge-costs which exceed D with a large value C, say $C = n\max\{d(u,v)\colon u,v \in V\}$. This does not violate the simplex condition for the cost-function and there exists a solution for $(\|\cdot\|_\infty, \mathrm{diam})$-2-CLUSTER of value D for the original graph if and only if the hypergraph with adjusted costs has a $(\|\cdot\|_1^w, \mathrm{diam})$-2-CLUSTER solution of value less than C. Relating to an easier problem, we can do a little better. If we remove the hyperedges which exceed D instead of changing their cost, we arrive at a hypergraph which still satisfies the subset condition ($\mathrm{diam}(\{u,v\}) \leq \mathrm{diam}(\{u,v,w\})$ for any $u, v, w \in V$) and we are only interested in any perfect matching, regardless of its weight. The computation of such a perfect matching is the problem called SIMPLEX COVER [19][1]. The augmenting-path strategy from [16] for 2-GATHERING[2], can be used to solve SIMPLEX COVER in time $\mathcal{O}(m^2)$, where m is the number of hyperedges of the input graph.

▶ **Corollary 7.** $(\|\cdot\|_\infty, \mathit{diam})$- *and* $(\|\cdot\|_\infty^w, \mathit{diam})$-2-CLUSTER *and if d satisfies the triangle inequality also* $(\|\cdot\|_\infty^w, \mathit{avg})$-2-CLUSTER *can be solved in* $\mathcal{O}(n^6 \log n)$ *time.*

▶ Remark. We would like to point out that SIMPLEX MATCHING is also an interesting way to solve a sort of geometric version of $(\|\cdot\|_1^w, \mathrm{avg})$-2-CLUSTER, originally introduced as MICROAGGREGATION in [6], 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 $\frac{1}{t}(x_1 + \cdots + x_t)$. 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 $2c(\{v_i, v_j, v_k\}) = \frac{4}{3}(c(\{v_i, v_j\}) + c(\{v_j, v_k\}) + c(\{v_i, v_k\}))$. This construction gives a polynomial-time algorithm to solve 2-MICROAGGREGATION which improves on the 2-approximation from [7].

As powerful as SIMPLEX MATCHING may seem, the estimated worst-case running-time is fairly large. We believe that an augmenting path strategy which is specifically tailored to the above problems can yield significant improvement. Observe that similar construction for $(\|\cdot\|_1^w, \mathrm{rad})$-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

[1] This covering problem is equivalent to $\{K_2, K_3\}$-PACKING an old, well studied generalisation of the classical matching problem [5].
[2] Confusingly, 2-GATHERING in [16] is not equivalent to the r-GATHERING problem from [1] with $r = 2$.

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.

5 Complexity Results

In [1], the problem r-GATHER, which is $(\|\cdot\|_\infty, \text{rad})$-$k$-CLUSTER with $r = k$ with Euclidean distance, was shown to be NP-complete for $k \geq 7$. In [3] 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_i calculated by: $\min_{x \in V} \max_{y \in P_i} d(x,y)$) which is not allowed in our formal definition. We establish in the following a different reduction from the EXACT-t-COVER problem which shows NP-hardness for all our variants of k-cluster and extends for all measures f which are strictly monotone with respect to radius, diameter or average distortion. With EXACT-t-COVER we refer to the problem of deciding for a given collection $C = \{S_1, \ldots, S_r\}$ of subsets of a universe $X = \{x_1, \ldots, x_n\}$ with $|S_i| = t$ for all i, if there exists $C' \subset C$ such that $|C'| = n/t$ and $\bigcup_{S \in C'} S = X$, which is NP-hard for all $t \geq 3$ [9].

▶ **Theorem 8.** *All variants of $(\|\cdot\|, f)$-k-CLUSTER are NP-hard for $k \geq 3$ even with the restriction to distances d which satisfy the triangle inequality.*

Proof (Sketch). 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 $|S_i| = t$. The graph G for $(\|\cdot\|, f)$-k-CLUSTER only contains edges of weight one and vertices $u_1, \ldots, u_n$ representing $x_1, \ldots, x_n$ and, for all $i \in \{1, \ldots, r\}$, we have vertices $w_i^1, \ldots, w_i^{k-1}$ representing an arbitrary fixed partition $P_1^i, \ldots, P_{k-1}^i$ of S_i with $|P_{i_j}| = k-1$ for all j, and some additional vertices v_j for sets which are not in the cover. Edges connect u_j to with w_{i_z} if $u_j \in P_z^i$. Other edges are included depending on f. We want a solution $C \subset \{S_1, \ldots, S_r\}$ with $|C| = n/t$ for EXACT-t-COVER to translate to the k-sets of vertices $\{w_i^z, u_j \colon x_j \in P_z^i\}$ for all i with $S_i \in C$. Assigning v_j to the set $\{w_i^1, \ldots, w_i^{k-1}\}$ for i with $S_i \notin C$ then partitions the remaining vertices. There is a k-clustering which only uses these types of clusters for w_i^z if and only if $S_1, \ldots, S_r$ is an exact cover.

For $f = \text{diam}$, we use $\ell := r - \frac{n}{t}$ vertices $v_1, \ldots, v_\ell$ and turn each of the sets $\{u_1, \ldots, u_n\}$ and $w_i^1, \ldots, w_i^{k-1}$ for $i \in \{1, \ldots, r\}$ into a clique, and connect each v_h with $h \in \{1, \ldots, \ell\}$ to all w_i^z ($i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k\}$). With this, there exists an exact cover for $S_1, \ldots, S_r$ if and only if a there exists a k-cluster of maximum diameter one.

For $f \in \{\text{rad}, \text{avg}\}$, we use r vertices $v_1, \ldots, v_r$ and edges (v_i, w_i^z) for $i \in \{1, \ldots, r\}$ and $z \in \{1, \ldots, k-1\}$ and further include vertices y_i^j for $i \in \{1, \ldots, \frac{n}{t}\}$ and $j \in \{1, \ldots, k-1\}$ with edges (y_1^i, y_h^i) and (y_1^i, v_j) for each $i \in \{1, \ldots, \frac{n}{t}\}$, $h \in \{2, \ldots, k-1\}$ and $j \in \{1, \ldots, r\}$. With this construction there exists an exact cover for $S_1, \ldots, S_r$ if and only if there is a clustering such that all clusters have cardinality k and radius one.

In particular, there exists an exact cover for $S_1, \ldots, S_r$ if and only if there exists a k-cluster with global cost 1, k and $2n + (k-1)r + \frac{n}{k-1}$ for radius with norm $\|\cdot\|_\infty, \|\cdot\|_\infty^w$ and $\|\cdot\|_1^w$, respectively and $\frac{k-1}{k}$, $k-1$ and $2n + \frac{1}{k}(tr - n)$ for average distortion with norm $\|\cdot\|_\infty, \|\cdot\|_\infty^w$ and $\|\cdot\|_1^w$, respectively. ◀

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 $k = 2$.

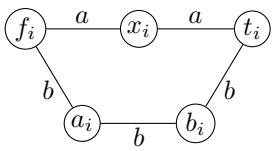
▶ **Theorem 9.** *$(\|\cdot\|_1^w, rad)$-2-CLUSTER is APX-hard, even with the restriction to distances d which satisfy the triangle inequality.*

Proof (Sketch). We reduce from Vertex Cover restricted to cubic graphs which is APX-hard by [13]. Let $G = (V, E)$ with $V = \{v_1, \dots, v_n\}$ be the input for Vertex Cover, we define $G' = (V', E')$ by $V' := \{v_i^1, v_i^2 \colon 1 \leq i \leq n\} \cup \{v_e \colon e \in E\}$ and $E' = \{\{v_i^1, v_i^2\} \colon 1 \leq i \leq n\} \cup \{\{v_i^1, v_e\} \colon v_i \in e\}$ with weights $w_E(\{v_i^1, v_i^2\}) = 1$ and $w_E(\{v_i^1, v_e\}) = 2$. With these definitions, G has a vertex cover of cardinality k if and only if there exists a solution for $(\|\cdot\|_1^w, \text{rad})$-2-cluster with global cost $2n + 2k + 2m$. Since $m = 3n/2$ and $k \geq n/2$ for a cubic graph, this reduction preserves non-approximability. ◀

The reduction above can not be altered for the cases of $(\|\cdot\|, f)$-2-cluster with some ∞-norm which were not shown to be polynomial-time solvable so far. We therefore consider a completely different problem for these cases to show:

▶ **Theorem 10.** *$(\|\cdot\|_\infty^w, avg)$-, $(\|\cdot\|_\infty, avg)$- and $(\|\cdot\|_\infty^w, rad)$-2-cluster are all NP-hard, for the latter two even with the restriction to distances d which satisfy the triangle inequality.*

Proof (Sketch). Reduction from $(3, 3)$-SAT, i.e., satisfiability with at most three variables in each clause and where each variable occurs (positively or negatively) in at most three clauses, which remains NP-hard by [18]. Let $v_1, \dots, v_n$ be the variables and $c_1, \dots, c_m$ be the clauses. We construct G by introducing for each v_i the subgraph displayed on the below.



For each clause c_j we introduce a vertex y_j connected with edges of weight b to t_i if v_i is a literal in c_j and to f_i if $\bar{v}_i$ is a literal in c_j. With $a = \frac{1}{2}$, $b = \frac{1}{3}$ for $(\|\cdot\|_\infty, \text{rad})$-, $a = 2, b = \frac{3}{2}$ for $(\|\cdot\|_\infty, \text{avg})$- and $a = 1$, $b = \frac{1}{2}$ and also additional edges $\{y_i, y_j\}$ for all $i \neq j$ of weight one for $(\|\cdot\|_\infty^w, \text{avg})$-2-cluster, the clause is satisfiable if and only if the clustering-problem has a solution of global cost one. ◀

6 Approximation results

We will only consider the case where d satisfies the triangle inequality in this section. This restriction is not just reasonable but in some sense necessary to achieve any kind of approximation. If we reconsider the reduction from Theorem 8 and turn the constructed graph G into a complete graph with additional edges of a large weight w, the difference in global cost in case of "yes"- or "no"-instance of Exact-t-Cover increases with w, which implies:

▶ **Proposition 11.** *If d violates the triangle inequality, there is no constant-factor approximation for $(\|\cdot\|, f)$-k-cluster in time polynomial in $|V|$, unless $P = NP$.*

A closer look at the metric given by the shortest paths for the original construction from Theorem 8, reveals that the global cost differs by a factor of two between "yes"- and "no"-instance for some problem-variants. Explicitly this means:

▶ **Proposition 12.** *There is no $(2 - \varepsilon)$-approximation in polynomial time for $(\|\cdot\|, f)$-k-cluster with $f \in \{rad, diam\}$ and $\|\cdot\| \in \{\|\cdot\|_\infty, \|\cdot\|_\infty^w\}$ for any $\varepsilon > 0$ unless $P = NP$, even if d satisfies triangle inequality.*

Known approximation results for clustering with size constraints include a 9-approximation from [3] 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; [11] 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 2.

Other known approximation results however also apply here and can even be altered to yield results for other problem-variants. The problem $(\|\cdot\|_\infty, \text{rad})$-$k$-cluster is discussed under the name r-gather in [1], where r takes the role of k. The concept for the 2-approximation presented there can be altered, even simplified, and also used to compute a 2-approximation for $(\|\cdot\|_\infty, \text{diam})$-$k$-cluster.

▶ **Theorem 13.** *$(\|\cdot\|_\infty, rad)$- and $(\|\cdot\|_\infty, diam)$-k-cluster are 2-approximable for all $k \geq 2$.*

Proof (Sketch). We try all values D that occur as pairwise distances $d(u,v)$ for $u, v \in V$ for the following greedy strategy: Start with $V_1 := V$ and iteratively, until $V_i = \emptyset$, choose $c_i \in V_i$, build clusters $P(c_i) := \{v \in V_i \colon d(c_i, v) \leq D\}$ and set $V_{i+1} = V_i \setminus P(c_i)$. This yields a partition of V into a finite number of clusters $P(c_i)$. If some cluster $P(c_i)$ has less than k vertices, consider $S(i,j) = \{v \in P(c_j) \setminus \{c_j\} \colon d(v, c_i) \leq D\}$ and move $\min\{|S(i,j)|, |P(c_j)| - k\}$ vertices from $S(i,j)$ to $P(c_i)$ for each $j \in \{1, \dots, i-1\}$ until $|P(c_i)| \geq k$. If this procedure is successful, we arrive at a k-cluster for V with maximum radius D and maximum diameter $2D$. This procedure is successful for $D = 2r^*$ and $D = D^*$ if r^* and D^* are optimal values for $(\|\cdot\|_\infty, \text{rad})$- and $(\|\cdot\|_\infty, \text{diam})$-$k$-cluster respectively. ◀

▶ Remark. A greedy procedure for $(\|\cdot\|_\infty, \text{avg})$-$k$-cluster could build up the sets $P(c_i)$ by successively adding $\operatorname{argmin}\{d(v, c_i) \colon v \in V_i \setminus P(c_i)\}$ 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 $P(c_j)$.

In [12] results from [10] for the so-called Proper Constraint Forest Problem are used to compute an $8(k-1)$-approximation for Microaggregation. We will use a different result from [10]: a 2-approximation for Lower Capacitated Tree Partitioning with capacity k which is the problem of computing a spanning forest of minimal cost for which each connected component has cardinality at least k. A spanning forest is characterised by a set of edges and its cost is defined as the sum of the weights of these edges.

▶ **Corollary 14.** *$(\|\cdot\|_1^w, avg)$-k-cluster is $2k$-approximable for all $k \geq 2$.*

▶ Remark. For $k = 2$, Theorem 4 showed that $(\|\cdot\|_1^w, \text{avg})$-$k$-cluster can be solved in polynomial time which also translates to Lower Capacitated Tree Partitioning with capacity $k = 2$; tree partitioning with capacity two is equivalent to weighted edge-cover.

Essential for the result above is the fact that components of a minimal spanning forest do not contain paths of length $2k$ or more. This property implies the existence of a central vertex which can reach all vertices in its component in at most k steps and allows to bound the average distortion. This property does not prevent a component from containing arbitrarily many vertices. An algorithm for $(\|\cdot\|_1^w, \text{diam})$- or $(\|\cdot\|_1^w, \text{rad})$-$k$-cluster requires such an

upper bound on the cardinality to prove an approximation factor. We therefore consider Lower Capacitated Path Partitioning, the restriction of Lower Capacitated Tree Partitioning to paths as connected components. With triangle inequality, [10] provides a 4-approximation for this problem and it is clear that minimal solutions can be assumed to have connected components with at most $2k-1$ vertices each, which yields:

▶ **Corollary 15.** $(\|\cdot\|_1^w, diam)$-k-cluster *is* $(8k-7)$-*approximable for all* $k \geq 2$.

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:

$$\text{avg}(P_i) \leq \text{rad}(P_i) \leq \text{diam}(P_i) \leq 2\text{rad}(P_i)\,. \tag{1}$$

This relation with Corollary 15 immediately yields:

▶ **Proposition 16.** $(\|\cdot\|_1^w, rad)$-k-cluster *is* $(16k-14)$-*approximable for all* $k \geq 2$.

By definition, the two ∞-norms also relate optimal values in the following way for every choice of $f \in \{\text{rad,diam,avg}\}$, where we denote by $\text{opt}(G, d, \|\cdot\|, f, k)$ the global cost of an optimal solution for $(\|\cdot\|, f)$-k-cluster on G with distance d:

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

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

▶ **Corollary 17.** $(\|\cdot\|_\infty^w, diam)$-$k$-cluster *is* 4-*approximable and* $(\|\cdot\|_\infty^w, rad)$-$k$-cluster *is* 8-*approximable for all* $k \geq 2$.

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:

▶ **Corollary 18.** $(\|\cdot\|_\infty^w, avg)$-$k$-cluster *is* $(4k-2)$-*approximable for all* $k \geq 2$.

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

$$|P|\text{avg}(P) = \min_{c\in P} \sum_{p\in P} d(c,p) \geq \min_{c\in P} \max_{u,v\in P} d(u,c) + d(v,c) \geq \max_{u,v\in P} d(u,v) = \text{diam}(P)\,.$$

Theorem 13 and Proposition 1 produce a 2-approximation for $(\|\cdot\|_\infty, \text{diam})$-$k$-cluster for which each set contains at most $2k-1$ vertices. The weighted ∞-norm of the average distortion of this partition 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 in an even more surprising way. The solutions for $k=2$ derived in Section 4 for two different problem-variants are combined 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 $(\|\cdot\|_1, \text{avg})$-2-cluster.

▶ **Theorem 19.** *The problem* $(\|\cdot\|_1^w, diam)$-4-cluster *can be approximated in polynomial time within a factor of* $\frac{35}{6}$.

Proof (Sketch). Consider as input any graph $G = (V, E)$ with induced distances d. First, compute an optimal solution $P_1, \dots, P_s$ for $(\|\cdot\|_1^w, \text{diam})$-2-CLUSTER, for which the optimal value $\|(\text{diam}((P_1), \dots, \text{diam}(P_s))\|_1^w$ is at most $D^* := \text{opt}(G, d, \text{diam}, \|\cdot\|_1^w, 4)$, simply because any 4-cluster is also a 2-cluster. Next, consider the complete graph $G' = (P, P^2)$ with vertices $P := \{p_1, \dots, p_s\}$ and edge-weights w defined by $w(p_i, p_j) := \min\{d(u, v) \colon u \in P_i, v \in P_j\}$. It can be shown that $D^* \geq 3 \cdot \text{opt}(G', w, \text{avg}, \|\cdot\|_1^w, 2)$ and use an optimal solution $S_1, \dots, S_q$ for $(\|\cdot\|_1^w, \text{avg})$-2-CLUSTER on G', such that $|S_i| \leq 3$ for all i by Corollary 3. The partition $S = \{\bigcup_{p_i \in S_j} P_i \colon 1 \leq j \leq q\}$ is a 4-cluster for G. If $S_q = \{p_i, p_j, p_k\}$ with center p_i for some $i, j, k \in \{1, \dots, s\}$ with $|P_j| = 3$, we replace the cluster $P = P_i \cup P_j \cup P_k$ in S by the two clusters $P' := P_j \cup \{u_i\}$ and $P'' := P \setminus P'$, where we choose $u_i \in P_i$ such that $w(p_i, p_j) = \min\{d(u_i, v) \colon v \in P_j\}$. These new clusters satisfy:

$$|P'|\text{diam}(P') \leq 4(\text{diam}(P_j) + w(p_i, p_j)) < 2|P_j|\text{diam}(P_j) + 4w(p_i, p_j) \quad \text{and}$$

$$|P''|\text{diam}(P'') \leq \tfrac{5}{2}|P_i|\text{diam}(P_i) + \tfrac{5}{2}|P_k|\text{diam}(P_k) + 5w(p_i, p_k)$$

Consider any set $R \in S$ which is not the result of splitting up a cluster. Worst case is $R = P_i \cup P_j \cup P_k$ with p_i as center of $S_q = \{p_i, p_j, p_k\}$, we know that $|R| \leq 7$ and $\text{diam}(R) \leq \text{diam}(P_i) + \text{diam}(P_j) + \text{diam}(P_k) + w(p_i, p_j) + w(p_i, p_k)$, hence:

$$|R|\text{diam}(R) \leq \tfrac{7}{2}(\|P_i|\text{diam}(P_i) + |P_j|\text{diam}(P_j) + |P_k|\text{diam}(P_k)) + 7(w(p_i, p_j) + w(p_i, p_k))\,.$$

Overall, this yields:

$$\sum_{R \in S} |R|\text{diam}(R) \leq \tfrac{7}{2} \sum_{i=1}^{r} |P_i|\text{diam}(P_i) + 6 \sum_{R \subset P_i \cup P_j} w(p_i, p_j) + 7 \sum_{R = P_i \cup P_j \cup P_k} w(p_i, p_j) + w(p_i, p_k)$$

$$\leq \tfrac{7}{2}\,\|(\text{diam}((P_1), \dots, \text{diam}(P_s))\|_1^w + 7 \sum_{i=1}^{q} |S_i|\text{avg}(S_i) \leq \tfrac{7}{2}D^* + \tfrac{7}{3}D^* = \tfrac{35}{6}D^*\,. \qquad \blacktriangleleft$$

▶ Remark. Equation 1 translates the above result to a $\frac{35}{3}$-approximation for $(\|\cdot\|_1^w, \text{rad})$-4-CLUSTER. Since the approximation-ratios from Theorem 19 are significantly better than the path-partitioning approximation from Corollary 15 (factor 25 and 50 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 in order 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 with the following results for $(\|\cdot\|, f)$-2-CLUSTER:

$k = 2$	rad	diam	avg
$\|\cdot\|_\infty$	in P (EDGE COVER) Th.5	in P (SIMPLEX COVER) Cor.7	NP-complete Th.10
$\|\cdot\|_\infty^w$	NP-complete Th.10	in P (SIMPLEX COVER) Cor.7	NP-complete Th.10
$\|\cdot\|_1^w$	APX-hard Th. 9	in P (SIMPLEX MATCHING) Cor.6	in P (WEIGHTED EDGE COVER) Th.4

The restriction to distances d which satisfy the triangle inequality already simplified exact solvability for the general NP-hard problem $(\|\cdot\|_\infty^w, \text{avg})$-2-CLUSTER which turned out to be solvable with SIMPLEX COVER in this case. We further showed that this restriction is necessary for approximations in time polynomial in the number of objects and derived a number of approximation strategies, mostly based on different other graph-problems. Our approximation-ratios (which are the best and/or only ones known) are:

	rad	diam	avg
$\|\cdot\|_\infty$	2 Th.13	2 Th.13	?
$\|\cdot\|_\infty^w$	8 Cor.17	4 Cor.17	$4k-2$ Cor.18
$\|\cdot\|_1^w$	$16k-14$ Prop.16	$8k-7$ Cor.15	$2k$ Cor. 14

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.

References

1 G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and An Zhu. Achieving anonymity via clustering. *ACM Transactions on Algorithms*, 6(3), 2010.

2 E. Anshelevich and A. Karagiozova. Terminal Backup, 3D Matching, and Covering Cubic Graphs. *SIAM J. Comput.*, 40(3):678–708, 2011.

3 A. Armon. On min-max r-gatherings. *Theoretical Computer Science*, 412(7):573–582, 2011.

4 J.-W. Byun, A. Kamra, E. Bertino, and N. Li. Efficient k-anonymization using clustering techniques. In R. Kotagiri, P. R. Krishna, M. Mohania, and E. Nantajeewarawat, editors, *Advances in Databases: Concepts, Systems and Applications*, volume 4443 of *LNCS*, pages 188–200. Springer, 2007.

5 G. Cornuéjols, D. Hartvigsen, and W. Pulleyblank. Packing subgraphs in a graph. *Operations Research Letters*, 1(4):139–143, 1982.

6 J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

7 J. Domingo-Ferrer and F. Sebé. Optimal Multivariate 2-Microaggregation for Microdata Protection: A 2-Approximation. In J. Domingo-Ferrer and L. Franconi, editors, *Privacy in Statistical Databases, PSD'06*, volume 4302 of *LNCS*, pages 129–138. Springer, 2006.

8 J. Edmonds and E. L. Johnson. Matching, euler tours and the chinese postman. *Mathematical Programming*, 5:88–124, 1973.

9 F. Ergün, R. Kumar, and R. Rubinfeld. Fast approximate pcps. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing, May 1-4, 1999, Atlanta, Georgia, USA*, pages 41–50, 1999.

10 M. Goemans and D. Williamson. A general approximation technique for constrained forest problems. *SIAM J. Comput.*, 24(2):296–317, 1995.

11 S. Guha, A. Meyerson, and K. Munagala. Hierarchical placement and network design problems. In *In Proceedings of the 41th Annual IEEE Symposium on Foundations of Computer Science, FOCS'00*, pages 603–612. IEEE Computer Society, 2000.

12 M. Laszlo and S. Mukherjee. Approximation Bounds for Minimum Information Loss Microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1643–1647, 2009.

13 C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43:425–440, 1991.

14 P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, November 2001.

15 A. Schrijver. *Combinatorial Optimization*. Springer, 2003.

16 A. Shalita and U. Zwick. Efficient algorithms for the 2-gathering problem. *ACM Transactions on Algorithms*, 6(2), 2010.

17 K. Stokes. On computational anonymity. In *Privacy in Statistical Databases – UNESCO Chair in Data Privacy, International Conference, PSD 2012, Palermo, Italy, September 26-28, 2012. Proceedings*, pages 336–347, 2012.

18 C. Tovey. A Simplified NP-complete Satisfiability Problem. *Discrete Applied Mathematics*, 8(1):85–89, 1984.

19 D. Xu, E. Anshelevich, and M. Chiang. 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*, pages 186–190, 2008.