Nordic Journal of Computing 1(1994), 246–263.

SOME GEOMETRIC CLUSTERING PROBLEMS*

ULRICH PFERSCHY
TU Graz, Institut für Mathematik B
Kopernikusgasse 24
A-8010 Graz, Austria
pferschy@ftug.dnet.tu-graz.ac.at

RÜDIGER RUDOLF
TU Graz
Institut für Mathematik B
Kopernikusgasse 24
A-8010 Graz, Austria

GERHARD J. WOEGINGER
TU Graz, Institut für Theoretische Informatik,
Klosterwiesgasse 32/II
A-8010 Graz, Austria

Abstract. This paper investigates the computational complexity of several clustering problems with special objective functions for point sets in the Euclidean plane. Our strongest negative result is that clustering a set of $3k$ points in the plane into k triangles with minimum total circumference is NP-hard. On the other hand, we identify several special cases that are solvable in polynomial time due to the special structure of their optimal solutions: The clustering of points on a convex hull into triangles; the clustering into equal–sized subsets of points on a line or on a circle with special objective functions; the clustering with minimal cluster-distances. Furthermore, we investigate clustering of planar point sets into convex quadrilaterals.

ACM CCS Categories and Subject Descriptors: F.2.2

1. Introduction

Problem statement.
Let P be a set of points in the plane. A partitioning of P into k disjoint (possibly empty) sets $C_1, C_2, \ldots, C_k$ is called a *clustering*, and the individual sets C_i are called its *clusters*. In cluster analysis, the points represent properties (data) of real-world objects, and the aim is usually to collect "similar" objects (points which are close to each other) in the same cluster, and to put objects which are very "different" into different clusters.

The definition of "similarity" of objects is crucial for every clustering process. In a general setup we let W be some weight function that assigns a real weight to any set of finite point sets $C_1, \ldots, C_k$ in the plane (Examples for W are the maximum diameter of all C_i or the sum of the circumferences of

*This research was partially supported by the Christian Doppler Laboratorium für Diskrete Optimierung and by the Fonds zur Förderung der wissenschaftlichen Forschung, Project P8971-PHY.

Received November 1993. Accepted February 1994.

the convex hulls of all C_i or the distances between all pairs of points in the same point set). Intuitively, W is a measure of the quality of the clustering $C_1, \ldots, C_k$. Then the *planar clustering problem for* W is defined as follows.

> **INSTANCE:** A set P of m points in the plane; integers k, n_l and n_u; a rational number d.
> **QUESTION:** Is there a clustering for P into k sets $C_1, C_2, \ldots, C_k$ such that $n_l \leq |C_i| \leq n_u$ and such that $W(C_1, C_2, \ldots, C_k) \leq d$ holds?

Clearly, this problem could be defined in higher dimensions, but we confine our interest to the plane. Usually, not all of the numbers k, n_l and n_u are specified; sometimes they are specified but not as part of the input. Sometimes there are additional restrictions on the clusters (e.g. the convex hulls are required to be pairwise disjoint). The special case where $n_l = n_u = |P|/k$ (i.e. all clusters contain the same number of points) occurs frequently in practical problems and is called *balanced clustering.*

Related results.
In general, the above problem is NP-complete. Supowit [16] has shown the NP-completeness if W assigns the maximum diameter of all C_i and if k is part of the input. The related problem of minimizing the maximum radius, which is also known as the k-center problem in the area of location problems, is also NP-complete (Megiddo and Supowit [11]). NP-completeness can also be shown for minimizing the maximum cluster area and for minimizing the sum of all cluster areas, as follows from a result of Megiddo and Tamir [12] that it is NP-complete to decide whether a set of points can be covered by a given number of lines. Some special cases that are solvable in polynomial time can be found in [1, 3, 13]. For more information, the reader is referred to Johnson's NP-Completeness Column [9] and to the article by Brucker [2].

Our results.
A problem where one could expect to find some useful characterization of optimal clusterings, is the partitioning of points into triangles. However, this special case of balanced clustering of $m = 3k$ points into clusters C_i, $i = 1 \ldots k$ of size three with a minimal sum of all triangle circumferences turns out to be NP-complete in general. On the other hand, we get a polynomial algorithm for finding the optimal clustering if the set of points is restricted to the boundary of a convex set.

A property which is often discussed in the treatment of geometric clustering problems is the "convex separability" of clusters in an optimal solution, which means that the convex hulls of each two optimal clusters can be separated by a line. In this context the partitioning of points on a line and on a sphere was investigated by Boros and Hammer [1]. For the case of a balanced clustering which minimizes the sum of all euclidean distances between points in the same cluster we show that points on a line can be clustered optimally such that the convex hulls of the clusters are disjoint. The same

does not hold for arbitrary point sets in the plane as can be shown by a simple counterexample. Hence, we try to find layouts of points narrowing the "gap" between these two setups. We show that points on a circle can in fact be separated in a way similar to points on a line for a special objective function. The case of points on the boundary of a convex set remains open.

A rather different and more general clustering problem which minimizes the sum of all distances between points in *different* clusters which is equivalent to *maximizing* the sum of distances of points in the same cluster can be solved quite easily in any metric space due to the special structure of the optimal solution.

Extending the inspection of clusters of size three (triangles) we also discuss clusters of size four (quadrilaterals), especially convex quadrilaterals. In this case an existence result can be given, namely that every set of $4k$ points in the plane *can* be clustered into at least $k-1$ convex quadrilaterals.

Relation to cut problems.
Clustering problems are in a general way related to *cut problems*. The extent of this relationship is determined by the nature of the weight function W. In fact, if we define W as

$$W(C_1,\ldots,C_k) := \sum_{p,q\in C_1} d(p,q) + \ldots + \sum_{p,q\in C_k} d(p,q)$$

i.e. the sum of all distances between points in the same cluster as we do in Sections 2 and 3, we get a clustering problem equivalent to the well–known *max–cut problem*. Its purpose is to *maximize* the sum of all distances of points in *different* clusters, points whose connection lines cross the "cut" dividing the clusters from each other. References and applications can be found in the book of Lengauer [10].

The objective function in the mentioned case of minimization of distances between different clusters defines a clustering problem equivalent to a *min–cut problem*. This problem can be solved in the general case, where the "distances" are arbitrary "edge weights", by an algorithm due to Goldschmidt and Hochbaum [8] in $O(m^{k^2})$ time. Many approximation algorithms (e.g. Saran and Vazirani [15]) and identifications of special cases are known for this problem.

Organization of the paper.
In Section 2 we deal with the partitioning of points into triangles minimizing the sum of circumferences. Other balanced clustering problems are treated in Sections 3 and 4 minimizing the sum of distances of points in the same cluster for points on a line in the former and minimizing the sum of the circumferences of the convex hulls of the clusters for points on a circle in the latter.

The distances between clusters are minimized for general point sets in Section 5. Clusterings whose subsets are convex quadrilaterals are discussed together with some combinatorial results in Section 6. We close the paper with a short discussion and remarks in Section 7.

2. Clustering into Triangles

In this section, we investigate the problem of clustering a planar set P of $m = 3k$ points into k triangles such that the sum of all triangle circumferences is minimized. We will show that this problem is NP-complete in general. However, if we restrict P to be a set of points lying on the boundary of a convex set, the optimum clustering can be found in polynomial time, $O(m^4)$.

EXAMPLE 1. Consider the following set of six points a, b, c, a', b', and c': Points a, b and c form an equilateral triangle of side length 1. Points a', b', c' are at distance $\varepsilon < 1/100$ from the center of the line segment $\overline{ab}$. a' is inside and b' and c' are outside of the triangle Δabc. It is easily checked that the optimum sum of circumferences clustering into two triangles consists of the clusters $\{a, b, c\}$ and $\{a', b', c'\}$, for ε sufficiently small.

Since the convex hulls of these two clusters intersect, this example demonstrates that the optimum clustering is not necessarily crossingfree. This is surprising since most geometric minimization problems have crossingfree optimum solutions, e.g. Minimum Matching, the shortest Traveling Salesman Tour, Minimum Maximum Diameter Clustering etc.

Next, we will give our NP-completeness result. We will make use of *rectilinear planar layouts* of planar graphs. A *rectilinear planar layout* of a planar graph $G = (V, E)$ maps the vertices in V to horizontal line segments and the edges of G to vertical line segments, with all endpoints of segments at positive integer coordinates. Two horizontal segments are connected by a vertical segment if and only if the corresponding vertices are adjacent in the graph. The following proposition is due to Rosenstiehl and Tarjan [14].

PROPOSITION 1. *Given a planar graph $G = (V, E)$, a rectilinear planar layout of G can be computed in time polynomial in the size of G. Moreover, the height and the width of the layout are both linear in the size of G. W.l.o.g. we may assume that all horizontal segments are at different (integer) heights.* □

Our NP-completeness proof will be done by a reduction from the following very special version of the exact cover problem that was shown to be NP-complete by Dyer and Frieze [5].

> PLANAR EXACT COVER BY 3-SETS (Planar X3C)
>
> **Input.** A set Q with $|Q| = 3q$; a set T of triples from $Q \times Q \times Q$ such that (i) every element of Q occurs in at most three triples and such that (ii) the induced graph G is planar. (This induced graph G is defined as follows: It contains a vertex for every element of Q and for every triple in T. There is an edge connecting a triple to an element if and only if the element is a member of the triple. Clearly, G is bipartite with vertex bipartition Q and T).

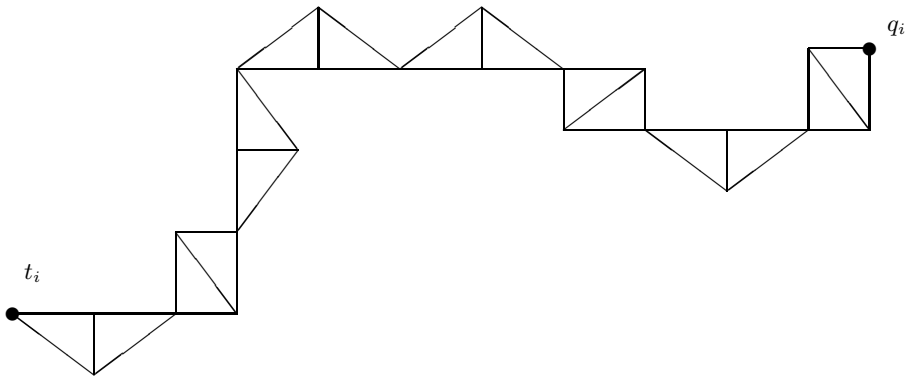

Fig. 1: A chain of diamonds connecting t_i to q_i

Question. Does there exist a subset of q triples in T which contains all the elements of Q ?

Hence, let Q and $T \subset Q \times Q \times Q$ constitute an instance of planar X3C. We will construct a point set $P(Q,T)$ of $3k$ points (the exact value of k will be determined later) that allows a clustering into triangles with total circumference at most $12k$ if and only if the planar X3C instance has a solution.

In a first step, we compute a rectilinear planar layout for the underlying undirected graph of G according to Proposition 1. Then we multiply all coordinates by a factor of 1000 in order to ensure that points on distinct horizontal (vertical) segments are sufficiently far away from each other.

Next, we define the point set $P(Q,T)$. Our main tool is the right-angled triangle Δ_0 with side lengths 3, 4 and 5; it will allow us to keep all points in $P(Q,T)$ at integer coordinates. For every element of Q, $P(Q,T)$ will contain a so-called *element point*. For every triple in T, $P(Q,T)$ contains three so-called *triple points* forming a so-called *triple triangle*. The triple triangles are copies of Δ_0 such that the two sides of lengths 3 and 4 are axes-parallel. The element points and the triple triangles are placed somewhere at the corresponding line segments in the rectilinear layout (because of our multiplication, there is ample space to place them).

In the next step, we consider some triple $t = (q_1, q_2, q_3)$ in T and the corresponding three triple points t_1, t_2 and t_3 that form a triangle Δ_0. For $1 \leq i \leq 3$, the point t_i is connected to the point q_i by a *chain of diamonds* as depicted in Figure 1. A *diamond* consists of two copies of Δ_0 that are glued together either by their sides of length 5 (rectangular diamond) or

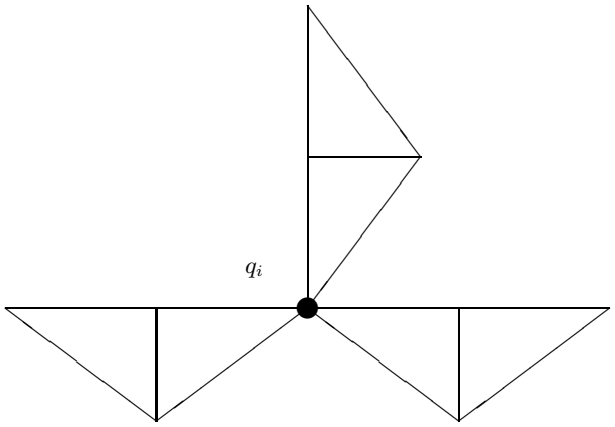

Fig. 2: Placement of diamonds around an element point q_i

by their sides of length 3 (triangular diamond). All diamonds are placed in such a way that the two shorter sides of the triangles are axes-parallel. No two rectangular diamonds occur consecutively in a chain. The chains of diamonds (roughly) follow the line segments corresponding to the two vertices t_i and q_i and to the connecting edge in the graph G.

There are two problems we have to be careful with. (1) The first problem is that we do not want to produce other triangles of circumference 12 outside of these chains of diamonds (e.g. if two distinct chains come very close to each other). Again, because of our multiplications in the beginning, there are sufficiently many degrees of freedom to route the chains far away from each other. Since every element occurs in at most three triples, it is also possible to keep the chains sufficiently far from each other if they meet in an element point (see Figure 2). The same holds for triple points. (2) The other problem is that it is not a priori clear that triple and element points can indeed be connected by such chains of diamonds: Triangular shaped diamonds shift the path by ± 8 units in x-direction and 0 units in y-direction (or 0 units in x-direction and ± 8 units in y-direction); rectangular shaped diamonds shift the path by a vector of $(\pm 3, \pm 4)$ or $(\pm 4, \pm 3)$. Once more, our multiplication in the beginning of the construction removes this problem. The main idea is to use only a small number of rectangular shaped diamonds in order to reach the correct remainders for the shift if divided by 8, and to use many triangular shaped diamonds for the long distances. As the distances are sufficiently large, we can mix rectangular shaped and triangular shaped diamonds without ending up in problem (1).

Connecting every triple triangle to its corresponding three element triangles completes the construction of the point set $P(Q, T)$. It is easy to check

that $|P(Q,T)|$ is divisible by three, say $|P(Q,T)| = 3k$ and that the construction can be performed in polynomial time. It remains to prove that $P(Q,T)$ can be partitioned into triangles of total circumference at most $12k$ if and only if the exact cover problem allows a solution.

(If) Assume that a clustering of total length $\leq 12k$ exists. By our construction, no three points in $P(Q,T)$ form a triangle with circumference < 12; consequently, the clustering must consist of k triangles all with circumference exactly 12. We claim that the subset T' of T that contains all triples for which the corresponding triple points form a cluster constitutes a solution to the planar X3C instance.

Consider some element $q_i \in Q$. The corresponding element point is contained in exactly one cluster, and this cluster belongs to a chain of diamonds. In this chain of diamonds, every other triangle must form a cluster; therefore, the corresponding point t_i on the other end of the chain *cannot be covered* by any triangle cluster in the chain, and the corresponding three triangle points must form another cluster. The clusters in the other (one or two) chains going away from q_i cover the corresponding triangle points, and these triangle points cannot form a cluster in the clustering.

This way we may assign to each $q_i \in Q$ a unique triple in T. On the other hand, if we assign one q_i to some triple, the other two elements in this triple must be assigned to this triple, too. Clearly, this yields a solution to the X3C.

(Only if) Now assume that the planar X3C has a solution $T' \subseteq T$. We construct a clustering as follows. All triangles corresponding to triples in T' are clusters, all triangles in $T \setminus T'$ are not. This completely determines which triangles on the chains have to be used in the clustering. Since T' is a solution of X3C, every element point is in exactly one cluster.

Summarizing, we have proved the following theorem.

THEOREM 1. *For a set P of $3k$ points with integer coordinates in the plane and an integer d, it is NP-hard to decide whether P can be clustered into k triangles with total circumference at most d.* □

REMARK. Since we do not know whether the above clustering problem is in NP, we only could prove an NP-hardness result. However, if we use *discretized* distances and circumferences (i.e. the circumference of Δxyz is defined by $\lceil\overline{xy}\rceil + \lceil\overline{xz}\rceil + \lceil\overline{yz}\rceil$), then the discretized clustering problem clearly is in NP. All 'important' triangles in our construction above have integer side lengths. Therefore, it follows that the discretized clustering problem into triangles is NP-complete.

REMARK. Our construction also shows that finding a clustering into triangles that minimizes the largest cluster circumference is NP-hard.

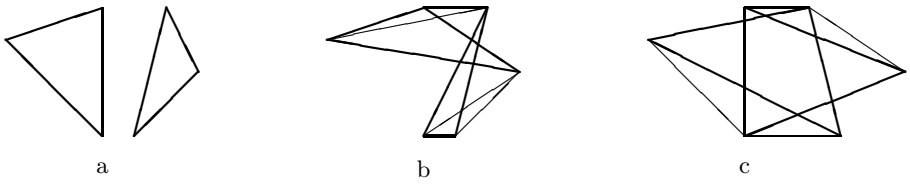

Fig. 3: All different situations with triangles of 6 convex points: Thick lines show the original triangles, thin lines the added edges.

Next, we consider the restriction where the set P consists of points on the boundary of a convex set. We will show that this restriction makes the problem easy.

LEMMA 1. *Let P be a set of six points in convex position. Then the clustering into two triangles with minimum total circumference is crossingfree.*

PROOF. Given 6 points in convex position, there are only three combinatorially distinct ways to construct two triangles (confer to Figure 3).

We only have to show that Cases b and c are not optimal. By applying the quadrangle inequality twice in both cases as indicated in the picture, one easily constructs a pair of non-intersecting triangles with smaller total circumference. □

THEOREM 2. *For a set P of $3k$ points on the boundary of a convex set in the plane a clustering into k triangles with minimum total circumference can be found in $O(k^4)$ time.*

PROOF. Lemma 1 implies that the optimum clustering is crossingfree. Hence, we may apply a dynamic programming approach to find the optimum clustering in the following way.

Let $p_1, p_2, \ldots, p_{3k}$ denote the points of P sorted clockwise around the convex set. We introduce a two-dimensional array $\text{TRI}[i,j]$, $1 \leq i,j \leq 3k$. The entry $\text{TRI}[i,j]$ concerns the subset P_{ij} of P that lies between p_i and p_j in clockwise direction on the convex boundary (inclusively the points p_i and p_j). In case $|P_{ij}|$ is divisible by three, $\text{TRI}[i,j]$ contains the length of the shortest clustering of P_{ij} into triangles. Otherwise, $\text{TRI}[i,j] = \infty$.

We show how to compute all entries of $\text{TRI}[*,*]$ in $O(k^4)$ overall time. This is done in a bottom-up fashion. First, we set all $\text{TRI}[i,j] = \infty$ for which $|P_{ij}|$ is not divisible by three. Then we go through k rounds. In round x, we compute $\text{TRI}[i,j]$ for the sets P_{ij} containing exactly $3x$ points. Round 1 is easy: P_{ij} contains three points, and the clustering is unique. For $x \geq 2$, we

test all possible triangles in P_{ij} containing the point p_i and combine them with the already calculated optimum values for the rest of P_{ij}.

$$\begin{aligned}\textsc{Tri}[i,j] \quad &= \min_{p_a,p_b \in P_{ij}} \quad \{\ \overline{p_a p_b} + \overline{p_b p_i} + \overline{p_i p_a} + \textsc{Tri}[i+1, a-1] \\ &\qquad + \textsc{Tri}[a+1, b-1] + \textsc{Tri}[b+1, j]\ \}\end{aligned}$$

If we implement the above equation, we have to take care that $a \neq b$ holds, and in case some of p_i, p_a, p_b and p_j are neighbors on the convex boundary (e.g. p_a and p_b), then the corresponding term (e.g. $\textsc{Tri}[a+1, b-1]$) must not be considered in the sum. Clearly, the computation of every entry in $\textsc{Tri}[*,*]$ takes at most $O(k^2)$ time; this gives the claimed overall time complexity of $O(k^4)$. □

3. Balanced Clustering of Points on a Line

It is shown that the clustering of $m = nk$ points on a line into k clusters of equal size, i.e. a balanced clustering, with a minimum sum of all distances between points of the same subset consists of k disjoint segments of the line each containing n points. The same problem without restrictions to the size of the clusters was treated by Boros and Hammer [1], who showed that every optimal clustering is *nested* i.e. $\forall i \neq j\ C_i \cap (\text{conv } C_j) = \emptyset$ or $C_j \cap (\text{conv } C_i) = \emptyset$.

THEOREM 3. *Let $P = \{p_1, \ldots, p_{2n}\}$ be a set of points on a line. The balanced clustering of P into two sets, which minimizes*

$$W(C_1, C_2) := \sum_{p,q \in C_1} d(p,q) + \sum_{p,q \in C_2} d(p,q)$$

consists of the first n points detected by scanning the line starting from one endpoint and the remaining n points nearer to the other endpoint of the line.

PROOF. Let $C = (C_1, C_2)$ be an arbitrary balanced clustering of P. We choose an arbitrary point $z \notin P$ such that there are exactly n points of P to the right of z and exactly n points of P to the left of z. We represent each p_i by its distance from the left endpoint. Hence we have $d(p_i, p_j) = |p_i - p_j|$.

Let $C_1^L = \{a_1, \ldots, a_k\}$, $C_2^L = \{b_{k+1}, \ldots, b_n\}$ such that $p - z < 0\ \forall p \in \{C_1^L \cup C_2^L\}$ and $C_2^R = \{b_1, \ldots, b_k\}$, $C_1^R = \{a_{k+1}, \ldots, a_n\}$ such that $q - z > 0\ \forall q \in \{C_1^R \cup C_2^R\}$.

The distance between two sets is defined by $d(A,B) := \sum_{a \in A} \sum_{b \in B} d(a,b)$. We have to show that the clustering $(C_1^L \cup C_2^L, C_1^R \cup C_2^R)$ yields a smaller value of W than C:

$$d(C_1^L, C_1^R) + d(C_2^L, C_2^R) \geq d(C_1^L, C_2^L) + d(C_1^R, C_2^R) \tag{1}$$

Inequality (1) can be written as

$$\begin{aligned}
&(n-k)\sum_{i=1}^{k}|a_i-z|+k\sum_{j=k+1}^{n}|z-a_j| \\
&\quad+k\sum_{i=k+1}^{n}|b_i-z|+(n-k)\sum_{j=1}^{k}|z-b_j| \\
\geq\ &\sum_{i=1}^{k}\sum_{j=k+1}^{n}|a_i-b_j|+\sum_{i=k+1}^{n}\sum_{j=1}^{k}|a_i-b_j| \qquad (2)
\end{aligned}$$

Using elementary inequalities like $||a|-|b|| \leq |a-b|$ we get

$$\sum_{i=1}^{k}\sum_{j=k+1}^{n}(|a_i-b_j|-|a_i-z|)+\sum_{i=k+1}^{n}\sum_{j=1}^{k}(|a_i-b_j|-|b_j-z|)\leq$$

$$\sum_{i=1}^{k}\sum_{j=k+1}^{n}\Big||a_i-b_j|-|a_i-z|\Big|+\sum_{i=k+1}^{n}\sum_{j=1}^{k}\Big||a_i-b_j|-|b_j-z|\Big|\leq$$

$$\sum_{i=1}^{k}\sum_{j=k+1}^{n}|z-b_j|+\sum_{i=k+1}^{n}\sum_{j=1}^{k}|a_i-z|=$$

$$k\sum_{j=k+1}^{n}|z-b_j|+k\sum_{i=k+1}^{n}|a_i-z|$$

Moving terms around yields inequality (2). □

COROLLARY 1. *With the conditions of Theorem 3 the balanced clustering of nk points on a line into k clusters consists of sets $C_1,\dots,C_k$ such that $int(conv\ C_i)\cap int(conv\ C_j)=\emptyset$ for all $i,j\in\{1,\dots,k\}, i\neq j$.*

PROOF. Let $C_1,\dots,C_k$ be an arbitrary balanced clustering. If for any pair $C_i,\ C_j$ $\mathrm{int}(\mathrm{conv}\ C_i)\cap\mathrm{int}(\mathrm{conv}\ C_j)\neq\emptyset$ holds we can decrease the value W of the clustering by exchanging elements of C_i and C_j according to Theorem 3. Successive modification of all overlapping pairs of sets yields our statement. □

REMARK. Obviously, the computation of an optimal clustering with respect to the conditions given above can be done in $O(m\log m)$ time by sorting the points and scanning the line.

4. Balanced Clustering of Points on a Circle with Minimal Circumference

It is shown that the balanced clustering of $m = nk$ points on a circle into k clusters with a minimum sum of circumferences consists of k non–intersecting segments each containing n points. The circumference of a set of points is the circumference of its convex hull.

THEOREM 4. *Let $P = \{p_1, \ldots, p_{2n}\}$ be a set of points on a circle of arbitrary radius. If P is partitioned into two sets of equal size such that the sum of their circumferences is minimized then the two sets can be separated by a line i.e. their convex hulls are disjoint.*

PROOF. See Appendix. □

COROLLARY 2. *A balanced clustering of nk points on a circle minimizing*

$$W(C_1, \ldots, C_k) := \sum_{i=1}^{k} \textit{ circumference } (C_i)$$

consists of sets $C_1, \ldots, C_k$ such that int(conv C_i) $\cap$ int(conv C_j) $= \emptyset$ for all $i, j \in \{1, \ldots, k\}, i \neq j$.

PROOF. Let $C_1, \ldots, C_k$ be an arbitrary balanced clustering. If for any pair C_i, C_j $\text{int}(\text{conv } C_i) \cap \text{int}(\text{conv } C_j) \neq \emptyset$ holds we can diminish the sum of the circumferences by exchanging elements of C_i and C_j according to Theorem 4. Successive modification of all intersecting pairs of sets yields our statement. □

REMARK. An optimal clustering can be found by applying dynamic programming as in the proof of Theorem 2. This yields an $O(m^{n+1})$ algorithm which is polynomial for every fixed n.

5. Clustering with Minimal Clusterdistances

We treat the general problem of finding an arbitrary clustering of m points in any metric space into non–empty subsets minimizing the sum of all distances between points of *different* subsets. Hence, our objective is to minimize

$$\tilde{W}(C_1, \ldots, C_k) := \sum_{\substack{i,j=1 \\ i \neq j}}^{k} \sum_{\substack{p \in C_i \\ q \in C_j}} d(p, q).$$

It is shown that the optimal solution is achieved by a partition into $k - 1$ sets of only one point and one set containing all other $m - k + 1$ points.

THEOREM 5. *Let M be a metric space and $P = \{p_1, \ldots, p_m\}$ a set of points in M. If P is partitioned into two subsets such that $\tilde{W}$ is minimized then the resulting clustering consists of one single point $\tilde{p} \in P$ defined by*

$$\tilde{p} = \arg\min_{p_i} \{\sum_{j=1}^{m} d(p_i, p_j)\}$$

and the set of all other points $P \setminus \{\tilde{p}\}$.

PROOF. Let $C = (C_1, C_2)$ be an arbitrary clustering of P with $|C_1| \leq |C_2|$. We show that there exists a single point $\bar{p} \in C_1$ such that the value of the clustering $K = (\bar{p}, P \setminus \{\bar{p}\})$ is less than or equal to $\tilde{W}(C_1, C_2)$. Obviously $\tilde{p}$ is the best possible selection for any $\bar{p}$ and we are done.

Let $C_1 = \{p_1, \ldots, p_k\}$ and $C_2 = \{p_{k+1}, \ldots, p_m\}$ with $k \leq m/2$. We denote the distance of p_i and p_j by $d(p_i, p_j) =: d(i, j)$. Changing from clustering C to clustering $K_i = (p_i, P \setminus \{p_i\})$ the distances between p_i and all other points in C_1 are added to the value of $\tilde{W}(C)$ and we will denote this amount by

$$A_i := \sum_{j=1}^{k} d(i, j).$$

All distances between $C_1 \setminus \{p_i\}$ and C_2 are deleted from $\tilde{W}(C)$ denoted by

$$D_i := \sum_{\substack{j=1 \\ j \neq i}}^{k} \sum_{\ell=k+1}^{m} d(j, \ell).$$

We will show that $\sum_{i=1}^{k} A_i \leq \sum_{i=1}^{k} D_i$ and therefore there exists some j such that $A_j \leq D_j$. Choosing $\bar{p} = p_j$ satisfies our claim.

$$\sum_{i=1}^{k} A_i = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} d(i, j) \leq \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} \left(d(i, k+j) + d(k+j, j)\right).$$

Multiplying the first term by $k - 1$ and evaluating the second term yields

$$\sum_{i=1}^{k} A_i \leq (k-1) \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} d(i, k+j) + (k-1) \sum_{j=1}^{k} d(j, k+j) =$$

$$(k-1) \sum_{i=1}^{k} \sum_{j=1}^{k} d(i, k+j) = \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} \sum_{\ell=k+1}^{2k} d(j, \ell) \leq \sum_{i=1}^{k} D_i.$$

□

COROLLARY 3. *With the conditions of Theorem 5 the clustering into k subsets minimizing $\tilde{W}$ consists of $k-1$ single points $p_{i_1}, \ldots, p_{i_{k-1}}$ and the set of the remaining $m-k+1$ points.*

PROOF. Let $C_1, \ldots, C_k$ be an arbitrary clustering. Obviously, in an optimal clustering any pair (C_i, C_j), $i, j \in \{1, \ldots, k\}$ has to be an optimal partition of $\hat{P} := C_i \cup C_j$ into two sets. Applying Theorem 5 successively yields our statement. □

REMARK. The computation of $\tilde{p}$ by a straight forward algorithm takes $O(m^2)$ time. Iterative application of this selection process yields an $O(km^2)$ algorithm.

6. Clusterings into Convex Parts

In this section, we deal with clusterings into *convex* parts. Among several related results we will show that any set P of $4k$ points in the plane can be clustered into k *convex* quadrilaterals if it has at least five points on its convex hull.

Let us define $\phi(n)$, $n \geq 4$, to be the smallest number ϕ such that any set of $m = kn$ points in the Euclidean plane with at least ϕ points on the convex hull can be partitioned into k convex clusters $C_1, \ldots, C_k$, each of cardinality n. (We will assume that all point sets treated in this section are in general position i.e. they have no three collinear points). The numbers $\phi(n)$ are closely related to the Erdős-Szekeres numbers $\mathrm{ES}(n)$. $\mathrm{ES}(n)$ is defined as the smallest integer number e such that every planar set of e points or more contains a convex n-gon as subset. It is known (see Erdős and Szekeres [6]) that $\mathrm{ES}(4) = 5$, $\mathrm{ES}(5) = 9$ and that

$$2^{n-2} + 1 \leq \mathrm{ES}(n) \leq \binom{2n-4}{n-2} + 1.$$

In what follows, we will show that the numbers $\phi(n)$ exist for all $n \geq 4$ and that there are positive reals c_1, c_2 such that

$$c_1 \mathrm{ES}(n)/n \leq \phi(n) \leq c_2 n \mathrm{ES}(n)$$

holds. For $n = 4$ we give the exact value $\phi(4) = 5$. This leaves open the nice possibility of $\phi(n) \equiv \mathrm{ES}(n)$.

First we will give the general bounds on $\phi(n)$ in terms of $\mathrm{ES}(n)$.

THEOREM 6. *For $n \geq 4$, $\lfloor (\mathrm{ES}(n) - n - 1)/(n-1) \rfloor < \phi(n) \leq (n-2)\lceil (\mathrm{ES}(n) - 1)/2 \rceil + n + 1$ holds.*

PROOF. (Lower bound). Define $f_1(n) = \lfloor (\mathrm{ES}(n) - n - 1)/(n-1) \rfloor$ and let P_1 be a set of $(n-1)f_1(n) + n < \mathrm{ES}(n)$ points in the plane that does not contain any convex n-gon as subset. Put a scaled copy of P_1 into the unit

circle and choose a set P_2 of $f_1(n)$ equidistant points on the unit cycle. Then the set $P_1 \cup P_2$ is not partitionable into convex n-gons: Otherwise, there exist $f_1(n)+1$ convex n-gons and at least one of them must be contained in P_1, a contradiction.

(Upper bound). Define $f_2(n) = (n-2)\lceil(\mathrm{ES}(n)-1)/2\rceil + n + 1$ and let P_3 be any planar point set with $h \geq f_2(n)$ points on the convex hull and i points in the interior of the hull, $h+i = kn$. We assume that P_3 is not partitionable into k convex n-gons and derive a contradiction.

First, we repeatedly produce convex n-gons in the interior of the convex hull until the number of interior points is less than $\mathrm{ES}(n)$. Then we consider any pair of interior points x and y. The line going through x and y divides the points on the convex hull into two parts, one of them containing at least $n-2$ points. These $n-2$ points together with x and y form a convex n-gon. Repeating this procedure we derive a feasible partition. □

THEOREM 7. $\phi(4) = 5 = \mathrm{ES}(4)$ *holds.*

PROOF. (reductio ad absurdum) We will assume that there exists a set P of $4k$ points which has at least five points on its convex hull and which is not partitionable into convex quadrilaterals (we will call such point sets *bad* point sets); from this we will derive a contradiction. Indeed, assume bad point sets exist and consider some bad point set P_4 with the minimum number of points.

We observe that P_4 cannot contain more than 9 points: Otherwise, we could select a subset $P_5 \subseteq P_4$ consisting of 5 points such that in $P_4 \backslash P_5$ there remain at least 5 hullpoints. Using $\mathrm{ES}(4) = 5$, we find a convex quadrilateral in P_5. If we remove this convex quadrilateral from P_4, we end up with a smaller bad set, a contradiction.

Therefore, P_4 contains at most 9 and at least 5 points; this implies $|P_4| = 8$. We will distinguish whether the convex hull of P_4 consists of eight, seven, six or five points and we will find a feasible partition in each case.

8. There is nothing to show, any partition will do.
7. Consider any line through the only inner point x of P_4. The line divides the convex hull into two parts, one of these two parts contains three points of P_4. The point x together with these three points gives one convex quadrilateral, the remaining four points form the other one.
6. Similarly as in case (7), we consider the line through the inner points x and y. One of the originating parts contains at least two points, and these two points together with x and y give a feasible convex quadrilateral.
5. In this case, we consider the three lines determined by the three inner points x, y and z. They partition the plane into one finite and six infinite regions (see Figure 4). Three of the infinite regions are wedges, three of them are three-sided. All hullpoints lie in these six infinite regions.

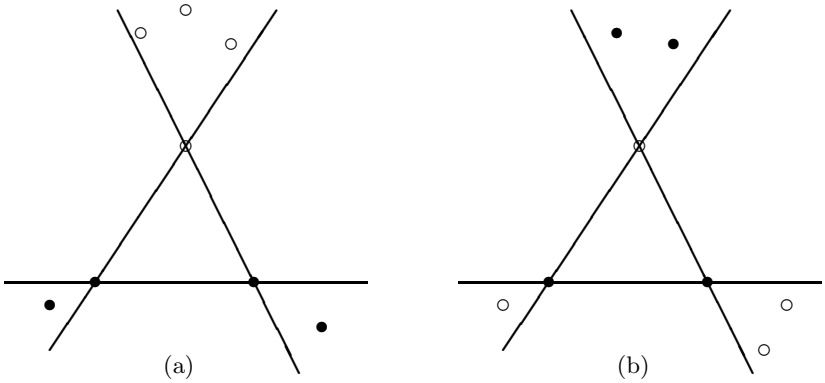

Fig. 4: Partitions of the plane by three lines

A hull point in one of the three-sided regions can be used to form a feasible convex quadrilateral with a, b and c. Three hullpoints in one wedge can be put together with the nearest inner point to produce a feasible quadrilateral (see Figure 4a). It remains to treat configurations of the form in Figure 4b. But here we simply combine two points in a wedge with the two farthest inner points and again we find a feasible partition. Consequently, P_4 is partitionable into convex quadrilaterals.

To complete the proof of Theorem 7, we must show that $\phi(4) > 4$. To see this we consider a set P_6 of eight points lying on the two branches of some hyperbola. Let one branch contain three and the other branch contain five points of P_6. Then the convex hull of P_6 consists of exactly four points, and obviously P_6 is not partitionable into two convex quadrilaterals. □

COROLLARY 4. *Every planar point set of* $4k$ *points in general position can be clustered into* k *clusters of cardinality four in such a way that at least* $k-1$ *of the clusters are convex quadrilaterals.* □

Summarizing, five points on the convex hull are sufficient to make a point set partitionable into convex quadrilaterals. We do not know an easy characterization for partitionable point sets with only three or four points on the convex hull. The problem remains open to find a *polynomial time* algorithm for solving this problem.

7. Discussion

In this paper we treated various cases of the general geometric clustering problem. On one hand we showed that it is NP-complete to find a partitioning of $3k$ points in the plane into k triangles with a minimal sum of circumferences or a minimal largest cluster circumference. On the other hand we

gave a polynomial algorithm for solving the same problem for points lying on the boundary of a convex set.

Identifying reasonable special cases of the balanced clustering problem, whose optimal solution can be characterized in such a way that its computation is more or less easy, we showed that the convex hulls of the optimal clusters are disjoint in the case of points on a line (minimizing the sum of distances of points in the same cluster) and of points on a circle (minimizing the sum of circumferences of the clusters). Other versions of objective functions and special positions of points remain to be considered, especially points in convex position.

A result in a more general context was that a partition into k non-empty clusters minimizing the sum of distances between points in different clusters consists of $k-1$ singleton sets and one set of all remaining points.

Exploiting the relation to cut problem, the *terminal cut problem* (see e.g. Cunningham [4]) could be translated into a *terminal clustering*, where each cluster has to contain a special *terminal point*. Hence, the clustering defines not only a partition of all points but also a partition of k special terminal points into k subsets. Problems of this kind occur frequently in location theory and seem to be worth to be investigated.

References

[1] E.Boros and P.L.Hammer, On clustering problems with connected optima in Euclidean spaces, *Discrete Mathematics* **75**, 1989, 81–88.

[2] P.Brucker, On the complexity of Clustering Problems, in: R. Henn, B. Korte, and W. Oettli, eds., *Optimization and Operations Research*, Springer Verlag 1977, pp. 45–54.

[3] V.Capoyleas, G.Rote and G.J.Woeginger, Geometric Clusterings, *J. Algorithms* **12**, 1991, 341–356.

[4] W.H.Cunningham, The optimal multiterminal cut problem, *DIMACS Series in Discrete Math. and Theoretical Comp. Sc.* **5**, 1991, 105–120.

[5] M.E.Dyer and A.M.Frieze, Planar 3DM is NP-complete, *J. Algorithms* **7**, 1986, 174–184.

[6] P.Erdős and G.Szekeres, A combinatorial problem in geometry, *Compositio Math.* **2**, 1935, 463–470.

[7] M.R.Garey and D.S.Johnson, *Computers and Intractability*, Freeman, San Francisco, 1979.

[8] O.Goldschmidt and D.S.Hochbaum, Polynomial algorithm for the k–cut problem, *Proc. of 29th Symposion on Foundations of Comp. Sci.*, 1988, 444–451.

[9] D.S.Johnson, The NP-Completeness Column: An Ongoing Guide, *J. Algorithms* **3**, 1982, 182–195.

[10] T.Lengauer, *Combinatorial Algorithms for Integrated Circuit Layout*, J.Wiley, Chichester, 1990.

[11] N.Megiddo and K.J.Supowit, On the complexity of some common geometric location problems, *SIAM J. Computing* **13**, 1984, 182–196.

[12] N.Megiddo and A.Tamir, On the complexity of locating linear facilities in the plane, *Oper. Res. Lett.* **13**, 1982, 194–197.

[13] C.Monma and S.Suri, Partitioning points and graphs to minimize the maximum or the sum of diameters, *Proc. Sixth Int. Conf. Theory and Appl. of Graphs*, Wiley, 1988.

[14] P.Rosenstiehl and R.E.Tarjan, Rectilinear planar layout of planar graphs and bipolar orientations, *Discr. Comp. Geometry* **1**, 1986, 343–353.
[15] H.Saran and V.V.Vazirani, Finding k–cuts within twice the optimal, *Proc. of 32th Symposion on Foundations of Comp. Sci.*, 1991, 743–750.
[16] K.J.Supowit, Topics in Computational Geometry, Ph. D. Thesis, 1981, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Report UIUCDCS-R-81-1062.

Appendix A. Proof of Theorem 4

PROOF. Let $C = (C_1, C_2)$ be an arbitrary balanced clustering of P. If C_1 and C_2 cannot be separated by a line we assume that the segment with the smallest number of points of the same cluster is a subset of C_1 and denote it by C_1^X. Adding C_1^X to C_2 and replacing it by points of C_2 adjacent to C_1 obviously decreases the number of intersections between the convex hulls. It remains to be shown that thereby the sum of circumferences is not increased.

We will use an angle representation of the chords and some trigonometric equalities. The parts of the clusters are represented by their angels $a, b, \ldots, k$ as shown in Figure 5. W.l.o.g. we assume that $a + b \geq h + i$ and that the radius of the circle $r = 1/2$.

There are two possible situations for intersecting clusters:

Case I $\text{int}(\text{conv } C_1 \setminus C_1^X) \cap \text{int}(\text{conv } C_2) = \emptyset$:

In this case sectors j and k contain only points from C_1. In order to show that a crossingfree clustering has a smaller sum of circumferences we use $|\text{chord } (\alpha)| = 2r \sin \frac{\alpha}{2}$ and have to prove

$$\begin{aligned}&\sin 1/2(a+b+c) + \sin 1/2(e+\ldots+i) + \sin 1/2(c+d+e) + \\ &\sin 1/2(g) + \sin 1/2(a+i+j+k) \geq \sin 1/2(a+\ldots+g) \\ &+ \sin 1/2(a+g+\ldots+k) + \sin 1/2(c) + \sin 1/2(e) + \sin 1/2(i).\end{aligned}$$

Inserting the terms $\sin 1/2(-d+h+i), \sin 1/2(-d)$ and $\sin 1/2(-h)$ and using the basic formulas for the summation and subtraction of the sine and cosine function ($\sin x + \sin y = 2 \sin \frac{x+y}{2} \cos \frac{x-y}{2}, \cos x - \cos y = -2 \sin \frac{x+y}{2} \sin \frac{x-y}{2}$) we get

$$\begin{aligned}&\sin 1/2(a+b+c) + \sin 1/2(e+\ldots+i) - \sin 1/2(a+\ldots+g) \\ &- \sin 1/2(-d+h+i) \\ &+ \sin 1/2(c+d+e) + \sin 1/2(-d) - \sin 1/2(c) - \sin 1/2(e) \\ &+ \sin 1/2(a+i+j+k) + \sin 1/2(g) - \sin 1/2(a+g+\ldots+k) - \sin 1/2(-h) \\ &+ \sin 1/2(-d+h+i) + \sin 1/2(-h) - \sin 1/2(i) - \sin 1/2(-d) = \\ &2 \sin 1/4(a+b+c+e+\ldots+i)[\cos 1/4(a+b+c-e\ldots-i) \\ &- \cos 1/4(a+b+c+2d+e+f+g-h-i)] \\ &+2 \sin 1/4(c+e)[\cos 1/4(c+2d+e) - \cos 1/4(c-e)] \\ &+2 \sin 1/4(a+g+i+j+k)[\cos 1/4(a-g+i+j+k) \\ &- \cos 1/4(a+g+2h+i+j+k)] \\ &+2 \sin 1/4(-d+i)[\cos 1/4(-d+2h+i) - \cos 1/4(d+i)] =\end{aligned}$$

Fig. 5: Intersecting clustering (Case II): C_1^X is exchanged with sector h.

$$
\begin{aligned}
&4\sin 1/4(a+b+\underline{c+e}+\ldots+i)\sin 1/4(a+b+\underline{c+d}-h-i)\\
&\sin 1/4(\underline{d+e}+f+g)\\
&+4\sin 1/4(a+g+\underline{i}+j+k)\sin 1/4(a+\underline{h+i}+j+k)\sin 1/4(g+\underline{h})\\
&-4\sin 1/4(c+e)\sin 1/4(c+d)\sin 1/4(d+e)\\
&-4\sin 1/4(-d+i)\sin 1/4(h+i)\sin 1/4(-d+h)>0,
\end{aligned}
$$

because the sum of all angels is 2π and the sinus function is increasing between 0 and $\pi/2$. The underlined expressions in the last sum mark the angels, which are compared with the negative terms.

Case II $\operatorname{int}(\operatorname{conv} C_1 \setminus C_1^X) \cap \operatorname{int}(\operatorname{conv} C_2) \neq \emptyset$:

In this case we have to show

$$
\begin{aligned}
&\sin 1/2(a+b+c)+\sin 1/2(e+\ldots+i)+\sin 1/2(c+d+e)+\\
&\sin 1/2(g)+\sin 1/2(i+j)\geq \sin 1/2(a+\ldots+g)+\sin 1/2(g+\ldots+j)\\
&+\sin 1/2(c)+\sin 1/2(e)+\sin 1/2(i).
\end{aligned}
$$

The same procedure as in Case I can be applied. We set $k := 0$ in the proof and note that the angle a makes no difference in the final comparison.

This exchange can be performed iteratively until C_1 and C_2 are separable by a line. □