Simplicial Depth for Multiple Query Points

Luis Barba[1], Stefan Lochau[2], Alexander Pilz*[3], and Patrick Schnider[4]

1 **Department of Computer Science, ETH Zürich, Switzerland**
`luis.barba@inf.ethz.ch`
2 **Department of Mathematics, ETH Zürich, Switzerland**
`stefan.andrea.lochau@alumni.ethz.ch`
3 **Institute of Software Technology, Graz University of Technology, Austria**
`apilz@ist.tugraz.at`
4 **Department of Computer Science, ETH Zürich, Switzerland**
`patrick.schnider@inf.ethz.ch`

Abstract

We consider the following generalization of *simplicial depth*: for a set of n data points P and a set of k query points Q, the simplicial depth of Q with respect to P is the number of simplices spanned by P that contain at least one point of Q. We study this generalization for point sets in the plane. For two query points we give bounds on the maximal simplicial depth, as well as an $O(n \log(n))$ time algorithm to compute the simplicial depth. For a general number of query points we prove a bound on on the maximal simplicial depth if the data point set is in convex position. Finally, we give an $O((n+k)^{7/3}\text{polylog}(n+k))$ time algorithm to compute the simplicial depth of arbitrary query point sets with respect to arbitrary data point sets.

1 Introduction

Suppose we are given a set of n data points in $\mathbb{R}^d$, which we would like to represent with just a few points. For just one representative in $\mathbb{R}^1$, this could be the median. One way to view the median is as the "deepest" point in the set of data points. Given a set $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^1$ of n reals, it is quite intuitive to formalize a notion of "depth" w.r.t. P (the data points): for a given $q \in \mathbb{R}$, we merely count how many points of P are on each side of q and take the minimum of these two numbers. Then finding the median equals finding a point of maximal depth. Defining higher-dimensional medians requires to generalize not only the median itself, but the entire notion of depth. Several such depth measures have been introduced over time, most famously Tukey depth [21] (also called halfspace depth), simplicial depth [18], or convex hull peeling depth [5]; see, e.g., the survey by Aloupis [3]. Here, we consider simplicial depth:

▶ **Definition 1.1** (Simplicial depth)**.** For a finite point set $P \subset \mathbb{R}^d$ and a query point q, the *simplicial depth* $\sigma_P(q)$ is the number of open simplices with $d+1$ vertices in P that contain q.

The definition is attributed to Liu [18][1]; however, special cases were also addressed prior to her article (e.g., already in 1955 by Kárteszi [14]). In $\mathbb{R}^2$, Boros and Füredi [7] showed that for any set P of size n in general position there exists a point q with $\sigma_P(q) \geq n^3/27 + O(n^2)$, and there are sets where every point has simplicial depth of at most $n^3/27 + n^2$. Clearly, for n points in general position, not all triangles can hit a single point. The simplicial depth cannot be more than $\binom{\lfloor \frac{n+2}{2} \rfloor}{3} + \binom{\lceil \frac{n+2}{2} \rceil}{3} = \frac{n^3}{24} - \frac{n}{6}$, and there are sets that allow for

* Supported by a Schrödinger fellowship of the Austrian Science Fund (FWF): J-3847-N35.
[1] While in [18] the simplices are closed, we follow, e.g., [7, 10, 13, 16] and consider them open. Still, this will not make a significant difference herein, as we usually require the sets to be in general position.

35th European Workshop on Computational Geometry, Utrecht, The Netherlands, March 19–20, 2019.
This is an extended abstract of a presentation given at EuroCG'19. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear eventually in more final form at a conference with formal proceedings and/or in a journal.

such a depth [7]. From an algorithmic point of view, Gil, Steiger, and Wigderson [13] and, independently, Khuller and Mitchell [17] showed that in $\mathbb{R}^2$ the simplicial depth of a query point can be computed in $O(n\log(n))$ time and that all simplicial depths of the points in P can be found in $O(n^2)$ time. For the simplicial depth of a query point in $\mathbb{R}^3$ and $\mathbb{R}^4$, Cheng and Ouyang [9] found $O(n^2)$ and $O(n^4)$ time algorithms, respectively. Improving on a previous general $O(n^d\log(n))$ time bound by Afshani, Sheehy, and Stein [1], Pilz, Welzl and Wettstein [20] gave an $O(n^{d-1})$ time algorithm for all $d \geq 3$. The problem is #P-complete and W[1]-hard if the dimension is part of the input [1]. The best known algorithm for finding a point with maximal simplicial depth in a given set in $\mathbb{R}^2$ takes $O(n^4)$ time [4].

In this work, we consider the use of multiple query points $q_1, \ldots, q_k$, instead of just one. The idea is to find a higher-dimensional analogue to quantiles, which further describe samples in $\mathbb{R}^1$. So we extend the definition by replacing the query point q by any of the points $q_i,\ i \in \{1, \ldots, k\}$. That is, we count the simplices containing at least one of the query points.

▶ **Definition 1.2** (Simplicial depth of multiple query points)**.** Let $P \subset \mathbb{R}^d$ and $Q \subset \mathbb{R}^d$ be two finite point sets. Then the *simplicial depth* of Q with respect to P is $\sigma_P(Q)$, the number of open simplices with $d+1$ vertices in P that contain at least one point $q \in Q$. A simplex that contains at least one of the query points Q *hits* Q.

Indeed, the two query points that maximize the simplicial depth for a one-dimensional data set are the 1/3 and 2/3-quantiles, which accounts for this way of generalization. The idea of generalizing quantiles by generalizing depth measures to several query points has already been considered for the Tukey depth [19].

The problem of stabbing triangles spanned by a point set was studied by Katchalski and Meir [16], as well as Czyzowicz, Kranakis and Urrutia [10], who independently proved that, for an n-point set P with h extreme points, $2n-2-h$ many points are sufficient and necessary to stab every triangle spanned by P (i.e., to have simplicial depth of $\binom{n}{3}$).

2 Two query points

In this section we focus on finite point sets P and two query points q_1, q_2 in the plane. We assume that all points in $P \cup Q$ are in general position, i.e., no three points are collinear.

2.1 Computing the depth of two query points

We argue that computing the simplicial depth of two query points is also in $O(n\log n)$. W.l.o.g., assume that the query points lie on the x-axis and that q_1 has a smaller x-coordinate than q_2. Rather than computing the simplicial depth directly, we consider all $\binom{n}{3}$ simplices and subtract the number of those which do *not* hit the query points.

We partition $P = U \dot\cup L$, where U and L are the points above and below the x-axis, respectively. We thus have $\binom{|U|}{3}$ triangles above and $\binom{|L|}{3}$ below the x-axis. The triangles intersecting the x-axis have one vertex on one side and two on the other side. For each point $p \in L$, let $s_{\mathrm{R}}(p)$ be the number of points in U to the right of the line pq_2, let $s_{\mathrm{L}}(p)$ be the number of points in U the left of pq_1, and let $s_{\mathrm{B}}(p) = |U| - s_{\mathrm{L}}(p) - s_{\mathrm{R}}(p)$. For a point $p \in U$ the functions are defined analogously with the roles of L and U swapped. Thus, we get

$$\sigma_P(q_1,q_2) = \binom{|P|}{3} - \binom{|U|}{3} - \binom{|L|}{3} - \sum_{p\in P}\left[\binom{s_{\mathrm{R}}(p)}{2} + \binom{s_{\mathrm{B}}(p)}{2} + \binom{s_{\mathrm{L}}(p)}{2}\right] . \qquad (1)$$

It remains to compute the values of s_{L} and s_{R} efficiently. To this end, we first sort the points radially around q_1 and q_2 in $O(n\log n)$ time. Then, for each point $p \in U$, we can

count the points of L to the left of q_1p in overall $O(n)$ time, by considering the points in clockwise order around q_1 and maintaining this number (i.e., starting with $|L|$, decreasing the number when reaching a point of L, and storing it when reaching a point of U). We thus obtain a table for s_{R} for all points in U, and do the analogous for the remaining values.

2.2 Bounds for two query points

We provide an upper bound for the simplicial depth of two query points. Let $m_1 = |L|$ and $m_2 = |U|$, i.e., $m_1 + m_2 = n$. First, consider a fixed point $l \in L$ and all triangles it forms with two points of U. For such a triangle lu_1u_2, we define C as the closed cone formed by the rays $\overrightarrow{lu_1}$ and $\overrightarrow{lu_2}$, and call $d := |C \cap U| - 1$ the *span* of the triangle. Then d is one plus the number of points strictly between u_1 and u_2 in the radial ordering around l. We now count how many of these triangles with a certain span d hit $\{q_1, q_2\}$.

We note that, for a fixed l, at most d triangles with span d can hit a query point, so at most $2d$ of these can be hitting. On the other hand, for each d there are at most $m_2 - d$ many triangles, hitting or not. We can do this for every lower and upper point, and sum up to get an upper bound on the simplicial depth as follows:

$$\sigma_P(q_1, q_2) \leq \sum_{l \in L} \sum_{d=1}^{m_2-1} \min\{2d, m_2 - d\} + \sum_{u \in U} \sum_{d=1}^{m_1-1} \min\{2d, m_1 - d\}$$

This implies a simpler bound of $\sigma_P(q_1, q_2) \leq \frac{1}{3} m_1 m_2 (m_1 + m_2 + 4)$, which is maximized for $m_1 = m_2 = n/2$ (see the full version), where we get the following:

▶ **Theorem 2.1.** *Let $P \subseteq \mathbb{R}^2$ be a set of n data points and q_1, q_2 be two query points, all in general position. Then*

$$\sigma_P(q_1, q_2) \leq n^3/12 + n^2/3.$$

There are indeed point sets in convex position that allow for a similar simplicial depth: consider P as the vertex set of a regular n-gon for $n = 2m$ (see Figure 1 for an illustration). Place both q_1 and q_2 on the intersection c of the long diagonals and move them slightly outwards such that they do not lie on any line through two points of P but such that the line through them contains c. Then $\{q_1, q_2\}$ has simplicial depth $\sigma_P(q_1, q_2) = \frac{m^3}{3} + m^2 - \frac{4m}{3}$. With $n = 2m$, this translates to $\frac{n^3}{24} + \frac{n^2}{4} - \frac{2n}{3}$, which is roughly half of the bound in Theorem 2.1. Further, recall that the maximal simplicial depth of a single point is $\frac{n^3}{24} - \frac{n}{6}$. Comparing to this, we only improve by addition of a quadratic term. Nevertheless, for up to 14 data points in convex position, computational experiments have shown that this construction is optimal. We conjecture that this holds for all sets in convex position. Note that while for one query point it is not hard to see that the simplicial depth is maximized for data point sets in convex position, the same cannot be said for two query points. In fact, the same question can be asked for any number of query points: is the simplicial depth of k query points always maximized by data point sets in convex position?

3 More query points

3.1 Upper bound for data points in convex position

In this section we give an upper bound to the simplicial depth of any set of k query points $Q = \{q_1, \ldots, q_k\}$ in a given point set $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^2$ when P is in convex position. We assume that $P \cup Q$ is in general position and Q is in $\mathrm{conv}(P)$, the convex hull of P.

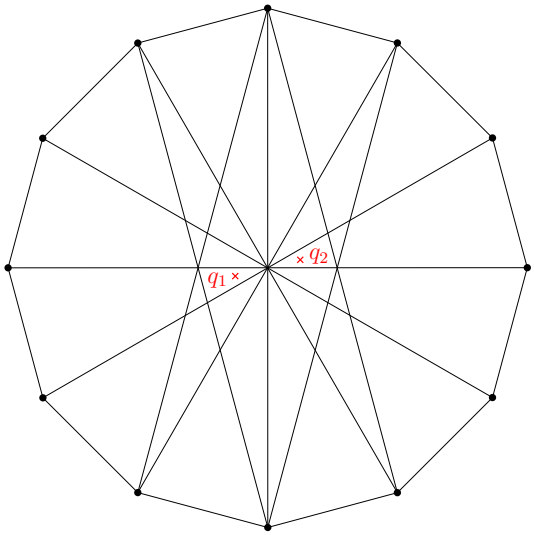

Figure 1 A point set of size $2m = 12$ with conjectured maximal simplicial depth 100 for two query points.

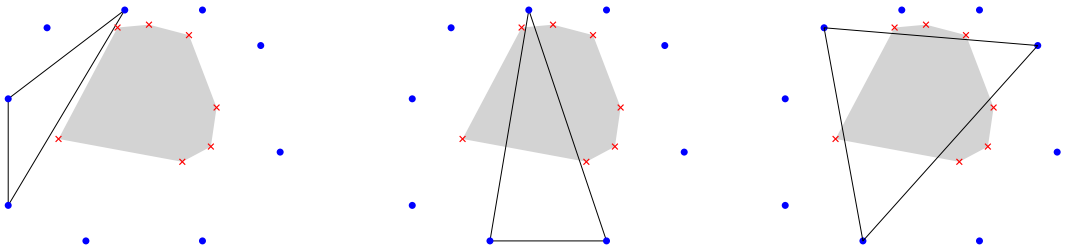

Figure 2 The different types of non-hitting triangles.

We again count the triangles not containing any points of Q. Let $\mathcal{S}$ be the set of triangles spanned by points of P. We partition $\mathcal{S}$ into those triangles that do hit Q, $\Sigma := \{S \in \mathcal{S} : S \cap Q \neq \emptyset\}$, and those which do not hit Q, $\Delta := \{S \in \mathcal{S} : S \cap Q = \emptyset\}$. Then $\sigma_P(Q) := |\Sigma| = \binom{n}{3} - |\Delta|$. We devise a lower bound on $|\Delta|$. Let $S \in \Delta$ be an arbitrary non-hitting triangle. Note that $S \setminus \text{conv}(Q)$ has at most three connected components, each of which contains at least one original vertex of S. We partition Δ into the triangles that get split into i parts by $\text{conv}(Q)$, Δ_i for $i \in \{1, 2, 3\}$. We see that (see Figure 2):

(1) If $S \setminus \text{conv}(Q)$ has only one component, then S and $\text{conv}(Q)$ are disjoint (as S does not hit Q). S has a unique pair of vertices w, w' whose supporting line separates S and Q.

(2) If $S \setminus \text{conv}(Q)$ has two components, we see that there are two vertices of S in one and a single vertex p in the other component. This vertex p is again unique.

(3) If $S \setminus \text{conv}(Q)$ has three components, we discard it and potentially worsen our bound.

Let $t_i(p)$ be the number of data points on the right side of the ray $\overrightarrow{pq_i}$, where the elements of Q are indexed according to their radial order around p. We get that:

$$\sigma_P(q_1,\dots,q_k) \le \binom{n}{3} - \sum_{p\in P}\left(\underbrace{\frac{1}{2}\binom{t_1(p)}{2} + \sum_{i=1}^{k-1}\binom{t_{i+1}(p)-t_i(p)}{2} + \frac{1}{2}\binom{n-1-t_k(p)}{2}}_{=:f_p(t_1,\dots,t_k)}\right).$$

We count each simplex that does not intersect $\mathrm{conv}(Q)$ "a half times" for w and w'. The function $f_p(t_1,\dots,f_k)$ is convex on $\{t\in\mathbb{R}^k \mid 0\le t_1\le\cdots\le t_k\le n-1\}$, so we can bound it from below separately using convex optimization techniques. (Intuitively, the value of f_p is small if the number of points between two consecutive points of Q is roughly the same; we provide a formal reasoning in the full version.) With this, we obtain an upper bound of

$$\sigma_P(q_1,\dots,q_k) \le \frac{n^3k}{6(k+3)} - \frac{3n^2}{2(k+3)} - \frac{5n}{24}.$$

For $k=2$, we can compare this to Theorem 2.1 – we go from the older $\frac{1}{12}n^3+\frac{1}{6}n^2$ bound to $\frac{1}{15}n^3-\frac{3}{10}n^2-\frac{5}{24}n$ and improve asymptotically by a factor of $\frac{5}{4}$. (But recall that the new bound is for P in convex position only.) Comparing thisto $\binom{n}{3}$ we get the following theorem.

▶ **Theorem 3.1.** *Let $P\subseteq\mathbb{R}^2$ be n points in convex position, and let $q_1,\dots,q_k\in\mathbb{R}^2$ be k query points. Then at most a fraction of $1-\frac{3}{k+3}+O\left(\frac{1}{n}\right)$ of the simplices with vertices in P can contain any of the k query points.*

3.2 Algorithmic aspects

▶ **Theorem 3.2.** *The simplicial depth of a set $Q\subset\mathbb{R}^2$ w.r.t. a set $P\subset\mathbb{R}^2$, all in general position, with $N=|P|+|Q|$ can be computed in $O(N^{7/3}\mathrm{polylog}(N))$ time.*

Proof. We use an approach similar to one of computing the number of empty triangles in a point set (see [12]). For a point $p\in P$, we define a simple polygon R_p that contains every triangle spanned by p and two other points of P not hitting Q. Let B be a bounding box of P. Shoot a ray from every point $q\in Q$ in the opposite direction of p until hitting B and add two edges for R_p in B starting at q with a small angle separated by the ray. See Figure 3. Now two points of $P\setminus\{p\}$ see each other in R_p iff they form a triangle with p not hitting Q.

Ben-Moshe et al. [6] construct the visibility graph of points inside a simple polygon. While enumerating the edges of this graph is too costly here, their method can be adapted to count them. They argue that edges of the visibility graph that cross an edge e separating the polygon (that is not necessarily a diagonal of the polygon) correspond to bichromatic crossings in an arrangement of red and blue segments: A point and the part on e which the point sees defines a (possibly empty) wedge, whose dual is a line segment; two points define an edge of the visibility graph crossing e iff their dual segments intersect; the red segments correspond to points in one sub-polygon defined by e, and the blue segments to points in the other sub-polygon. These segments can be given in $O(N\log(N))$ time using standard machinery for polygon visibility and point-line duality. Agarwal [2, Theorem 6.1] shows how to count bichromatic crossings of n red and blue segments in $O(n^{4/3}\mathrm{polylog}(n))$ time.

As in [6, Section 2.1], we divide R_p into two parts, each containing at least a constant fraction of the union of points and R_p's vertices (e.g., 1/3 is doable by adapting an approach from [8] not even using that R_p is star-shaped). Then, we recursively count the edges of the visibility graph in the two sub-polygons and add the number of edges crossing the diagonal using Agarwal's algorithm [2, Sect. 6]. As this is the dominating task in each iteration, we can use induction on the number of points to show that this algorithm requires $O(N^{4/3}\mathrm{polylog}(N))$ time for computing the number of triangles with a point p not hitting Q. ◀

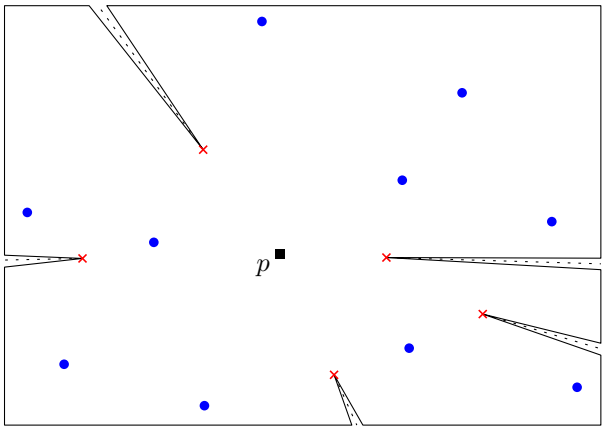

Figure 3 The polygon R_p of a point $p \in P$. Any two points of $P \setminus \{p\}$ (blue dots) that see each other correspond to a triangle with p not containing a point of Q (red crosses).

We do not know about the complexity of finding a set Q, $|Q| = k$, with maximal simplicial depth for a given integer k and data set P. Using a straight-forward reduction from monotone planar 3-SAT [11], we show in the full version that extending a given set to have a point in each triangle of P is NP-hard.

References

1 Peyman Afshani, Donald R. Sheehy, and Yannik Stein. Approximating the simplicial depth. *CoRR*, abs/1512.04856, 2015. URL: `http://arxiv.org/abs/1512.04856`.

2 Pankaj K. Agarwal. Parititoning arrangements of lines II: applications. *Discrete & Computational Geometry*, 5:533–573, 1990. `doi:10.1007/BF02187809`.

3 Greg Aloupis. Geometric measures of data depth. In Regina Y. Liu, Robert Serfling, and Diane L. Souvaine, editors, *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, pages 147–158. DIMACS/AMS, 2003.

4 Greg Aloupis, Stefan Langerman, Michael A. Soss, and Godfried T. Toussaint. Algorithms for bivariate medians and a Fermat-Torricelli problem for lines. *Comput. Geom.*, 26(1):69–79, 2003. `doi:10.1016/S0925-7721(02)00173-6`.

5 Vic Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, 139(3):318–355, 1976. URL: `http://www.jstor.org/stable/2344839`.

6 Boaz Ben-Moshe, Olaf A. Hall-Holt, Matthew J. Katz, and Joseph S. B. Mitchell. Computing the visibility graph of points within a polygon. In Jack Snoeyink and Jean-Daniel Boissonnat, editors, *Proc. 20th ACM Symposium on Computational Geometry (SoCG 2004)*, pages 27–35. ACM, 2004. `doi:10.1145/997817.997825`.

7 E. Boros and Z. Füredi. The number of triangles covering the center of an n-set. *Geometriae Dedicata*, 17(1):69–77, Oct 1984. `doi:10.1007/BF00181519`.

8 Bernard Chazelle. A theorem on polygon cutting with applications. In *23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982*, pages 339–349. IEEE Computer Society, 1982. `doi:10.1109/SFCS.1982.58`.

9 Andrew Cheng and Ming Ouyang. On algorithms for simplicial depth. In *Proc. 13th Canadian Conference of Computational Geometry (CCCG 2001)*, pages 53–56, 2001.

10 Jurek Czyzowicz, Evangelos Kranakis, and Jorge Urrutia. Guarding the convex subsets of a point-set. In *Proc. 12th Canadian Conference of Computational Geometry (CCCG 2000)*, pages 47–50, 2000.

11 Mark de Berg and Amirali Khosravi. Optimal binary space partitions for segments in the plane. *Int. J. Comput. Geometry Appl.*, 22(3):187–206, 2012.

12 David P. Dobkin, Herbert Edelsbrunner, and Mark H. Overmars. Searching for empty convex polygons. *Algorithmica*, 5(4):561–571, 1990. doi:10.1007/BF01840404.

13 Joseph Gil, William Steiger, and Avi Wigderson. Geometric medians. *Discrete Mathematics*, 108(1):37 – 51, 1992. doi:10.1016/0012-365X(92)90658-3.

14 Franz Kárteszi. Extremalaufgaben über endliche Punktsysteme. *Publ. Math. Debrecen*, 4:16–27, 1955.

15 William Karush. Minima of functions of several variables with inequalities as side constraint. *M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois*, 1939. URL: https://dissexpress.proquest.com/dxweb/results.html?QryTxt=&By=&Title=&pubnum=TM11033.

16 Meir Katchalski and Amram Meir. On empty triangles determined by points in the plane. *Acta Mathematica Hungarica*, 51(3-4):323–328, 1988.

17 Samir Khuller and Joseph S.B. Mitchell. On a triangle counting problem. *Information Processing Letters*, 33(6):319 – 321, 1990. doi:10.1016/0020-0190(90)90217-L.

18 Regina Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18(1):405 – 414, 03 1990. doi:10.1214/aos/1176347507.

19 Alexander Pilz and Patrick Schnider. Extending the centerpoint theorem to multiple points. In Wen-Lian Hsu, Der-Tsai Lee, and Chung-Shou Liao, editors, *29th International Symposium on Algorithms and Computation (ISAAC 2018)*, volume 123 of *LIPIcs*, pages 53:1–53:13. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2018. URL: http://www.dagstuhl.de/dagpub/978-3-95977-094-1, doi:10.4230/LIPIcs.ISAAC.2018.53.

20 Alexander Pilz, Emo Welzl, and Manuel Wettstein. From Crossing-Free Graphs on Wheel Sets to Embracing Simplices and Polytopes with Few Vertices. In Boris Aronov and Matthew J. Katz, editors, *33rd International Symposium on Computational Geometry (SoCG 2017)*, volume 77 of *LIPIcs*, pages 54:1–54:16. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.SoCG.2017.54.

21 John W. Tukey. Mathematics and picturing data. *Proc. Int. Congr. Mathematics, Vancouver*, 2:523 – 531, 1975.