

# Path Length in Proximity Graphs as a Data Depth Measure

Kathryn Seyboth <sup>\*†</sup>

Eynat Rafalin<sup>\*</sup>

Diane Souvaine<sup>\* ‡</sup>

## 1 Introduction

The statistical analysis method of *data depth* is a center-outward ordering of points based on their centrality in a data set. Generally accepted depth measures such as halfspace depth [6] assume unimodality of the data set, finding a single center even in the case of several well-spaced clusters of points.

This paper defines a proximity graph depth measure and its contours, presents a notion of set approximation that is sensitive to the multimodality of a data set, and defines a new proximity graph with linear-time dependence on dimension that performs consistently under our depth measure and can be manipulated to achieve different results.

## 2 Proximity Graph Depth

**Definition:** The [proximity graph] depth of a point  $p$  relative to a set  $S = \{p_1 \dots p_n\}$  is the minimum number of edges in the [proximity graph] of  $S \cup p$  traversed from  $p$  to reach the convex hull of  $S \cup p$ <sup>1</sup>.

We coded algorithms for computation of depths and contours in the Delaunay Triangulation (DT) [1], Gabriel Graph (GG) [2], and our newly-defined Lune Graph (Section 5) under this measure<sup>2</sup>.

In  $R^2$ , the overall time complexity of depth assignment is  $O(n \log n)$  for the DT and  $O(n^2)$  for the GG. For higher dimensions,  $O(n^{\lceil \frac{d}{2} \rceil})$  time is required for the DT and the GG has a complexity of  $O(dn^3)$ .

The DT and GG exhibit sensitivity to pockets of high density in point sets. This is a useful feature for data sets with several clusters, and several local maxima, each corresponding to a set of points likely to be a part of the same cluster.

Unfortunately, distributions that are too close behave as a single mode, and those separated by large empty regions appear unimodal, as the dearth of points between the clusters prevents paths from travelling inward quickly from the convex hull.

<sup>\*</sup>Department of Computer Science, Tufts University, Medford, MA 02155 partially supported by NSF grant CCF-0431027, kseyboth, erafalin, dls@cs.tufts.edu. For full paper, see [4].

<sup>†</sup>Partially supported by Tufts Provost's Summer Scholars and CRA Distributed Mentor Program.

<sup>‡</sup>2005-6 Radcliffe Inst. Fellow and MIT Visiting Scientist.

<sup>1</sup>Green & Sibson suggested using path length to the convex hull along DT edges as a depth measure [3], but included only DT and did not include analysis or experimental results.

<sup>2</sup>The Relative Neighbor Graph [5] demonstrates inconsistent performance under this depth function and we omit it.

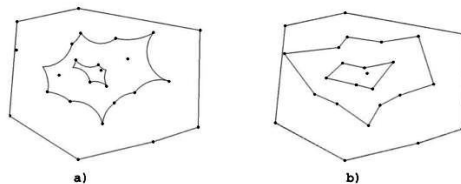


Figure 1: Depth contours and simplified depth contours of the DT for a set of 25 points

## 3 Depth Contours

Depth Contours [6], nested regions of increasing depth, serve as a topological map of the data. The  $j$ th depth contour consists of all points in space of depth  $\geq j$ .

### Delaunay Depth

The depth of a point in the DT measure is its depth upon its hypothetical addition to the set. Each triangle in the DT corresponds to an empty circle circumscribing it. A point added to a triangulated set will be connected to all defining points of the circles in which it lies.

No DT circle with a defining point of depth  $< j - 1$  contains any point of depth  $\geq j$ . Therefore the borders of the contours, those places where the depth of points change, are defined by arcs of the Delaunay circles (Fig. 1a). In a multimodal set some contours are not continuous: each local maximum may have its own section of the  $j$ th contour.

The boundaries of the DT contours can be simplified using straight lines rather than arcs (Fig. 1b). The  $j$ th Delaunay simplified contour of  $S$  is the area inside the CH of  $S$  which is not inside any Delaunay triangle with a defining point of depth  $< j - 1$ .

### Gabriel Depth

A segment  $p_1 p_2$  in the GG of set  $S$  with  $p_1, p_2 \in S$  is defined by an empty circle for which  $p_1 p_2$  is the diameter. A point  $p$  added to  $S$  connects to points  $q \in S$  for which the circle of diameter  $pq$  is empty.

Every  $q \in S$  has an associated convex *region of influence* containing those points which, if added to  $S$ , would be connected to  $q$ : the intersection of half-planes containing  $q$  that are delimited by a line through some  $r \in S$  and perpendicular to the segment  $qr$ . The  $j$ th GG depth contour of  $S$  is the area inside the convex hull of  $S$  but not in the region of influence of any point of depth  $< j - 1$ .

## 4 Medians and Seeds

**Definition:** The **median** of a set  $S$  under some depth measure  $D : S \rightarrow \mathbb{R}$  is the set of points  $M$  such that  $\forall p \in M, D(p) \geq D(q) \forall q \in S$ .

The median can be viewed as the center of the set. The use of a group of points as the median assumes the unimodality that is common in depth measures.

The *mode* of a set is the most common value obtained in a set of observations [7]. We use the term mode flexibly and refer to a *bimodal* (or *multimodal*) point set as one having two (or more) local maxima.

Medians are not the best estimators in multimodal situations; points not of maximum depth may be local maxima. Estimating a set using a simple median can ignore many local maxima that represent several modes. Instead, we define *seeds*.

**Definition:** A **seed** of a set  $S$  under some depth measure  $D : S \rightarrow \mathbb{R}$  is a connected set of points  $T \subset S$  st  $\forall p, q \in T, D(p) = D(q)$  and  $\forall r \in S, r \notin T$  adjacent to some  $u \in T, D(r) < D(u)$ .

Finding seeds requires a recursive search to locate all points of the seed and all points in  $S$  that are connected to that seed, requiring only linear running time in the size of the proximity graph.

### 4.1 Improvements

Proximity graph measures may be overly sensitive to density differences and create more seeds than desired. Two strategies combat erroneous seeds:

**Significant Maxima:** Only seeds that are local maxima and whose neighbors would be maxima were points of the seed not there are included. Essentially, the seed is a local maxima by 2 depths rather than one<sup>3</sup>.

If the number of seeds required is known or limited, it is possible to change this definition to require larger and larger peaks: e.g. adding the requirement that the removal of the seed and the second depth would leave the third group as a seed.

**Convex-Hull Path Origins:** Another option is to count the number of convex hull breadth-first search roots that could have initiated paths to the seeds. If many points could have created the path that labelled a point, the seed is relatively central compared to a seed for which there were only one or two path origins.

Once each seed has a value, it is possible to weed out weak seeds, e.g. by using only those with value over half of the maximum.

## 5 Lune Graph

The GG tends to find more seeds than desirable, but DT computation relies exponentially on dimension, making it virtually useless for high-dimensional data sets. A graph denser than the GG but with the same linear dependence on dimension is needed. We define a new graph that

<sup>3</sup>This fix can eliminate seeds when there are two seeds at the center of a mode which are separated by only one point. Neither seed in the double peak is significant, so that mode is eliminated from notice. A simple check can fix this problem.

combines the speed of the GG with a performance similar to the DT.

**Definition:** The  **$k$ -lune** of points  $p$  and  $q$  in  $R^d$  is the intersection of the  $d$ -spheres passing through  $p$  and  $q$  and centered on the perpendicular bisector of the segment  $pq$  at a distance from  $pq$  of  $\frac{|pq|}{k}$ .

**Definition:** Two points  $p$  and  $q$  of a set  $S$  are connected in the  **$k$ -Lune Graph** of  $S$  iff the  $k$ -lune( $p, q$ ) contains no point  $r \in S$ .

Because a smaller volume must be empty to form a segment in the Lune graph than in the GG, there are more segments and it is denser. The higher  $k$  is, the larger the  $k$ -lunes are, and the sparser the graph. The GG is the  $\infty$ -Lune graph, and as  $k \rightarrow 0$ , the  $k$ -Lune graph approaches the complete graph.

The  $k$ -Lune graph can be computed in  $O(dn^3)$  time by checking all points for containment in the  $k$ -lune of all  $O(n^2)$  pairs of points.

The performance of the  $k$ -Lune graph varies according to the value of  $k$ : a high value yields a broad range of depth with more seeds; a lower value produces only a few local maxima.

Contours for the Lune graph are defined similarly to those of the GG, based on a points' regions of influence (Section 3). The  $j$ th  $k$ -Lune depth contour of  $S$  is the area inside the CH of  $S$  which is not in the region of influence of any point of depth  $< j - 1$ .

## 6 Conclusion

Depth measures based on path length in proximity graphs are sensitive to the possibility of multiple modes in the data; the depth function can correctly locate several centers of the data set. The GG and Lune graph depths are particularly useful because their complexity depends linearly on dimension. These depth measures may contribute to improving clustering algorithms.

## References

- [1] S. Fortune. Voronoi diagrams and Delaunay triangulations. In *Handbook of disc. and comp. geom.*, pages 377–388. CRC Press, 1997.
- [2] K.R. Gabriel and R.R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18:259–278, 1969.
- [3] P. J. Green and R. Sibson. Computing dirichlet tessellations in the plane. *The Computer Journal*, 21(2):168–173, 1978.
- [4] K. Seyboth, E. Rafalin, and D. L. Souvaine. Path length in proximity graphs as a data depth measure. TUFTS-CS technical report tr-2005-5, Tufts University, 2005.
- [5] G.T. Toussaint. The relative neighborhood graph of a finite planar set. *Pat. Recog.*, 12:261–268, 1980.
- [6] J.W. Tukey. Mathematics and the picturing of data. In *Proc. Int. Cong. of Math.*, pages 523–531, 1974.
- [7] Erik W. Weisstein. Mode. From MathWorld, <http://mathworld.wolfram.com/Mode.html>.