Towards Robust Topology of Sparsely Sampled Data

Carlos D. Correa, Member, IEEE, and Peter Lindstrom, Member, IEEE

Abstract—Sparse, irregular sampling is becoming a necessity for reconstructing large and high-dimensional signals. However, the analysis of this type of data remains a challenge. One issue is the robust selection of neighborhoods—a crucial part of analytic tools such as topological decomposition, clustering and gradient estimation. When extracting the topology of sparsely sampled data, common neighborhood strategies such as $k$-nearest neighbors may lead to inaccurate results, either due to missing neighborhood connections, which introduce false extrema, or due to spurious connections, which conceal true extrema. Other neighborhoods, such as the Delaunay triangulation, are costly to compute and store even in relatively low dimensions. In this paper, we address these issues. We present two new types of neighborhood graphs: a variation on and a generalization of empty region graphs, which considerably improve the robustness of neighborhood-based analysis tools, such as topological decomposition. Our findings suggest that these neighborhood graphs lead to more accurate topological representations of low- and high-dimensional data sets at relatively low cost, both in terms of storage and computation time. We describe the implications of our work in the analysis and visualization of scalar functions, and provide general strategies for computing and applying our neighborhood graphs towards robust data analysis.

Index Terms—Neighborhood graphs, topology, sparsely sampled data.

1 INTRODUCTION

With the increasing rate of acquisition and simulation capabilities, sparse sampling of data sets is becoming a necessity for reconstructing complex and possibly high-dimensional signals. In scientific simulations, sparse and irregularly distributed samples are required to reconstruct three-dimensional scalar or vector fields in certain regions of interest without the need to store and process large regular grids. In other cases, such as the exploration of high-dimensional functions generated in uncertainty quantification, sampling on regular grids is prohibitive. The need for sparse sampling may also arise from the acquisition process, such as in LIDAR scanning, or from simulation based on Monte Carlo or particle-in-cell methods, e.g., in the study of magnetic fields in a tokamak.

Many techniques have been devoted to the analysis and rendering of data on regular grids, and currently there are efforts to find efficient counterparts for irregular and sparse data in low dimensions. However, the lack of an underlying structure for the samples renders these techniques inapplicable or difficult to scale to higher dimensions. One example is the computation of neighborhoods for sparsely sampled data in arbitrary dimensions. The neighborhood of a sample plays a crucial role in analyzing and visualizing scalar fields, as demonstrated by recent data exploration approaches \cite{25,39}, which show applications of clustering and regression in high-dimensional spaces.

Common to these approaches is the generation of a topological representation of the data. Topological representations, such as the contour tree \cite{10} and the Morse-Smale complex \cite{19}, are valuable aids in understanding scalar functions. They help describe a function in terms of connected components and segment the data into regions of uniform level set or gradient behavior. Moreover, representations like the contour tree act as visual structures that can be drawn in the 2D plane without occlusion problems.

We address the problem of finding good sample neighborhoods that aid in extracting the underlying topology of the sparsely sampled data. The problem of finding a good neighborhood can be formulated as follows: given a scalar function $f$, a finite sample set $V \subset \mathbb{R}^d$, and a sample $x \in V$, find the set of neighboring points in $V$ that best describe the behavior of $f$ near $x$ to allow an accurate classification of $x$ as an extremum, a saddle point, or a regular point. For instance, a sample is classified as a maximum if its function value exceeds those of its neighbors. Intuitively, to ensure an accurate classification, we wish for each sample to have neighbors in roughly $2d$ “sufficiently different” directions (as would be the case on a regular grid), so that the positive
and negative gradient directions are represented by nearby samples.

To understand the importance of choosing a good neighborhood, consider Fig. 1, where we show a 2D terrain, sampled irregularly using 700 (top) and 4,000 (bottom) random points. A common choice for extracting the critical points of this terrain is to create a neighborhood graph consisting of the \( k \) nearest neighbors \((k\text{NN})\) to each point. In Fig. 1(b), we see the neighborhood graph for \( k = 6 \) neighbors per sample point, reflecting the expected number of neighbors in a triangulation of the samples. Maxima of this function are shown as red or green spheres. Green points correspond to the ground truth maxima of the analytical function that describes the terrain. Red points are false maxima that appear due to poor neighbor connectivity, and can be regarded as topological noise. We observe that, as we increase the number of sample points, the number of false maxima increases proportionally. To counteract this problem, we may use higher quality neighborhoods, such as the Delaunay triangulation (DT) [22] or less dense subsets of the DT such as the Gabriel graph (GG) [24]. Fig. 1(c) shows that the GG may reduce the topological noise.

In this paper, we present a number of contributions in computing neighborhood graphs that lead to more accurate extraction of topological representations of sparsely sampled data at similar computational cost and graph complexity. These are:

- **The natural empty region graph**: a subset of the so called empty region graphs (ERG) that guarantees certain neighborhood and space partitioning properties useful for local analysis. ERGs are neighborhood graphs where two points are connected if a geometric region around them, called the empty region, does not contain any other point [8].

- **The relaxed empty region graph**: a variation on the empty region graphs [8] that results in topological representations that are less sensitive to the sparsity and distribution of the samples than their original counterparts, as shown in Fig. 1(e).

- **The stochastic empty region graph**: a generalization that introduces the notion of the likelihood of samples being neighbors.

To facilitate choosing an empty region with desired neighbor pruning properties, we introduce the notion of the *umbra* of an empty region. We describe the neighborhoods associated with our new graphs for samples in arbitrary dimensions, and outline general procedures to obtain them efficiently. We have studied the impact of these neighborhoods on the extraction of extrema in scalar fields and report our findings here. We discuss implications and applications of our methods in the analysis and visualization of scalar fields in arbitrary dimensions.

## 2 Related Work

### Analysis and Visualization in High Dimensions

The exploratory analysis of multi-dimensional functions demands the use of various data and visual analysis tools, including regression [17], response surface fitting [6] and generalized additive models [28]. For large-scale and high-dimensional data, these models are applied together with data reduction strategies such as clustering [3], projection and multidimensional scaling [11] to data obtained via sampling [48]. Some of the most common sampling strategies are random, Latin Hypercube [35, 47] and importance sampling [46], as well as centroidal Voronoi tessellation [18].

The visualization of high-dimensional functions, however, remains a challenge. The most common approaches include the projection of data into scatterplots in 2D or 3D subspaces [15], star coordinates [31], Chernoff faces [13], Andrews plots [1] and parallel coordinates [29]. A key challenge in visualization is finding the best projection of the data. Asimov presents the grand tour approach [2], which provides a sequence of 2D subspaces chosen for viewing. Dimension reduction techniques and manifold learning approaches are designed to extract a low-dimensional manifold embedded in a higher-dimensional space, which can be visualized more effectively in a 2D graphical display, as surveyed extensively in [21, 51].

### Topological Analysis

A series of techniques have been proposed to extract and represent the topology of scalar fields, including contour trees [53], Reeb graphs [44], and Morse-Smale complexes [19, 37]. Such representations assist analytic tools such as regression [25], feature extraction [26], classification [23, 39] and clustering [12, 38].

Efficient algorithms have been proposed for computing these structures for low-dimensional data, often confined to regular grids. Carr et al. [10] present an algorithm for computing contour trees in arbitrary dimensions. Pascucci et al. propose a robust algorithm for computing Reeb graphs on high-dimensional manifolds [42]. Harvey and Wang [27] construct 2D topological terrains of high-dimensional scalar functions using nearest neighbor graphs. Osterling et al. [40] follow a similar approach for scattered data. Gerber et al. [25] propose an approximate representation of the Morse-Smale complex for high-dimensional scalar functions. Carr and Snoeyink [9] describe a method for computing the contour tree on arbitrary graphs. In their recent paper, Osterling et al. [40] hinted at the problem of finding local extrema using point clouds and suggested the use of inexpensive alternatives to the Delaunay triangulation. In this paper, we propose more robust alternatives based on generalizing empty region graphs.

### Neighborhood Graphs

Neighborhood or proximity graphs create a geometric structure that connects two points if they are close in some sense. These graphs have been well studied and include the relative neighborhood graph [30], the Gabriel graph [24], \( \beta \)-skeletons [33], \( \sigma \)-local graphs [5], \( \Theta \)-graphs [32], \( \gamma \)-neighborhood graphs [54] and Delaunay triangulations [22]. These graphs have been well studied in terms of their geometric properties [4, 8, 14], and have been applied in geographic analysis [34], pattern recognition [50], clustering [52], machine learning [49], normal estimation [41] and the extraction of contour trees [39]. In this paper, we present the first study of the use of generalized empty region graphs towards robust topology extraction.

## 3 Background

At the core of high-dimensional data analysis is the notion of a neighborhood graph. A neighborhood graph considers data points as vertices interconnected with edges that represent some measure of similarity, such as Euclidean distance. We assume that each sample carries a scalar function value \( f(x) \). In this paper we study several neighborhood graphs and their influence on the topology of \( f \).

### 3.1 Scalar Field Topology

A scalar function can be summarized concisely in terms of its topological decompositions. For example, the contour tree describes how contours appear, merge, split and disappear as the scalar isovalue is varied [53], and segments the domain into regions of homeomorphic level set components. For a piecewise linear function over a triangulation, the contour tree computation amounts to maintaining the connected components of the vertices \( \{ x \in V : f(x) > c \} \) and edges of the neighborhood graph given by the triangulation. Carr et al. [10] referred to such a graph augmented with function values as a *height graph*, and their contour tree algorithm uses only a height graph as input.

Morse theory allows us to identify critical points and decompose a scalar function into regions of uniform gradient flow [19]. The resulting segmentation, called the Morse-Smale complex, can be approximated in any dimension without the notion of either an interpolant or gradient by considering steepest ascent and descent paths over the edges of a neighborhood graph [25].

Both of these approaches classify a sample point as a maximum if its function value exceeds those of its neighbors (and equivalently for minima). Saddle points are samples with neighbors whose upward or downward paths lead to more than one maximum or minimum.

### 3.2 \( k \text{NN} \) Graphs

Consider a set of points \( V = \{ x_1, x_2, \ldots, x_n \} \) drawn randomly from the domain \( \Omega \subseteq \mathbb{R}^d \). Let \( d(p, q) \) represent the Euclidean distance between two points. Among the most common neighborhood graphs are the \( \epsilon \)-neighborhood graph \( G_\epsilon(V, E_\epsilon) \) and the (directed) \( k \) nearest neighbor graph \( G_k(V, E_k) \). These graphs are defined by

\[
pq \in E_\epsilon \iff d(p, q) \leq \epsilon, \tag{1}
\]

and

\[
pq \in E_k \iff q \in k\text{NN}(p). \tag{2}
\]
Fig. 2. $F$ measure of extremum detection for several 2D functions using a kNN neighborhood graph. Clearly, selecting the appropriate $k$ to achieve maximum precision and recall depends on the function.

where $kNN(p)$ is the set of $k$ nearest neighbors of a point $p$.

The problem of topology extraction is complicated when dealing with sparse and irregular samplings. Consider, for example, randomly sampling the 2D function $f(x) = |x|$, which has a single minimum at the origin, and computing the minima of $f$ by inspecting the neighbors of each sample in a $kNN$ graph. For sufficiently small values of $k$, the neighborhoods are so sparse that all $k$ neighbors of a sample $x$ may by chance have a higher function value, causing $x$ to be misclassified as a minimum, even though $x$ and its neighbors may be very far from the true minimum. In fact, it is not difficult to see that the expected number of false minima increases linearly with the number of random samples $n$: A sample $x$ will be classified as a minimum if all of its neighbors appear in the "outward" tangent halfspace defined by $x$ and $Vf$. For random samplings the neighbors are uniformly distributed in both directions, and therefore the likelihood of all $k$ neighbors having a higher function value is roughly $\frac{k}{2^k}$. For $n$ random samples, the expected number of minima detected is therefore $\frac{n}{2^k}$.

One can partially solve this problem by increasing the number of neighbors $k$. For the simple radial function $f$ above, fully connecting all sample points gives the correct answer. However, the functions we are interested in may have several local extrema, and over-connecting the neighborhood graph will conceal these extrema. These two issues of under- and over-connecting neighbors can be described as lack of precision (many false positives when $k$ is small) and lack of recall (many false negatives when $k$ is large) in the detection of extrema. Here precision is measured as the ratio of correctly detected extrema to the total number of detected extrema, and recall as the ratio of the correctly detected extrema to the number of true extrema.

A way to measure the quality of extrema detection is via the $F$ measure: the harmonic mean between precision and recall. Ideally, $F = 1$, and smaller values are the combined effect of low precision and/or low recall. Fig. 2 shows how the $F$ measure (y axis) for a number of 2D functions varies with the number of neighbors $k$. The ideal $k$ is chosen as the one that maximizes the $F$ measure. We see that: (1) the ideal $k$ is different for each function; (2) in some cases, such as the Schwefel function, no $k$ yields perfect accuracy, and (3) the impact of selecting a non ideal $k$ is different for each function, e.g., the Ackley function drops the $F$ measure to almost 0 above 30 neighbors, while the Camelback function stays at $F = 0.8$ up to 90 neighbors.

Fig. 2 reveals that picking the $k$ that achieves the most accurate topology is not trivial, and depends upon many factors, including the sampling strategy and the spatial frequency of the scalar function. One potential solution to this problem is to regard the false extrema as noise and to prune them with persistence-based topological simplification [26]. However, as we show in Section 5.3, the persistence of false extrema may well overlap with that of the true extrema and noise in the scalar field. A more robust solution is to make use of better neighborhoods, such as the Delaunay triangulation and its variants.

![Graph](image)

Fig. 3. Density and computational cost of Delaunay triangulation compared to the Gabriel graph in various dimensions. Although a good neighborhood, the cost of DT quickly becomes prohibitive.

![Graph](image)

Fig. 4. Empty region graphs. (a) B, C and F are neighbors of A, since the circles connecting them with A do not contain any other point. (b) Relaxed ERG. Since D is not a neighbor of A, E can be a neighbor, since its empty region only contains a non-neighbor of A.

### 3.3 Delaunay Triangulations

The Delaunay triangulation (DT) of a collection of points in two dimensions is a triangulation of the convex hull of the points in which the circumcircle of each triangle does not contain a sample point. The edges of this triangulation are a better description of the neighborhoods, since the neighbors represent adjacent Voronoi cells that collectively partition the space around the point. In fact, the Delaunay triangulation produces a good neighborhood graph for the detection of extrema. However, the average neighborhood size in a Delaunay triangulation grows exponentially with the number of dimensions [45]. Moreover, the $O(n^{1+\frac{1}{d-1}})$ computational cost of DT is also exponential in $d$. This is seen in Fig. 3, which plots for 10,000 random sample points both the average neighborhood size and the computational cost as a function of $d$.

### 3.4 Empty Region Graphs

As an alternative to DT, a number of simpler, less costly neighborhood graphs have been proposed, such as the relative neighbor graph (RING) and the Gabriel graph (GG), as surveyed by Jaromczyk et al. [30]. A family of these, known collectively as the empty region graphs, are more efficient to compute (e.g. they have $O(n^3)$ computational complexity) and produce similar or better neighborhoods. Fig. 3 compares the DT with the Gabriel graph and reveals that, although the neighborhood size of GG also grows exponentially with the number of dimensions, it grows much slower than the DT.

Empty region graphs are neighborhood graphs, in which two points are connected by an edge if a canonical region $R$ defined by the points does not contain any other point. More formally, the edges of an empty region graph $G(V,R) = (V,E)$ are given by

$$pq \in E \iff R(p,q) \cap V = \emptyset \quad (3)$$

where the region $R$ defines the neighborhood and is called the empty region. Fig. 4(a) depicts a small neighborhood using a disk as the canonical region $R$. B, C and F are neighbors of A, because their corresponding disks do not contain any other point. D is not a neighbor of A, since its disk contains C. E is not a neighbor either, since its disk contains D. In Section 4.2, we relax this condition to define a different
type of neighborhood graph, as shown in Fig. 4(b). For simplicity of presentation, we assume that $R$ is an open set, i.e., does not contain its boundary. Some of the most common ERGs are:

**Nearest Neighbor Graph (NNG).** This is the directed graph that results from the empty region $R(p,q)$ formed by the open $d$-ball centered on $p$ with radius $d(p,q)$.

$$pq ∈ E ⇔ ∀r ∈ V, d(p,r) ≥ d(p,q)$$ (4)

**Relative Neighborhood Graph (RNG).** This graph is defined by a lune-shaped region consisting of the intersection of two $d$-balls of radius $d(p,q)$, one centered on $p$ and the other centered on $q$, i.e.,

$$pq ∈ E ⇔ ∀r ∈ V, \max\{d(p,r),d(q,r)\} ≥ d(p,q)$$ (5)

**Gabriel Graph (GG).** This is the graph defined by a $d$-ball centered at $\frac{1}{2}(p+q)$ with diameter $d(p,q)$, i.e.,

$$pq ∈ E ⇔ ∀r ∈ V, d(p,r)^2 + d(q,r)^2 ≥ d(p,q)^2$$ (6)

**Diamond Graph (DG).** The DG empty region is formed by the intersection of two solid circular cones with axis $pq$, angle $θ$, and apexes at $p$ and $q$, respectively.

$$pq ∈ E ⇔ ∀r ∈ V, \max\{ξ rpq, ξ rqp\} ≥ θ$$ (7)

Unless otherwise stated, we use $θ = \frac{π}{4}$ as the canonical diamond graph.

**β-Skeleton.** The so-called lune-based β-skeleton is a one-parameter generalization of the RNG and GG, defined as follows:

- For $0 < β < 1$, the empty region is the intersection of all $d$-balls with diameter $d(p,q)/β$ that have $p$ and $q$ on the boundary.
- For $β ≥ 1$, the empty region is the intersection of two $d$-balls with diameter $βd(p,q)$ centered at $(1-\frac{β}{2})p + \frac{β}{2}q$ and $\frac{β}{2}p + (1-\frac{β}{2})q$.

It follows that $β = 2$ gives the RNG, while $β = 1$ is the GG. Thus, $β$ parameterizes a family of empty region graphs. Later on, we will exploit this property to define a probabilistic ERG. Note that geometric inclusion of one region within another also implies a partial order of the resulting neighborhood graphs (in terms of their edges), so that:

$$\text{RNG} ⊆ \text{GG} ⊆ \text{DG} ⊆ \left(β ≤ \frac{1}{\sqrt{2}}\right)$$-skeleton (8)

### 4 Generalized Empty Region Graphs

We have seen that the shape of the empty region $R$ directly determines which edges to connect in an empty region graph. Although $R$ could in principle be any set, we define certain desired properties of a neighborhood that reduce the ERGs to a family of what we call “natural” ERGs. To simplify exposition, we assume that $p$ is at the origin and that $d(p,q) = 1$.

**Natural Empty Region Graph (nERG).** A nERG $G(V,R)$ has an empty region $R$ with the following properties:

- $R$ is a subset of the unit ball. This ensures that $q$ can be excluded from being a neighbor of $p$ only by points $r ∈ R(p,q)$ closer than $q$ is to $p$. Thus, $\text{NNG} ⊆ \text{nERG}$ for any $R$.
- $R$ is symmetric about the hyperplane orthogonal to and bisecting $pq$. As a consequence, $G$ is an undirected graph.
- $R$ is a hypersolid of revolution around $pq$. Thus $R$ is coordinate-free and $G$ is invariant to affine transformations of $V$.
- $R$ has $p$ (and therefore $q$) on its boundary. This prevents a point at infinity from being a neighbor.
- $R$ is simply connected. This ensures that there are no pockets in space where possibly distant points are “safe” from the empty-region test.

As a consequence of our definition, the largest natural region is the one associated with the 2-skeleton (the RNG), and the smallest is the unit line segment, or the 0-skeleton.

### 4.1 Space Pruning Umbras

ERGs produce good neighborhoods because each neighbor prunes the space around it, thus limiting the number of possible neighbors in any given direction. To better understand this, consider the space that is pruned by a neighbor $q$ with respect to a sample point $p$. For any other point $s$ in the domain, $ps$ cannot be an edge if $q$ lies in the empty region $R(p,s)$. For the Gabriel graph, for instance, these points $s$ lie on the opposite side of the hyperplane through $q$ that is orthogonal to $pq$ (see Fig. 5). Because point $q$ shadows all the points in that region, we call it the umbra $U(p,q)$ of $q$ with respect to $p$. We define the umbra implicitly as follows:

**Umbra of an empty region $R$.** The umbra $U(p,q)$ associated with an empty region $R(p,q)$ is the region containing all points $s$ such that:

$$s ∈ U(p,q) ⇐⇒ q ∈ R(p,s)$$ (9)

Fig. 5 shows four ERGs and their umbras. We see that the RNG and the GG both prune the space in halves, the former being more restrictive than the latter. On the other hand, the β-skeleton ($β = \frac{1}{\sqrt{2}}$) prunes a quarter of the domain (in 2D), corresponding to the set of points closest to one of the four possible directions along the two dimensions. However, unlike in DG, this conical umbra has its apex at $q$ and not at $p$, and therefore the resulting graph can be dense. What follows is a method for computing ERGs based on the desired shape of the umbra region.

#### 4.1.1 ERG Construction via Umbra Inversion

Designing an empty region with desired neighborhood pruning properties can be challenging, and often times it is easier to prescribe the umbra. The umbra readily determines what regions of space are pruned by a given point, which may be hard to infer directly from the empty region. Fortunately empty regions and their corresponding umbra form a duality related by a homeomorphism—each point in $R$ maps to a
unique point in $U$. This mapping is given by the inversion transformation. A point $q^{-1}$ is the inverse of $q$ with respect to a hypersphere centered on $p$ with radius $r$ if:

$$d(p,q)d(p,q^{-1}) = r^2, \quad p \parallel pq^{-1}$$

(10)

For all points $r \in R(p,q)$, $r^{-1} \in U(p,q)$ where $r^{-1}$ is the inverse of $r$ with respect to the circle centered at $p$ with radius $d(p,q)$. Moreover, $r^{-2} = r$, and in two dimensions $r^{-1}$ with respect to the unit circle is simply the complex conjugate inverse of $r$.

**Proof.** To understand why points in the umbra and the empty region are related by inversion, let us define points $r \in R(p,q)$ and $s \in U(p,q)$. Therefore, $q \in R(p,s)$ (see Fig. 6(d) for the DG). $R(p,s)$ is denoted by the rotated gray diamond and $R(p,q)$ by the light blue diamond. Since $R(p,q)$ and $R(p,s)$ are similar and they are formed as solids of revolution around the corresponding edge, $\Delta rqp = \Delta qsp$. Thus, since they also share a common angle $\angle rqp = \angle qsp$, triangles $\triangle pqr$ and $\triangle psq$ are similar. This implies $d(p,q)/d(q,s) = d(p,r)/d(r,p)$, or $d(p,q)d(r,p) = d(q,s)^2$. Thus, $s = r^{-1}$ with respect to a circle centered at $p$ with radius $d(p,q)$.

Using inversion, one can construct an empty region starting from a parametric definition of the desired umbra. Natural ERGs, however, are symmetric. We can incorporate these symmetries in the umbra as well. The empty-region symmetry with respect to the edge $pq$ implies the same symmetry of the umbra, i.e. $R$ and $U$ are hypersolids of revolution. The orthogonal symmetry with respect to the hyperplane bisecting $pq$ corresponds to the inverse symmetry with respect to the circle centered at $q$ with radius $d(p,q)$.

Based on these ideas, we show how to construct the DG umbra. For simplicity, and without loss of generality, let $p = (0,0), q = (1,0)$. A portion of the boundary of the DG umbra is defined by the line $y = x$. Because the resulting empty region must be symmetric, we define the umbra only for points $(x,y)$ outside the unit disk centered on $q$, as depicted in Fig. 6(a). Reflection of the partial umbra with respect to the symmetry circle results in a circular arc with endpoints $(1,1)$ and $(0,0)$, as shown in Fig. 6(b). Finally, the region is completed by including its reflection across the $x$ axis (Fig. 6(c)). The diamond-shaped empty region is obtained from the umbra by inversion.

We here showed how to construct the empty region parametrically. The (complement of the) empty regions in our ERG definitions are all expressed implicitly, however, in the form $f(p,q,x) < 0$. It is easy to show that the implicit empty region $f(p,q,x) < 0$ maps to the implicit umbra $f(p,x,q) < 0$, and vice versa. Furthermore, any desired symmetries can be easily be enforced in an implicit representation.

Sparse ERGs like GG may be too restrictive in the way they prune space. But reducing the size of the empty region further comes at the cost of a rapid increase in graph complexity and potential over-smoothing. To alleviate this problem, we introduce two new families of neighborhood graphs below.

### 4.2 Relaxed Empty Region Graphs

We now relax the containment condition for sample points. We observe that in the original ERG all points $V$ around $p$ prune the space of potential neighbors, whether they are neighbors of $p$ or not, which may unnecessarily exclude otherwise good neighbors. For instance, it is possible to arrange points on a path around a sample $p$ as a cascading sequence in which each point shadows the next one, leaving us with only a single nearest neighbor. To address this, we relax the empty-region condition, so that only established neighbors shadow other points, as follows:

**Relaxed Empty Region Graph (rERG).** Let $N(p) = \{q : pq \in E\}$ denote the neighbors of a vertex $p$, and let $q_i$ denote the $i^{th}$ nearest sample to $p$, with $q_0 = p$. A rERG with empty region $R$ and umbra $U$ is defined in terms of the following recurrence:

$$N_1(p) = \{q\}$$

(11)

$$N_i(p) = N_{i-1}(p) \cup \{q_i\} \quad \forall r \in N_{i-1}(p)$$

(12)

$$N(p) = N_{n-1}(p)$$

(13)

where $n$ is the number of data points. In other words, we construct $N(p)$ by adding points in order of increasing distance, as long as they are not shadowed by any point already in $N(p)$. By this definition alone, rERGs are not symmetric. We thus define two variations: the symmetric rERG (sERG) and the mutual rERG (mrERG), as the union and intersection of the rERG and its transpose, respectively, i.e., $pq \in srERG \iff pq \in rERG \lor qp \in rERG$, and $pq \in mrERG \iff pq \in rERG \land qp \in rERG$.

Since we consider only a subset of the points in $V$ in the containment test, it is easy to see that $ERG(R) \subseteq rERG(R)$ for all empty regions $R$. In our experiments, we observe that relaxation only adds a few slightly longer edges than the original ERG, both for the mutual and symmetric graphs.

### 4.3 Stochastic Empty Region Graphs

There is a tradeoff between the accuracy of the topology extraction and the size of the neighborhood. For example, the Gabriel graph is relatively sparse, and although it produces fewer false extrema than the $k$-nearest neighbor graph, it is less precise than the $\beta$-skeleton for $\beta < 1$. However, the $\beta$-skeleton grows much faster in size. A graph with a better tradeoff may lie between those two. In the search for such a graph, we point out that empty region graphs are based on binary decisions that do not take into account how far inside the empty region a point is. Thus, slight perturbations in the sampling pattern may have large effects on the resulting ERG. Consider for example a point $r$ near, but inside, the boundary of an empty region $R(p,q)$, which invalidates $pq$ as an edge. However, it takes a small displacement to make $r$ appear outside $R$, and $pq$ now becomes an edge (assuming no other points lie within $R$). In fact, point $r$ may not invalidate the edge for a slightly different empty region shape. Now consider a point $r'$ in the middle of the empty region. After a small displacement, it is likely that $r'$ remains within the empty region. We say that the “neighborliness” of $p$ and $q$ is more sensitive to $r$ than it is to $r'$.

Based on this observation, we present a generalization of empty region graphs, called the stochastic ERG, which is a weighted neighborhood graph in which each edge has an associated probability:

**Stochastic Empty Region Graph (sERG).** Let $R_{\alpha}$ be a family of empty regions parameterized by a random variable $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Let $\chi_{\alpha}(x)$ denote the indicator function, i.e., $\chi_{\alpha}(x) = 1$ if $x \in R_{\alpha}$ and zero otherwise. For $R(p,q)$ uniformly drawn from $R_{\alpha}$, define $P(p,q,r)$ as the probability that $r \in V$ is not within $R(p,q)$:

$$P(p,q,r) = 1 - \frac{1}{\alpha_{\max} - \alpha_{\min}} \int_{\alpha_{\min}}^{\alpha_{\max}} \chi_{R(p,q)}(r) d\alpha$$

(14)

Then the probability that vertices $p$ and $q$ form an edge $pq$ is

$$P(pq) = \min_{r \in V} P(p,q,r)$$

(15)

The resulting weighted graph is called the stochastic empty region graph.

A realization $sERG^{\alpha}$ of this stochastic graph is drawn from
the probability function, such that:

$$pq \in sERG^* \iff P(pq) \geq u$$

with the random variable $u$ drawn from a probability distribution. For simplicity, we assume a uniform probability distribution and $u$ is chosen as a constant for the whole graph. Alternative definitions, such as picking $u$ adaptively, require further analysis and is beyond the scope of the paper. From this definition, we see that a (deterministic) empty region graph is a special case of a sERG with probability function $P(p, q, r) = 1 - \chi_{R(pq)}(r)$.

Although one may use any set of empty regions $\{R_\alpha\}$, this technique becomes practical when $R_\alpha$ has a natural parameterization, as is the case with the $\beta$-skeleton and the $\theta$-dependent diamond graph, because their empty regions are nested as the parameter value varies. Consequently, a point $x$ that lies between the inner and outer empty region $R_{\alpha_{\min}}$ and $R_{\alpha_{\max}}$, respectively, falls on the boundary of some empty region $R_\alpha$ with $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Finding the corresponding $r$ for a point $r \in V$ is straightforward for both of these ERGs. Consequently, when $r \in R_{\alpha_{\max}} \setminus R_{\alpha_{\min}}$, we may compute $P(p, q, r)$ in closed form as

$$P(p, q, r) = \alpha - \frac{\alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}$$

(17)

Otherwise, $P(p, q, r) = 0$ whenever $r \in R_{\alpha_{\min}}(p, q)$, and $P(p, q, r) = 1$ if $r \notin R_{\alpha_{\min}}(p, q)$.

Note that for all possible draws $sERG^*$ of the stochastic empty region graph, we have

$$ERG_{\alpha_{\min}} \subseteq sERG^* \subseteq ERG_{\alpha_{\max}}$$

(18)

This suggests that the topological accuracy of any particular realization $sERG^*$ is bounded by those of its enclosing ERGs. However, as we shall see, a remarkable result is that in the aggregate, when considering multiple draws, the precision and recall of a sERG may both exceed those of its bounding graphs.

An example stochastic empty region graph is shown in Fig. 7(middle), where the thickness of each edge is proportional to its probability. A random draw from this graph is shown in Fig. 7(right).

A topology extracted using an $sERG$ is a stochastic topology, where each extremum has a probability associated with it. Extrema with low probability are likely to be false extrema caused by neighborhood artifacts, while extrema with high probability are likely to be true extrema. To compute this probability, we perform a series of random draws from the sERG. The probability of a data point being an extremum amounts to the relative number of graphs in which this extremum appears. Sec. 5.2.1 discusses results of stochastic topologies.

### 5 Evaluation

To evaluate the impact of our neighborhood graphs, we study the graph density, accuracy in extremum detection and distribution of persistence.

---

#### Table 1

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>GG</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>GG-r</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>RNG</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>RNG-r</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>KNN-dense</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>KNN-lin</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

---

#### Fig. 8

ERG complexity as a function of dimensionality. RNG grows linearly with the number of dimensions, while GG and its supersets grow exponentially. Our rERGs grow similar to the corresponding ERGs, and the extra cost is not significant. $\beta = 1/\sqrt{2}$ for the $\beta$-Skeleton plot.

---

#### Fig. 9

F measure for several ERGs as a function of sample size $n$. Our relaxed diamond and Gabriel graphs consistently score well.

---

#### Fig. 10

Precision and recall plots for two functions with different spatial frequency using a fixed $n = 100,000$ samples. Relaxation greatly improves precision at only a small loss in recall.

---

An ideal graph should be as sparse as possible and result in perfect precision and recall. We tested different graphs on random samplings of a number of optimization functions from 2 to 9 dimensions. In our methodology, we use the performance of the Gabriel graph as a representative ERG and compare it to that of KNN. For low-dimensional functions, we compare to DG as a representative of a dense ERG and to sparse graphs such as RNG for high dimensions.

### 5.1 ERG Density

Establishing bounds on the number of edges in ERGs has been an active line of research. It is known that for random samplings the Gabriel graph has complexity $O(2^d)$ in $d$ dimensions [30]. Diamond graphs, being supersets of the Gabriel graph, also grow exponentially, but much slower than DT. Fig. 8 shows the average number of neighbors as a function of the number of dimensions for various ERGs, using a sample set of 100,000 random points. Based on our experiments, we observed that our relaxed ERGs do not increase the number of neighbors dramatically over any of the original ERGs. The ramifications of this result are important. Relaxed ERGs are usually faster to compute than their original counterparts due to fewer containment tests. Moreover, as we discuss in the following sections, the inclusion of longer edges along different directions considerably improves the precision of extrema detection.

### 5.2 Topological Precision and Recall

One of the advantages of using empty region graphs over $k$NN is the improvement of precision and recall in the detection of extrema. To classify a detected extremum as a true or false positive, and to find false negatives, we computed the ground truth for a number of analytic functions [43]. We include these extrema as part of the sample points to be able to classify each extremum as either a true or false positive. We then measure the harmonic mean of precision and recall—the F measure—to combine precision and recall in the same plot.
5.3 Persistence Distribution

Topological persistence, the difference in function value between two critical points, indicates if an extremum is part of the signal or is due to noise [20]. To evaluate persistence, we extracted the extremum for some 2D functions for which we know the exact location of each extremum. We then fit a probability density function to the detected true and false extremum, as shown in Fig. 11. Black and red points indicate the persistence of true and false extremum, respectively. If the probabilities of false and true extremum do not overlap, removing false extremum is easy, as shown in Fig. 11(a). Since in general it is unknown if an extremum is true or false, finding a good threshold is easier if the spread of false extremum is small, as shown for the GG and DG in Fig. 11(a). Fig. 11(b) shows a case where the false and true distributions overlap and finding a single threshold to remove all false extremum is not possible without removing other true extremum. In fact, the number of simplified true extremum is much smaller for the GG and DG than for kNN. More than half of the extremum are removed at persistence 0.2 for kNN, while only one sixth for the GG (at persistence 0.03) and none for DG.

We extended our study to higher dimensions, as shown in Fig. 12, for up to 100,000 random points along two to six dimensions. As in Fig. 10, we focus on sparse neighborhood graphs (RNG and GG), which lead to better recall than denser neighborhoods in high dimensions. We observe that the Gabriel graph has a distribution of persistence with lower mean (line plot) and considerably lower standard deviation (area plot) than kNN. In this case, we use \( k = 3d \). The decreasing trend is explained by the inability to detect many true or false extremum in high dimensions, i.e., a reduction in recall. For the relaxed graphs, this particular experiment yielded perfect precision, so the persistence plots (not shown) are horizontal lines with mean 0. Fig. 12(b) shows the result for Ackley’s path function, in which the number of extremum grows exponentially with the number of dimensions. We also notice a considerable difference in the standard deviation of persistence between kNN and GG. The RNG, however, results in a much higher average persistence. After six dimensions (keeping \( n \) constant), the sampling density does not suffice to extract false or true extremum, so the overall distribution of persistence due to topological noise decreases. For the relaxed Gabriel graph, the number of false extremum is considerably lower and the persistence curve (not shown) lies barely above 0 for two to four dimensions.

Although persistence simplification may not be robust to topological noise due to the neighborhood connectivity, ERGs and our variants reduce the chances of having false extremum with high persistence.

5.4 Sampling Quality

So far we have assumed that sample points are randomly distributed. With better sampling strategies, we increase the likelihood of extract-
We have demonstrated that local data analysis tools, such as those involved in determining whether a point is a local extremum or a regular point, are sensitive to the choice of neighborhood used to connect nearby sample points, one such tool being topological decomposition.

7 Discussion and Conclusion
We have demonstrated that local data analysis tools, such as those involved in determining whether a point is a local extremum or a regular point, are sensitive to the choice of neighborhood used to connect nearby sample points, one such tool being topological decomposition.
We found that:

**Natural ERGs generate neighborhoods that help extract critical points more robustly than kNN graphs.** We also observe that persistence-based simplification is more robust for ERGs than kNN graphs. One of the reasons is that kNN neighbors may not be well distributed in direction. When all neighbors appear in the same half-space (which is likely to happen in kNN graphs), the chance of finding a false extremum with arbitrarily large persistence increases. This has important implications in topological analysis and visualization of noisy data, and may suggest heuristics to determine the appropriate persistence thresholds based on the behavior of different ERGs.

**Relaxed ERGs improve precision but may lead to a reduction in recall.** We have seen that ERGs result in higher precision than kNN graphs for comparable \( k \) values (i.e., in 2D, GG and DG require about 4 and 6 neighbors, respectively). Relaxed ERGs, by definition, can only result in higher precision, since they include previously ignored neighbors along underrepresented directions. There may be an impact in recall, when the sampling density decreases. Certain cases may occur when rERGs add neighbors far from a point, resulting in undesired long edges, which limits recall. This is usually solved by restricting the search to a maximum number of neighbors or a maximum radius. On the other hand, there may be cases where precision is preferred to recall, e.g., in optimization problems, in which case relaxed ERGs are more practical.

**Persistence simplification alone is in general not sufficient for removing false positives.** However, the probability of being a critical point is, since it is obtained from a draw of “possible” graphs, which incorporates additional information from several likely neighborhoods. We conclude that a combination of persistence and probability thresholds leads to more accurate and precise topologies.

**As we attempt to sample in higher dimensions, the chances of finding interesting topology decrease rapidly.** Also posed as the curse of dimensionality, poor sampling of a high dimensional space results in missing local extrema. Even for the GG, the neighborhood becomes dense and soon connects all data points. Ideally, a neighborhood proportional to the number of dimensions will have a better chance of finding interesting features. Our results suggest that the relaxed RNG performs considerably better than the RNG when the sampling density is low.

The extraction of a good topological representation of a scalar field remains an open challenge, and more so in irregular and sparsely sampled data. This paper is a stepping stone towards robust topology analysis that empirically correlates the accuracy and precision of extrema detection with the choice of neighborhood graph.
REFERENCES