

## NUDGED ELASTIC BAND IN TOPOLOGICAL DATA ANALYSIS

HENRY ADAMS — ATANAS ATANASOV — GUNNAR CARLSSON

---

**ABSTRACT.** We use the nudged elastic band method from computational chemistry to analyze high-dimensional data. Our approach is inspired by Morse theory, and as output we produce an increasing sequence of small cell complexes modeling the dense regions of the data. We test the method on data sets arising in social networks and in image processing. Furthermore, we apply the method to identify new topological structure in a data set of optical flow patches.

### 1. Introduction

The analysis of large sets of high-dimensional data is a fundamental problem for all branches of science and engineering. Regression analysis can be used effectively when the data has a linear or low degree polynomial structure based on a choice of model for the data. However, it is often the case that the data is genuinely nonlinear and that there isn't an obvious choice for how to model it. The purpose of topological data analysis [5] is to provide methods which produce simple combinatorial representations of the data.

In this paper we construct a cell complex representation in which the cell structure depends on the density of the points in the data set. We adapt the mathematical formalism of Morse theory, which in its idealized form constructs a cell decomposition of a manifold using sublevel sets of a function on the manifold (called the “Morse function”), to the setting of point clouds, i.e. finite sets of points in Euclidean spaces. The specific features of our approach are as follows.

---

2010 *Mathematics Subject Classification.* 62-07, 37B30.

*Key words and phrases.* Topological data analysis, nudged elastic band, Morse theory.

- We use a density estimator as our analogue of the Morse function. One could also use other intrinsic functions on the geometry, such as notions of data depth, to obtain other compact representations.
- In order to sample cells from the analogue of the Morse skeleton of the Morse complex, we adapt the nudged elastic band method (NEB) from computational chemistry [30], [26], [27], [37] to our situation. NEB has been used to study high-dimensional conformation spaces of complicated molecules, and typically uses an energy function as the analogue of the Morse function.
- We produce an increasing sequence of cell complex models. In accordance with the idea of topological persistence, this increasing sequence gives a more accurate representation of the data than the choice of any single complex.
- We currently construct only the one-dimensional skeleton of the cell complex. Producing higher-order cells will require more difficult mathematics, since the minimization problems involved in the construction of such cells are challenging, and are related to minimal surface problems in geometric analysis.

In studying data sets computationally, one finds that outliers will generally obscure topological features. One approach for dealing with outliers is thresholding by density, i.e. studying superlevel sets of a function that estimates density. A difficulty is that the superlevel sets of density frequently require large numbers of points to represent them, since they are “codimension zero” subsets. Consider Figure 1, which contains a standard Morse theoretic picture: a sublevel set of a torus in  $\mathbb{R}^3$  with Morse function given by height. As is standard in Morse theory, the sublevel set is homotopy equivalent to the Morse skeleton, which is the dotted loop in Figure 1. In order to achieve an accurate representation of the homotopy type, sampling from the entire sublevel set will require many more points than sampling from the Morse skeleton. The goal of this paper is to demonstrate that one can effectively sample cells from the Morse skeleton of a density function on the data set, thereby obtaining a more economical representation of the topology of the data set.

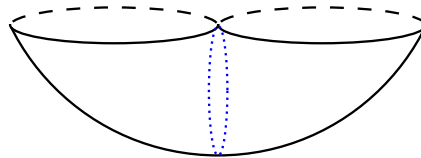


FIGURE 1. A sublevel set of a torus in  $\mathbb{R}^3$  with Morse function given by height. The dotted loop is the Morse skeleton.

We demonstrate how the method works on several nonlinear data sets. First, while it is common to cluster social network data, there is more understanding to be gained about the structure within a single cluster or about the relationships between different clusters. As illustrated with the social network data set, our method can find such relationships. Second, from a collection of optical image patches we recover a minimal geometric description of the three circle model, which has applications in image compression and in texture analysis [7], [34]. Third, we find the primary circle model for range image patches [1]. Fourth, we identify new topological structure in a data set of optical flow patches.

In Section 2 we survey related work, and in Section 3 we provide background material on CW cell complexes, Morse theory, and NEB. We describe our method in Section 4, present results on data sets in Section 5, and conclude in Section 6.

## 2. Related work

The analysis of high-dimensional point cloud data is an important yet challenging task for which a wide variety of tools exist. One approach is to map the data points to a lower dimension. This can be done linearly via projection pursuit [28] or principal component analysis [29], else nonlinearly via Isomap [40], Locally Linear Embedding [36], or Laplacian eigenmaps [2], among others. A second approach is to build a combinatorial representation of the data, for example a dendrogram or a cluster tree, and this is the direction we consider.

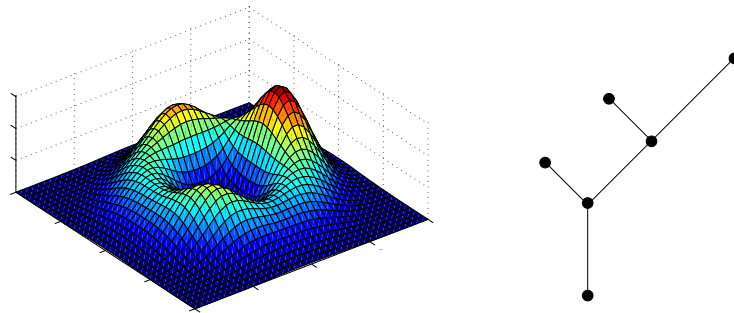


FIGURE 2. A probability density function on the left and the cluster tree for its superlevel sets on the right. The function has a circular base and three bumps of varying heights. For a certain choice of density threshold the superlevel set of the function has the topology of a circle, but this is not reflected in the cluster tree.

Given a point cloud data set drawn from an unknown probability density function, one may cluster the data by approximating the basins of attraction of the modes of density [42], [24, pp. 205]. The goal of [39] is to recover the cluster tree of density, a combinatorial model describing how the connected components

of the superlevel sets of density merge as the density threshold decreases. Regardless of its topology, a connected component of a superlevel set is always represented in the cluster tree by a single point. For example, some superlevel sets of the density function in Figure 2 have circular topology, but this is not apparent in the cluster tree. There is more information to be gained by considering the topology of the superlevel sets.

In topological data analysis one often models a superlevel set of density with a simplicial complex, such as a Čech, Vietoris–Rips, alpha, or witness complex [16], [12]. These simplicial complexes typically contain thousands of simplices, are too large to interpret by hand, and on their own are not useful descriptors of a superlevel set. Yet from these complexes one can estimate the homology groups of the superlevel set using persistent homology [15], [43], [13]. Though homology is a useful signature for describing the geometry of a superlevel set, in general there are several possible models with given homology groups, and determining which model best fits the data is a supervised step. Hence our method, which produces models matching the geometry of the data, plays a complementary role to persistent homology. For example, Carlsson et al. in [7] use persistent homology to identify an image processing data set with  $\text{rank}(H_0) = 1$  and  $\text{rank}(H_1) = 5$ , where  $H_i$  denotes the  $i$ -th homology group. There are many spaces satisfying these homological constraints, including a wedge sum of five circles. Carlsson et al. propose a particular space: a model containing three circles with four intersection points (Figure 10). In Section 5.2 we use our method to obtain this three circle model directly (Figure 11). Our model contains only four 0-cells and eight 1-cells, which is much fewer than the number of simplices needed in a Čech, Vietoris–Rips, alpha, or witness complex reconstruction. Applications of persistent homology often study only one superlevel set of density at a time due to the difficulty of multidimensional persistent homology [8], but exceptions employing kernel density estimators include [9], [17], [4].

There are a variety of adaptations of Morse theory to combinatorial [18] or applied settings. Closely related to the Morse complex that we consider is the Morse–Smale complex, which can be thought of as the intersection of the two Morse complexes for Morse function  $f$  and  $-f$ . The Morse–Smale complex can be efficiently computed for a function defined on a triangulated 2-dimensional or 3-dimensional domain [14], [3], [23], [22]. Furthermore, in [20], [21] a Morse–Smale complex is statistically approximated from a high-dimensional scalar field. A difference between [20], [21] and our setting is that their input is not a point cloud but instead a point cloud equipped with a function value at each point, thought of as a finite sampling from a scalar field. A second difference is that the scalar function in [20], [21] is treated symmetrically, meaning that minima are as important as maxima, and the cells in the complex encode how one can

descend from a maximum down to a minimum. Saddle points are not recovered, and any max-min pair is connected by at most one descending path. In our NEB approach, we equip the point cloud with a density function but we are not interested in the regions where this function is small or attains minima. Saddle points are of primary interest, as they correspond to high-density cells between maxima, and we are interested in detecting when there are multiple high-density paths between two maxima, for example as in Figure 11 (center). For an application of the nudged elastic band method to the study of configuration spaces of hard disks, see [6].

### 3. Background

We briefly introduce three topics: CW complexes, Morse theory, and the nudged elastic band method.

**3.1. CW complexes.** A CW complex is a type of cell complex. For  $k$  a nonnegative integer, a  $k$ -cell is the closed ball  $\{y \in \mathbb{R}^k \mid \|y\| \leq 1\}$  of dimension  $k$ . So a 0-cell is a point, a 1-cell is a line segment, a 2-cell is a disk, etc. A CW complex  $W$  is a topological space formed by the following inductive procedure. The 0-skeleton  $W^{(0)}$  of  $W$  is a set of 0-cells. The 1-skeleton  $W^{(1)}$  is formed by gluing the endpoints of 1-cells to the 0-skeleton, and can be thought of as a graph. Inductively, we form the  $k$ -skeleton  $W^{(k)}$  by gluing the boundaries of  $k$ -cells to the  $(k-1)$ -skeleton  $W^{(k-1)}$ . If  $W$  is a finite-dimensional CW complex, then this process terminates and we have  $W = W^{(k)}$  for some  $k$ . See Figure 3 for an example and [25] for further details.

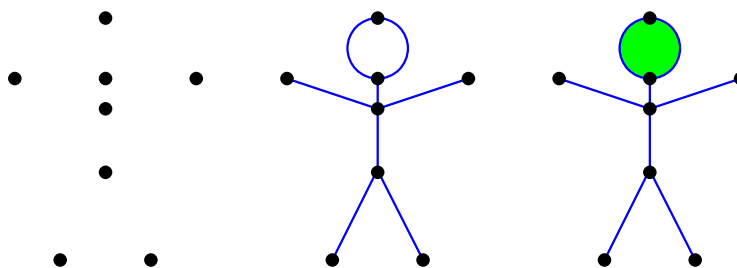


FIGURE 3. A stick figure represented as a CW complex containing eight 0-cells, eight 1-cells, and one 2-cell. The 0-skeleton is on the left, the 1-skeleton is in the center, and the full 2-skeleton is on the right.

**3.2. Morse theory.** The following introduction to Morse theory is informal; see [32] for a thorough treatment. Suppose  $M$  is a compact manifold of dimension  $d$  and Morse function  $f: M \rightarrow \mathbb{R}$  is smooth with non-degenerate critical points  $m_1, \dots, m_k \in M$  satisfying

$$a_0 < f(m_1) < a_1 < f(m_2) < \dots < a_{k-1} < f(m_k) < a_k.$$

The index  $\lambda_i$  of critical point  $m_i$  is the number of linearly independent directions around  $m_i$  in which  $f$  decreases. So a minimum has index 0, a maximum has index  $d$ , and a saddle point has index between 1 and  $d-1$ . Let  $M_a = f^{-1}((-\infty, a])$  be the sublevel set corresponding to  $a \in \mathbb{R}$ . Morse theory tells us that each  $M_{a_i}$  is homotopy equivalent to a CW complex with one  $\lambda_i$ -cell for each critical point  $m_i$ . In particular,  $M_{a_i}$  is homotopy equivalent to  $M_{a_{i-1}}$  with a single  $\lambda_i$ -cell attached. For instance,  $M_{a_1}$  is homotopy equivalent to a point and is obtained from  $M_{a_0}$ , the emptyset, by attaching a single 0-cell. Figure 4 contains an example.

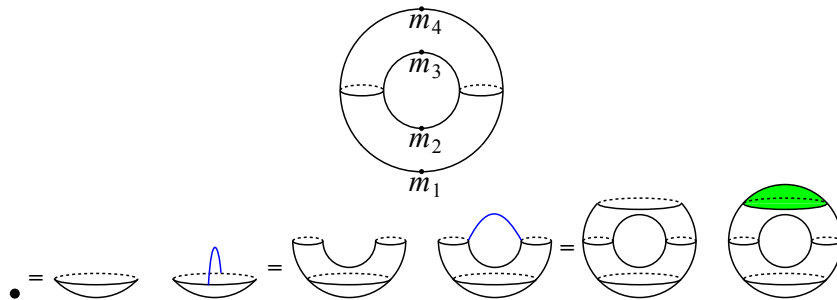


FIGURE 4. Morse theory example. (Top) Manifold  $M$  is a torus and function  $f: M \rightarrow \mathbb{R}$  is the height function. There are four critical points  $m_1, \dots, m_4$  with indices 0, 1, 1, and 2, respectively. Let  $a_0 < f(m_1) < a_1 \dots < f(m_4) < a_4$ . (Bottom) Moving from left to right, we attach cells of dimension 0, 1, 1, and 2 in order to obtain  $M_{a_i}$  from  $M_{a_{i-1}}$ .

Though Morse theory is traditionally stated in terms of sublevel sets, there is an equivalent formulation in terms of superlevel sets. The superlevel sets of  $f$  correspond to the sublevel sets of  $-f$ , so  $m_i$  is a critical point of  $f$  with index  $\lambda_i$  if and only if  $m_i$  is a critical point of  $-f$  with index  $d - \lambda_i$ . It follows that superlevel set  $M^{a_{i-1}} = f^{-1}([a_{i-1}, \infty))$  is obtained from  $M^{a_i}$  by attaching a single cell of dimension  $d - \lambda_i$ .

**3.3. Nudged elastic band.** To find saddle points in a high-dimensional space we use the nudged elastic band method (NEB) from computational chemistry [30], [26], [27], [37]. Consider a chemical system, e.g. several molecules, whose space of states is parametrized by  $\mathbb{R}^n$  and is equipped with a differentiable map  $E: \mathbb{R}^n \rightarrow \mathbb{R}$  encoding the potential energy of the system at each state. The local minima of  $E$  correspond to stable states, and chemists are interested in finding reaction paths between two stable states. A reaction path is a minimum energy path, whose points minimize the energy  $E$  in all directions perpendicular to the path, and which passes through at least one saddle point of index one.

The nudged elastic band method is used to find minimum energy paths. An initial piecewise linear path (called a band) connecting two local minima of

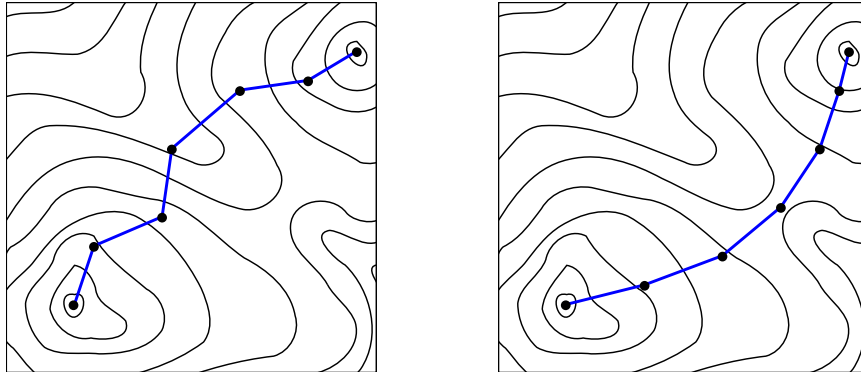


FIGURE 5. (Left) An initial band connecting two local minima. (Right) A minimum energy path found using the nudged elastic band method.

energy  $E$  is chosen, and forces move the band towards a minimum energy path. A gradient force moves each node in the band along the component of  $-\nabla E$  perpendicular to the band, and a spring force ensures that adjacent edges of the band are of approximately equal length, so that the band does not tear apart. More explicitly, the gradient and spring forces appear as the first two terms on the right hand side of equation (4.1) in Section 4.3. Extra forces can be added to smoothen or dampen the motion of the band. Once the forces have stabilized, the convergent band is the approximation to the minimum energy path found by NEB.

#### 4. Method

We suppose data set  $X \subset \mathbb{R}^n$  is a finite sampling from an unknown probability density function  $g: \mathbb{R}^n \rightarrow [0, \infty)$ . We are interested in the regions of high density in the data set, and therefore would like to understand the profile of superlevel sets

$$Y^a = g^{-1}([a, \infty)) = \{y \in \mathbb{R}^n \mid g(y) \geq a\}$$

containing all points in  $\mathbb{R}^n$  with density at least  $a \in [0, \infty)$ . The superlevel sets  $Y^a$  encode how the dense regions of data set  $X$  are organized.

We follow the ideas of Morse theory to build CW complex models  $Z^a$  approximating the superlevel sets  $Y^a$ . First we use  $X$  to build a differentiable density estimate  $f: \mathbb{R}^n \rightarrow [0, \infty)$  approximating the unknown probability density function  $g$ . The 0-cells of our models will be local maxima of  $f$ . The 1-cells will be paths of high density between 0-cells, found using NEB. To find dense 2-cells with boundaries in the 1-skeleton we use an adaptation of NEB. Given a cell  $e \subset \mathbb{R}^n$ , let its density be  $\inf\{f(y) \mid y \in e\}$ . We define  $Z^a$  to be the union of the cells with density greater than or equal to  $a$ .

Note that for  $a \geq b$  we have the inclusion of superlevel sets  $Z^a \subset Z^b$ . A feature of our approach is that we compute the CW complex model  $Z^a$  for many values of  $a$  at once. This allows the user to observe how  $Z^a$  changes as  $a$  decreases. In accordance with the idea of topological persistence [15], [43], [13], knowing  $Z^a$  over a range of values for  $a$  gives a better picture of the data set than knowing  $Z^a$  for any single choice of  $a$ . For example, in Section 5.1 we produce three different models – four vertices, a square loop, and a square disk – for a social network data set (Figure 8). Each model represents the data at a different density threshold, and together they provide a more complete understanding than any single CW complex model alone. In addition, our nested output can be used to measure the significance of topological features. Suppose a 1-cell with density  $a$  appears in  $Z^a$  to form a new loop, and suppose a 2-cell with density  $b \leq a$  appears in  $Z^b$  to fill this loop to a disk. Then  $a - b$  measures how long this loop persists, and if  $a - b$  is large then this loop is likely a significant feature of the data set, but if  $a - b$  is small then the loop may be the product of noise.

Several steps in our approach can be treated in a variety of ways. To name a few, we estimate density, find local maxima, generate random initial bands, simulate bands according to a formulation of NEB, and cluster convergent bands. We use simple methods to handle each of these steps, but we leave open the possibility of substituting more sophisticated methods. Our approach depends on the choice of several parameters, and the main parameter which we find necessary to tune is the standard deviation used to build a density estimator. We use a consistent choice for all other parameters across all of our data sets, which suggests the other parameters may not be as important to tune.

**4.1. Density estimation.** From data set  $X \subset \mathbb{R}^n$ , we build a differentiable density estimator  $f: \mathbb{R}^n \rightarrow [0, \infty)$  approximating the unknown underlying density as follows. Let  $\psi_{x,\sigma}: \mathbb{R}^n \rightarrow [0, \infty)$  be the probability density of a normal distribution centered at  $x \in \mathbb{R}^n$  with standard deviation  $\sigma > 0$ . More explicitly, the  $n \times n$  covariance matrix contains  $\sigma^2$  along the diagonal and 0 elsewhere. We use the kernel estimator  $f(y) = |X|^{-1} \sum_{x \in X} \psi_{x,\sigma}(y)$ . See [38] for other possible estimators, including different kernels or adaptive choices of standard deviation. Bandwidth estimators often aim to minimize the  $L^2$  error, but this is not necessarily optimal for recovering the topology of sublevel sets. We regard standard deviation  $\sigma$  as the main parameter the user must tune: one may increase  $\sigma$  to smoothen  $f$  or decrease  $\sigma$  to expose local detail.

**4.2. 0-cells.** To find the 0-cells of our model, we pick a uniformly random sample of points from  $X$  and flow each along the gradient of  $f$  towards a local maxima using the mean shift iterative procedure [19], [10]. If the size of  $X$  is greater than 2,000 then we select 2,000 uniformly random points from  $X$ , and if



the size of  $X$  is at most 2,000 then we use each point of  $X$  as a seed for mean shift. Let  $y_0 \in X$  be one of the random initial points. We define a sequence  $y_0, y_1, \dots$  by setting  $y_{i+1} = m(y_i)$ , where the mean shift function  $m: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is defined by

$$m(y) = \frac{\sum_{x \in X} \psi_{x,\sigma}(y)x}{\sum_{x \in X} \psi_{x,\sigma}(y)} .$$

The vector  $m(y) - y$  is proportional to the normalized gradient  $\nabla f(y)/f(y)$ , and the  $y_i$  converge to a local maxima of  $f$ . It is necessary to cluster the convergent points to identify which represent the same 0-cell, and we use single-linkage clustering. Since we are interested in CW complex models with more than one 0-cell, we select the clustering threshold parameter to lie in longest interval where a constant number of clusters greater than one is obtained. For the social network, optical image, range image, and optical flow examples in Sections 5.1–5.4, the longest interval with a constant number of clusters constitutes 66.0%, 97.8%, 99.6%, and 90.6% of the total parameter range with more than one cluster, respectively. We select the densest member from each cluster as a 0-cell in our model.

**4.3. 1-cells.** To find the 1-cells of our model we use NEB, which we now describe in our data analysis setting. Our formulation is similar to [30], [26], [27], [37]. A piecewise linear band is given by a list of nodes  $v_1, \dots, v_N$ , with endpoints  $v_1$  and  $v_N$  in our set of 0-cells. Forces act on the intermediate nodes  $v_i$  with  $1 < i < N$  while the endpoints remain fixed. The first task is to approximate a unit tangent vector  $\tau_i$  at each intermediate node. Define  $u_i^+ = v_{i+1} - v_i$  and  $u_i^- = v_i - v_{i-1}$ . We use a naïve tangent estimate  $\tau_i = (u_i^+ + u_i^-)/\|u_i^+ + u_i^-\|$  given by averaging, though more elaborate tangent estimators may be used [26].

At each intermediate node  $v_i$  we define a total force

$$(4.1) \quad F_i = c\nabla f(v_i)|_{\perp} + (\|u_i^+\| - \|u_i^-\|)\tau_i + \text{smoothing}.$$

The expression  $\nabla f(v_i)|_{\perp}$  is the component of  $\nabla f(v_i)$  perpendicular to the tangent  $\tau_i$ , and is called the gradient force. The gradient constant  $c$  adjusts the strength of the gradient force. To normalize with respect to the maximum gradient of the normal distribution, we set

$$c = \left( \sup_{y \in \mathbb{R}^n} \|\nabla \psi_{\bar{0},\sigma}(y)\| \right)^{-1} = (\sigma\sqrt{2\pi})^n \sqrt{e}.$$

The term  $(\|u_i^+\| - \|u_i^-\|)\tau_i$  in (4.1) is the spring force. This name is slightly misleading, as the spring force neither enforces a natural spring length on each edge nor minimizes the length of each edge. Instead, the spring force aims to equate the lengths of adjacent edges in the band.

The *smoothing* term in (4.1) is a force added to prevent kinks, which hinder convergence, from forming in the band. Let  $\theta_i$  be the angle between vectors  $u_i^+$  and  $u_i^-$ . Typically these vectors are close to parallel and  $\theta_i$  is close to zero. Let  $0 \leq \alpha < \beta \leq \pi$  be fixed angles, and let  $h_{\alpha,\beta}: [0, \pi] \rightarrow [0, 1]$  be a function which is zero for  $x \leq \alpha$ , one for  $x \geq \beta$ , and increasing continuously from zero to one as  $x$  increases from  $\alpha$  to  $\beta$ . We define our smoothing force to be  $h_{\alpha,\beta}(\theta_i)(u_i^+ - u_i^-)$ , which moves  $v_i$  in order to decrease  $\theta_i$  whenever  $\theta_i > \alpha$ . In this work we set

$$h_{\alpha,\beta}(x) = \begin{cases} 0 & \text{if } x \leq \alpha, \\ \frac{1}{2} \left( 1 - \cos \left( \pi \frac{x - \alpha}{\beta - \alpha} \right) \right) & \text{if } \alpha < x < \beta, \\ 1 & \text{if } x \geq \beta, \end{cases}$$

with  $\alpha = \pi/6$  and  $\beta = \pi/2$ . We obtain the same CW complex models for the social network, optical image, range image, and optical flow examples in Sections 5.1–5.4 using the smoothing force in [30], although occasionally kinks appear in the bands prior to convergence.

Evolving the band amounts to numerically solving the system of first order differential equations  $v'_i = F_i$ ; see Figure 6. In the chemistry setting it is appropriate to set acceleration proportional to the gradient of the potential energy. In our setting we do not view  $\nabla f$  as a force in the literal sense but only as an indication of which direction to move in order to maximize  $f$ , and hence we use the first derivative  $v'_i$  rather than the second derivative. Nevertheless, we have also had success with the second order equation and do not dismiss its use.

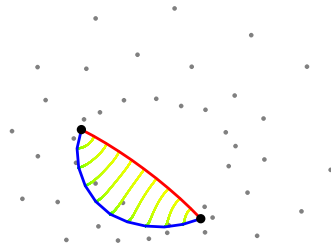


FIGURE 6. A sample 1-cell trial on a small data set. The initial band is in red and the convergent 1-cell is in blue. The lines which fade from yellow to green trace the paths of the intermediate nodes.

Between two 0-cells  $p$  and  $q$  in our model there may be no 1-cells, a single 1-cell, or multiple 1-cells. To find the 1-cells we generate a sample of initial bands joining  $p$  and  $q$ , as described in Appendix B. We evolve each band until it converges and discard non-convergent bands. We also discard convergent bands which pass too close to any other 0-cell  $r \neq p, q$ , as such a band should instead

appear as the concatenation of two others. To identify which convergent bands represent the same 1-cell, we perform single-linkage clustering on the bands with threshold parameter equal to 10% of the diameter of the data set. The social network, optical image, range image, and optical flow examples in Sections 5.1–5.4 are quite stable with respect to this choice – we obtain identical results for any percentage of the diameter between 1% and 45%. Our metric  $d_{p,q,N}$  on bands of  $N$  nodes starting at  $p$  and ending at  $q$  assigns to the two bands  $v_1, \dots, v_N$  and  $\tilde{v}_1, \dots, \tilde{v}_N$  the distance  $d_{p,q,N}(\{v_i\}, \{\tilde{v}_i\}) = (N - 2)^{-1} \sum_{i=2}^{N-1} d(v_i, \tilde{v}_i)$ . From each cluster we select the band with the highest density to represent the corresponding 1-cell. We estimate a band’s density as  $\min\{f(v_1), \dots, f(v_N)\}$ , though one could obtain a more accurate estimate by subdividing the band further or by using the climbing image method of [27].

We make the following consistent parameter choices for all of the data sets. We include  $N = 11$  nodes in the bands and say that a band has converged when  $(N - 2)^{-1} \sum_{i=2}^{N-1} \|v'_i\| < 10^{-4}$ . Increasing  $N$  or decreasing the convergence threshold improves the maximum density path approximation at the expense of computation time, but does not alter our CW complex models.

**4.4. Higher-dimensional cells.** In order to search for higher-dimensional cells, one can imagine adapting NEB. In Appendix C we describe a naïve approach along with its weaknesses. We are interested in better methods, but further work is needed.

### 5. Results

We apply our method to the four data sets in Table 1. From the social network, optical image, and range image data sets we extract small cell complexes which efficiently model the dense regions of the data. Moreover, we identify new topological structure in the data set of optical flow patches. See Appendix A for

	social network	optical image	range image	optical flow
size of data set $X \subset \mathbb{R}^n$	1,127	15,000	15,000	15,000
dimension $n$	5	8	24	16
standard deviation <sup>(1)</sup> $\sigma$	0.45	0.20	0.35	0.30

TABLE 1. Data set information

---

<sup>(1)</sup> This is not a property of the data set, but instead the standard deviation we use to estimate density.

further information on the data sets. Our code is available along with several of the data sets at the following webpage: <http://code.google.com/p/neb-tda>.

**5.1. Social network.** The National Longitudinal Study of Adolescent Health is a school-based study of American youth. In 1994–1995 a sample of high schools and middle schools was chosen, and when possible each high school was paired with a sister middle school from the same community. The students from each school were asked to list up to five of their closest female friends who attend their school or their sister school, and likewise for their male friends.

In [33], Moody creates network graphs from the survey data. Each student is represented by a vertex, and the edge between two students exists if each student listed the other as a friend. We analyze the graph for the students from “Countryside High School,” a pseudonym used by Moody, and its sister middle school. Of the 1,147 students from this community who participated in the survey, 20 students shared no connections with any others and were removed from the graph.

Figure 7 illustrates the following interesting structure in this graph. Most of the students can be placed into one of four groups, containing a majority of white high schoolers, nonwhite high schoolers, white middle schoolers, or nonwhite middle schoolers. There are many friendships between students in the same group. In addition, there are a significant number of friendships between groups of the same school or race category. However, there are very few friendships between groups with neither category in common.

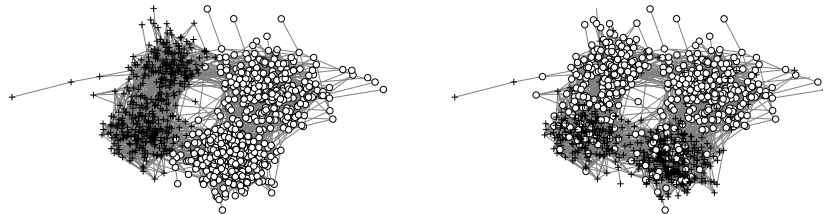


FIGURE 7. Social network for “Countryside High School” and its sister middle school. (*Left*) Cross vertices are students from the middle school and circle vertices are students from the high school. (*Right*) Cross vertices are white students and circle vertices are nonwhite students. A handful of students are without race data and their vertices are left unmarked.

We illustrate how our method can recover relevant structure in this graph. The shortest path distance between vertices  $v$  and  $w$  is the fewest number of edges one must cross to travel from  $v$  to  $w$ . To represent the graph as a point cloud we use stress majorization, an optimization strategy in multidimensional scaling [11], to embed the vertices of the graph as a data set  $X \subset \mathbb{R}^5$  in a manner distorting the shortest path metric as little as possible. The choice of embedding dimension is not critical in this example — we have tested our method on

embeddings in  $\mathbb{R}^n$  for  $2 \leq n \leq 8$  and obtained the same CW complex models, where higher choices of  $n$  allow for smaller distortion of the shortest path metric. Each point  $x \in X$  corresponds to a student, and we regard the resulting set  $X$  as our input data set.

Our method finds four 0-cells, which correspond to the four groups of students. We find four 1-cells which form a square loop. The 0-cells have densities in  $[1.69 \cdot 10^{-2}, 2.02 \cdot 10^{-2}]$  and the 1-cells in  $[1.07 \cdot 10^{-2}, 1.54 \cdot 10^{-2}]$  <sup>(2)</sup>. We find a 2-cell filling the loop with density  $0.79 \cdot 10^{-2}$ . Recall CW complex  $Z^a$  is defined to be the union of the cells with density greater than or equal to  $a$ , and see Figure 8 for three such CW complex models. For  $a \in (1.54 \cdot 10^{-2}, 1.69 \cdot 10^{-2})$  the model  $Z^a$  consists of four 0-cells. Hence we recover the groupings of students based on school and race. For  $a \in (0.79 \cdot 10^{-2}, 1.07 \cdot 10^{-2})$  the model  $Z^a$  is a square. The square reveals that groups sharing either the school or race category are more closely linked than groups sharing neither. This suggests that when making friends it is more difficult to cross two cultural barriers than one. For  $a < 0.79 \cdot 10^{-2}$  the model  $Z^a$  fills to a disk, which is an appropriate representation of the data at a sufficiently coarse scale. Note that this increasing sequence of CW complex models provides a better understanding of the data set than any single model alone.

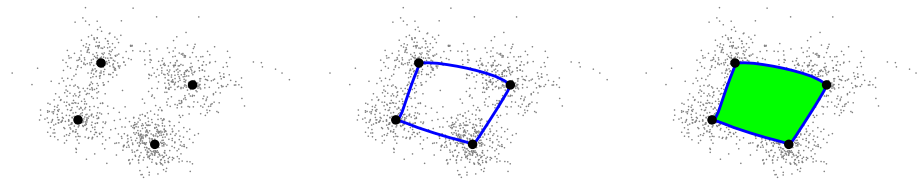


FIGURE 8. Social network data, projected to a plane using principal component analysis. Our CW complex model  $Z^a$  grows as the density threshold  $a$  decreases. From left to right, we have four 0-cells, a square loop, and a disk.

**5.2. Optical image patches.** The optical image database collected by van Hateren and van der Schaaf in [41] contains a variety of indoor and outdoor scenes. From this database, Lee et al. in [31] select a large random sample of  $3 \times 3$  patches, each thought of as a point in  $\mathbb{R}^9$ . The coordinates are the logarithms of grayscale pixel values, and they use the logarithm so that the relative reflectance of an object is invariant under changes in lighting intensity. Lee et al. define a norm measuring the contrast of a patch and select the high-contrast patches. They normalize each patch by subtracting from each pixel the average of the nine

<sup>(2)</sup> Occasionally we also find a 0-cell with density below  $6 \cdot 10^{-4}$ , but due to its low density value it does not affect our analysis.

coordinates, and by dividing by the contrast norm. They change to the Discrete Cosine Transform (DCT) basis  $\{e_1, \dots, e_8\}$ , which maps the patches to the unit sphere  $S^7 \subset \mathbb{R}^8$ . Note in Figure 9 that  $e_1$  and  $e_2$  are horizontal and vertical linear gradients while  $e_3$  and  $e_4$  are horizontal and vertical quadratic gradients. Let  $\mathcal{M}$  be the resulting set of high-contrast, normalized,  $3 \times 3$  patches. Each point of  $S^7$  is close to some point of  $\mathcal{M}$ , but the density of  $\mathcal{M}$  varies widely.

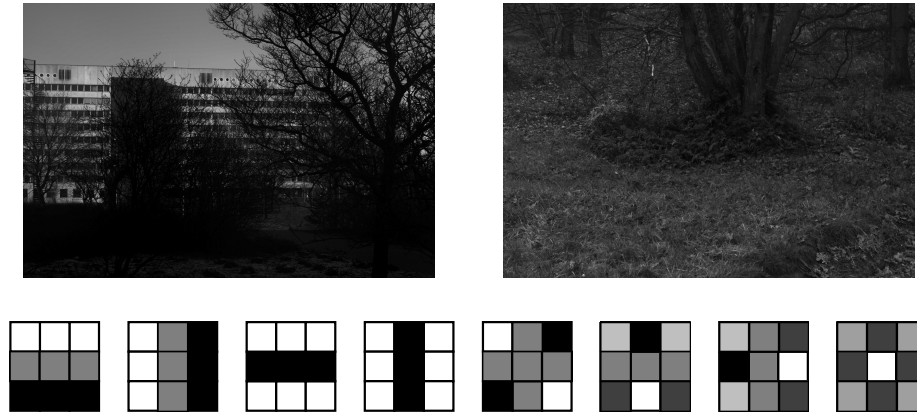


FIGURE 9. (Top) Sample optical images from the van Hateren database.  
(Bottom) Vectors  $e_1, e_2, \dots, e_8$  form the  $3 \times 3$  DCT basis.

Carlsson et al. in [7] use persistent homology to study  $\mathcal{M}$ . Using a family of density estimators they select a family of dense core subsets from  $\mathcal{M}$ . They apply persistent homology, and for a global estimate of density their core subset has  $\text{rank}(H_0) = \text{rank}(H_1) = 1$ , where  $H_i$  denotes the  $i$ -th homology group. This core subset lies near the primary circle  $\{\alpha e_1 + \beta e_2 \mid (\alpha, \beta) \in S^1\}$  containing linear gradient patches at all angles. With a more local estimate of density, the core subset has  $\text{rank}(H_0) = 1$  and  $\text{rank}(H_1) = 5$ . Carlsson et al. identify a three circle model matching this homology profile and the data; see Figure 10. This is a supervised step. In addition to the primary circle, the three circle model contains two secondary circles,  $\{\alpha e_1 + \beta e_3 \mid (\alpha, \beta) \in S^1\}$  and  $\{\alpha e_2 + \beta e_4 \mid (\alpha, \beta) \in S^1\}$ , which include quadratic gradients in the horizontal or vertical direction. The primary circle reflects nature's preference for linear gradients in all directions, and the secondary circles reflect nature's preference for the horizontal and vertical directions.

For our analysis we select a random input data set  $X \subset \mathcal{M}$  of size 15,000. We find four 0-cells located near the four most common patches  $\pm e_1$  and  $\pm e_2$ . Between each of the four 0-cell pairs  $\{e_1, e_2\}$ ,  $\{e_2, -e_1\}$ ,  $\{-e_1, -e_2\}$ , and  $\{-e_2, e_1\}$  we find a quarter-circular 1-cell. Together these form the primary circle. Between the pair  $\{e_1, -e_1\}$  we find two semicircular 1-cells forming the horizontal secondary circle, and between  $\{e_2, -e_2\}$  we find two semicircular 1-cells forming

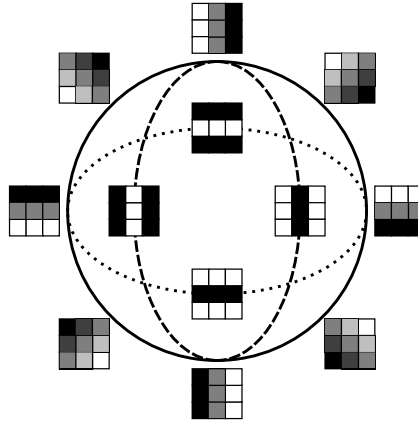


FIGURE 10. Three circle model. The solid outer circle is the primary circle and contains linear gradients. The dotted and dashed inner circles are the horizontal and vertical secondary circles which contain quadratic gradients. Each secondary circle intersects the primary circle twice, but the secondary circles do not intersect each other.

the vertical secondary circle. The 0-cells have densities in  $[2.11, 2.36]$  <sup>(3)</sup>, the primary circle 1-cells in  $[1.17, 1.28]$ , and the secondary circle 1-cells in  $[0.33, 0.38]$ . For  $a \in (1.28, 2.11)$  our model  $Z^a$  is the four most common patches, for  $a \in (0.38, 1.17)$  it is the primary circle, and for  $a < 0.33$  it is the three circle model (Figure 11). Indeed, in an unsupervised fashion we obtain the minimal cellular decomposition of the three circle model that is geometrically accurate.

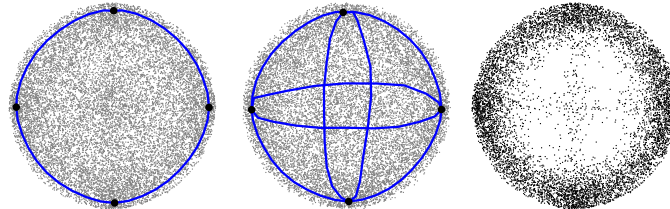


FIGURE 11. (Left) Optical image patches  $X \subset \mathbb{R}^8$  and primary circle complex  $Z^a$  for  $a \in (0.38, 1.17)$ , projected to the  $e_1e_2$  plane. (Center) Three circle model  $Z^a$  for  $a < 0.33$ , projected to the plane spanned by  $e_1 + e_4/4$  and  $e_2 + e_3/4$  so that the secondary circles are visible. (Right) The densest 8,500 points of  $X$ , projected to the  $e_1e_2$  plane. The primary circle appears clearly and the projections of the secondary circle patches form a faint cross.

Carlsson et al. discover a 2-dimensional Klein bottle surface that contains the three circle model as its backbone and which is a good example of how

<sup>(3)</sup> Occasionally we also find a 0-cell with density below 0.08, but due to its low density value it does not affect our analysis.

data set models can not only improve qualitative understanding but also lead to applications. As a low-dimensional manifold the Klein bottle can be used in image compression schemes [7], and in addition the Klein bottle model has been used to identify and analyze optical image textures [34]. In order to identify the Klein bottle using persistent homology, Carlsson et al. add a set  $Q \subset \mathcal{M}$  of optical image patches to their data set, where “the size of  $Q$  can be thought of as a measure of the failure of the density function to cut out the 2-dimensional manifold” [7]. As this is a first exploration of nudged elastic band in topological data analysis, we do not add any such points  $Q$  in order to try to corroborate the Klein bottle model.

**5.3. Range image patches.** A range image is captured by a laser scanner, and each pixel in a range image stores the distance between the laser scanner and the nearest object in the corresponding direction. In [31], Lee et al. select a large random sample of log-valued, high-contrast, normalized,  $3 \times 3$  range image patches from the Brown database, which contains a variety of indoor and outdoor scenes. They observe that the patches cluster near binary patches, where the binary values correspond to foreground and background.

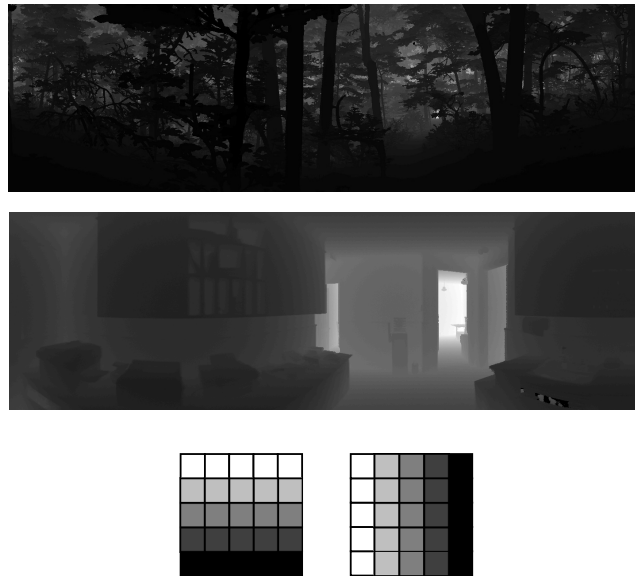


FIGURE 12. (*Top*) Sample range images from the Brown database. (*Bottom*) Vectors  $e_1$  and  $e_5$  from the  $5 \times 5$  DCT basis are horizontal and vertical linear gradients.

Though the largest clusters are arranged in the shape of a circle, the  $3 \times 3$  binary patches are too coarse to fill the full circle. In [1], Adams and Carlsson consider  $5 \times 5$  patches, preprocessed in a manner similar to [31]. They obtain



a large sample  $\mathcal{R}$  of high-contrast, normalized,  $5 \times 5$  range image patches. The  $5 \times 5$  DCT basis now contains 24 vectors  $\{e_1, \dots, e_{24}\}$ , where  $e_1$  and  $e_5$  are horizontal and vertical linear gradients (Figure 12). With persistent homology they find that a dense core subset of  $\mathcal{R}$  has  $\text{rank}(H_1) = 1$ , and they propose the range primary circle model  $\{\alpha e_1 + \beta e_5 \mid (\alpha, \beta) \in S^1\}$  to match this homology.

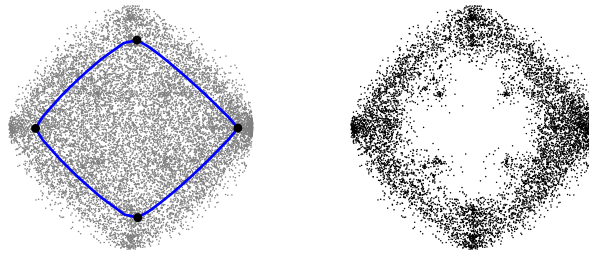


FIGURE 13. (Left) Range image patches  $X \subset \mathbb{R}^{24}$  and range primary circle  $Z^a$  for  $a \in (0.39, 0.51)$ , projected to the  $e_1 e_5$  plane. (Right) The densest 9,000 points of  $X$ .

For our Morse theory approach, we pick a random subset  $X \subset \mathcal{R}$  of size 15,000. We find four 0-cells near  $\pm e_1$  and  $\pm e_5$ . Three of these 0-cells have densities in  $[0.59, 0.75]$  while the 0-cell near  $e_1$  has density 1.47. This reflects the fact that many range patches are shots of the ground and hence are near the horizontal linear gradient given by  $e_1$ . We find four quarter-circular 1-cells forming a loop with densities in  $[0.51, 0.64]$ , and a 2-cell filling the loop with density 0.39. For  $a \in (0.75, 1.47)$  our model  $Z^a$  is the single 0-cell near DCT basis vector  $e_1$ , and for  $a \in (0.39, 0.51)$  our model  $Z^a$  is the range primary circle (Figure 13).

**5.4. Optical flow patches.** In this example we identify new topological structure in an optical flow data set. A video records a sequence of images, and its optical flow is the apparent motion in the sequence of images. At each frame, the optical flow is represented by a vector field with one vector per pixel that points to where that pixel appears to move for the subsequent frame. No instrument measures optical flow directly, and estimating optical flow from a video is an ill-posed problem. However, Roth and Black in [35] create a database of ground-truth optical flow by pairing range images with camera motions and by calculating the produced optical flow. The range images are from the Brown database, and the camera motions are retrieved from a database of videos from hand-held or car-mounted cameras.

With preprocessing steps analogous to [31] we create a large sample  $\mathcal{F}$  of high-contrast, normalized,  $3 \times 3$  optical flow patches. We change to the DCT basis

$\{e_1^u, \dots, e_8^u, e_1^v, \dots, e_8^v\}$ , where the superscript  $u$  denotes flow in the horizontal direction and  $v$  denotes the vertical direction; see Figure 14. We select  $X$  to be a random subset of  $\mathcal{F}$  of size 15,000.

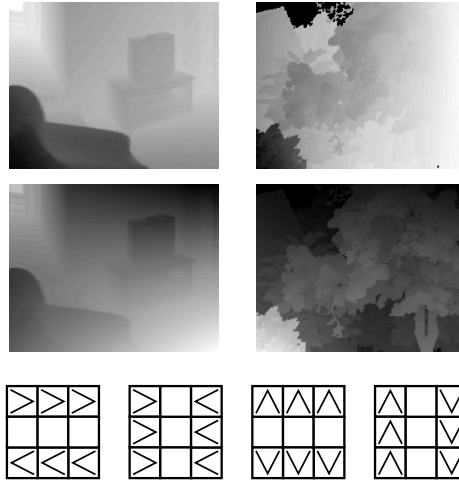


FIGURE 14. (*Top*) Two sample optical flows from the Roth and Black database. Horizontal components are on the top and vertical components are on the bottom. White corresponds to flow in the positive direction ( $+x$  or  $+y$ ) and black corresponds to the negative direction. (*Bottom*) Vector fields  $e_1^u$ ,  $e_2^u$ ,  $e_1^v$ , and  $e_2^v$  from the  $3 \times 3$  optical flow DCT basis. Compare with the patches  $e_1$  and  $e_2$  in Figure 9.

We apply our method and find four 0-cells near  $\pm e_1^u$  and  $\pm e_2^u$  and four quarter-circular 1-cells, all with densities in  $[0.71, 2.79]$ . These cells form a loop. We find a 2-cell filling the loop with density 0.39, and hence for  $a \in (0.39, 0.71)$  our model  $Z^a$  recovers the horizontal flow circle near  $\{\alpha e_1^u + \beta e_2^u \mid (\alpha, \beta) \in S^1\}$  (Figure 15). When applying a camera motion to a range image, the apparent motion of the foreground is faster than that of the background. In particular, applying rightward horizontal camera motion to the linear  $3 \times 3$  range patch  $\alpha e_1 + \beta e_2$  (with  $e_1$  and  $e_2$  from the  $3 \times 3$  DCT basis) produces the optical flow patch  $\alpha e_1^u + \beta e_2^u$ , and applying leftward horizontal camera motion to this same range patch produces the optical flow patch  $-\alpha e_1^u - \beta e_2^u$ . Hence the horizontal flow circle can be obtained by applying horizontal camera motion to range patches from the  $3 \times 3$  analogue of the range primary circle. Also, one expects horizontal camera motion to be more common than vertical motion in hand-held and car-mounted videos. Thus the topology of the optical flow data set combines important patterns from both the range image and camera motion databases: the horizontal flow circle can be interpreted as the image of the range primary circle after applying common horizontal camera motions.

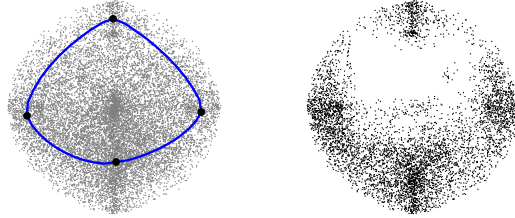


FIGURE 15. (Left) Optical flow patches  $X \subset \mathbb{R}^{16}$  and horizontal flow circle  $Z^a$  for  $a \in (0.39, 0.71)$ , projected to the  $e_1^u e_2^u$  plane. (Right) The densest 6,000 points of  $X$ .

REMARK 5.1. Any method in data analysis must tolerate noisy data points, namely points  $x \in X$  with low density. One may try to remove noisy points from the data set, but in general it is not easy to remove noise without also removing features. Because we use data set  $X$  primarily to build density estimate  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ , we believe that noise which is sparse enough to not significantly affect  $f$  will not significantly affect our CW complex output. As evidence for this claim, consider the large samples  $\mathcal{M}$ ,  $\mathcal{R}$ , and  $\mathcal{F}$  of optical image, range image, and optical flow patches. Both [7] and [1] remove noise from  $\mathcal{M}$  and  $\mathcal{R}$  by studying only core subsets of points with densities in the top 30%. However, in this paper we work with random subsets of  $\mathcal{M}$  and  $\mathcal{R}$  containing both dense and noisy points. As shown in Figures 11, 13, and 15, we extract small CW complex models without first attempting to discard noisy data points.

## 6. Conclusion

We have introduced a method for finding structure in high-dimensional Euclidean data sets. Following Morse theory, we use the nudged elastic band method to sample cells from the analogue of the Morse complex determined by the density function. We produce a nested family of CW complex models representing the dense regions of the data. We test the approach on social network, optical image, and range image data sets and find compact complexes revealing important nonlinear patterns. Moreover, we discover new topological structure, the horizontal flow circle, in a data set of optical flow patches.

### Appendix A. Additional data set information

In this appendix we provide further information about each of the data sets. Our Java code and the optical image, range image, and optical flow data sets are available at the following webpage: <http://code.google.com/p/neb-tda>.

**A.1. Social network.** More information on the social network data set is available at the Adolescent Health Study webpage <http://www.cpc.unc.edu/>

projects/addhealth. To obtain a copy of the non-linkable Adolescent Health Network Structure files, please contact [addhealth@unc.edu](mailto:addhealth@unc.edu).

**A.2. Optical image patches.** The optical image database collected in [41] contains 4,167 grayscale images from indoor and outdoor scenes, each  $1020 \times 1532$  pixels (Figure 9). More details and the database itself are available at <http://bethgelab.org/datasets/vanhateren/>. From this database, Lee et al. in [31] create a set  $\mathcal{M}$  of high-contrast, normalized,  $3 \times 3$  patches through the following preprocessing steps.

- (1) Lee et al. select a large random sample of  $3 \times 3$  patches with coordinates the logarithms of grayscale pixel values.

$$\begin{bmatrix} x_1 & x_4 & x_7 \\ x_2 & x_5 & x_7 \\ x_3 & x_6 & x_9 \end{bmatrix}$$

Each patch is represented by a vector  $x = (x_1, \dots, x_9)^T \in \mathbb{R}^9$ .

- (2) Lee et al. define a norm  $\| \cdot \|_D$  measuring the contrast of a patch. Two coordinates  $x_i$  and  $x_j$  of  $x$  are neighbors, denoted  $i \sim j$ , if the corresponding pixels in the  $3 \times 3$  patch are adjacent. Let

$$\|x\|_D = \sqrt{\sum_{i \sim j} (x_i - x_j)^2}.$$

Lee et al. select the patches with contrast norm in the top 20% of their sample.

- (3) Lee et al. normalize each patch by subtracting the average coordinate value and by dividing by the contrast norm. This maps the patches to a 7-dimensional ellipse.
- (4) Lee et al. change to the Discrete Cosine Transform (DCT) basis  $\{e_1, \dots, e_8\}$  for  $3 \times 3$  patches. The basis vectors are normalized to have contrast norm one, and so this maps the patches to a 7-dimensional sphere.

Let  $\mathcal{M}$  be the resulting set of high-contrast, normalized,  $3 \times 3$  optical patches. Three dense core subsets from  $\mathcal{M}$  are studied in [7] using persistent homology. In this paper we select a random data set  $X \subset \mathcal{M}$  of size 15,000.

**A.3. Range image patches.** The Brown range image database by Lee and Huang is a set of 197 range images from indoor and outdoor scenes, mostly  $444 \times 1440$  pixels (Figure 12). The operational range for the Brown scanner is typically 2–200 meters, and the distance values for the pixels are stored in units of 0.008 meters. The database can be found at the following webpage: <http://www.dam.brown.edu/ptg/brid/index.html>.

From the Brown database we obtain a space of range image patches through the following steps, which are similar to the procedures in [31], [1].

- (1) We randomly select about  $4 \cdot 10^5$  size  $5 \times 5$  patches from the images in the database. Each patch is represented by a vector  $x \in \mathbb{R}^{25}$  with logarithm values.
- (2) We compute the contrast norm

$$\|x\|_D = \sqrt{\sum_{i \sim j} (x_i - x_j)^2}$$

of each patch and select the patches with contrast norm in the top 20% of the entire sample.

- (3) We subtract from each patch the average of its coordinates and divide by the contrast norm.
- (4) We change to the DCT basis  $\{e_1, \dots, e_{24}\}$  for  $5 \times 5$  patches, normalized to have contrast norm one. This maps the patches to a 23-dimensional sphere.

Let  $\mathcal{R}$  be the resulting set of high-contrast, normalized,  $5 \times 5$  range patches. Our data set is a random subset  $X \subset \mathcal{R}$  of size 15,000.

**A.4. Optical flow patches.** The optical flow database by Roth and Black (see [35]) contains 800 optical flow fields, each  $250 \times 200$  pixels (Figure 14). This ground-truth optical flow is generated by pairing range images from the Brown range image database with camera motions. The camera motions are extracted from a database of 67 videos from hand-held or car-mounted video cameras, each approximately 100 frames long, using *boujou* software by 2d3 Ltd. available at <http://www.2d3.com>. The Roth and Black database is available at <http://www.gris.informatik.tu-darmstadt.de/~sroth/research/flow/downloads.html>.

From the Black and Roth database we create a space of optical flow patches, and our preprocessing is similar to that of [31].

- (1) We randomly choose  $4 \cdot 10^5$  size  $3 \times 3$  optical flow patches from the Roth and Black database. Each patch is a matrix of ordered pairs, where  $u_i$  and  $v_i$  are the horizontal and vertical components, respectively, of the flow vector at pixel  $i$ .

$$\begin{bmatrix} (u_1, v_1) & (u_4, v_4) & (u_7, v_7) \\ (u_2, v_2) & (u_5, v_5) & (u_8, v_8) \\ (u_3, v_3) & (u_6, v_6) & (u_9, v_9) \end{bmatrix}$$

We define  $u = (u_1, \dots, u_9)^T$  and  $v = (v_1, \dots, v_9)^T$  to be the vectors of horizontal and vertical flow components. We rearrange each patch to be a vector

$$x = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathbb{R}^{18}.$$

- (2) We compute the contrast norm

$$\|x\|_D = \sqrt{\sum_{i \sim j} \|(u_i, v_i) - (u_j, v_j)\|^2}$$

for all patches. We select the patches with contrast norm in the top 20% of the entire sample.

- (3) We normalize the patches to have zero average flow. More explicitly, given a patch
- $x$
- , let
- $\bar{u} \in \mathbb{R}^9$
- have each entry equal to
- $\frac{1}{9} \sum_{i=1}^9 u_i$
- , the average of the horizontal components. Let
- $\bar{v}$
- be defined similarly. We replace each patch
- $x$
- with

$$\begin{pmatrix} u - \bar{u} \\ v - \bar{v} \end{pmatrix}.$$

We then divide each patch by its contrast norm.

- (4) We change to the DCT basis
- $\{e_1^u, \dots, e_8^u, e_1^v, \dots, e_8^v\} \subset \mathbb{R}^{18}$
- for
- $3 \times 3$
- optical flow patches, where

$$e_i^u = \begin{pmatrix} e_i \\ \vec{0} \end{pmatrix} \quad \text{and} \quad e_i^v = \begin{pmatrix} \vec{0} \\ e_i \end{pmatrix}.$$

This maps the patches to a 16-dimensional sphere.

Let  $\mathcal{F}$  be the resulting set of high-contrast, normalized,  $3 \times 3$  optical flow patches. Our data set is a random subset  $X \subset \mathcal{F}$  of size 15,000.

## Appendix B. Initial bands

Let  $p$  and  $q$  be distinct 0-cells. Our method for generating the initial bands between  $p$  and  $q$  depends on whether data set  $X$  is a general data set in  $\mathbb{R}^n$  or whether  $X$  is normalized to lie on a unit sphere  $S^{n-1} \subset \mathbb{R}^n$ .

For a general data set  $X \subset \mathbb{R}^n$ , we pick a uniformly random vector  $y$  from the set of all unit vectors perpendicular to  $p - q$ , which is a sphere of dimension  $n - 2$ . We also pick  $r \in [0, d(p, q)]$  uniformly randomly. The resulting initial band is  $N$  evenly distributed nodes along the circular arc (or straight line, with probability zero) between the points  $p$ ,  $(p + q + ry)/2$ , and  $q$ .

If  $X$  is normalized to lie on a unit sphere  $S^{n-1} \subset \mathbb{R}^n$ , we generate initial bands lying near  $S^{n-1}$ . Though  $p$  and  $q$  are near dense regions of  $X$  they need not lie in  $X$  nor in  $S^{n-1}$ . Let  $\hat{p} = p/\|p\|$  and  $\hat{q} = q/\|q\|$ . We pick a uniformly random vector  $y \neq \hat{p}, \hat{q}$  in  $S^{n-1}$ . The plane defined by  $y, \hat{p}$ , and  $\hat{q}$  intersects  $S^{n-1}$  in a circle. Let  $\hat{p} = \hat{v}_1, \dots, \hat{v}_N = \hat{q}$  be the unique band that is evenly-spaced along this intersection circle, that starts at  $\hat{p}$ , that ends at  $\hat{q}$ , and that does not pass through  $y$ . We define our initial band to be  $p = v_1, \dots, v_N = q$ , where

$$v_i = \frac{(N - i)\|p\| + (i - 1)\|q\|}{N - 1} \hat{v}_i.$$

### Appendix C. Higher-dimensional cells

One can imagine adapting the NEB method to search for higher-dimensional cells. Suppose for  $k > 1$  that we have obtained the  $(k - 1)$ -skeleton of a CW complex model. Now given an initial  $k$ -cell, we would like to move it towards a maximum density  $k$ -cell, and we describe a naïve approach. We model a  $k$ -cell by a graph  $G = (V, E)$  that is the 1-skeleton of a regular CW complex homeomorphic to a  $k$ -cell. We choose an attaching map from the boundary of the  $k$ -cell to the  $(k - 1)$ -skeleton of the CW complex model; this fixes the nodes in  $V$  that lie in the boundary of the  $k$ -cell. Given a node  $\alpha \in V$ , let  $v_\alpha \in \mathbb{R}^n$  denote its position and let  $V_\alpha = \{\beta \in V \mid \{\alpha, \beta\} \in E\}$  denote its set of incident vertices. We estimate the  $k$ -dimensional tangent space at  $v_\alpha$  to be the span of the first  $k$  components of a principal component analysis on  $\{v_\beta - v_\alpha \mid \beta \in V_\alpha\}$ . If  $\alpha$  is not a boundary node, then we define

$$F_\alpha = c\nabla f(v_\alpha)|_\perp + \sum_{\beta \in V_\alpha} (v_\beta - v_\alpha)$$

to be the force at vertex  $\alpha$ . As before,  $\nabla f(v_\alpha)|_\perp$  is the component of  $\nabla f(v_\alpha)$  perpendicular to the tangent space and is called the gradient force. Gradient constant  $c$  adjusts the strength of the force. The term  $\sum_{\beta \in V_\alpha} (v_\beta - v_\alpha)$  is the spring force. We numerically solve the system of first order differential equations  $v'_\alpha = F_\alpha$ .

We point out several weaknesses in the straightforward approach above.

- The spring force does not generalize the dimension  $k = 1$  case.
- The gradient forces and spring forces do not both go to zero; instead, they balance against one another. This means that the cell does not converge exactly to the maximum density cell. A possible remedy is to project the spring force to the tangent space. One may then need to add an appropriate smoothing force to prevent kinks from forming in the cell.
- It may be preferable to have forces depend not only on the underlying 1-skeleton graph of the  $k$ -cell but also on its higher dimensional cells.
- One may want an adaptive representation of a cell whose triangulation changes as the cell moves. Otherwise, the choice of an initial triangulation may affect the subsequent motion of the cell.
- If the  $(k - 1)$ -skeleton of the CW complex is complicated, then it is not clear which attaching maps to choose as the potential boundaries for  $k$ -cells.

We are interested in improved generalizations of NEB to higher-dimensional cells.

Nevertheless, we use this approach to search for 2-cells in of our data sets.

We use a particular CW complex homeomorphic to a 2-cell, containing 201

vertices, 400 edges, and 200 2-cells, whose 1-skeleton is a web-shaped graph with 20 nodes on each of 10 concentric rings (Figure 16). We place the boundary nodes evenly-spaced on a chosen loop in the 1-skeleton; these nodes remain fixed. For the initial location of the 2-cell, we place the center node at the average of the boundary nodes, and we linearly interpolate between the boundary and center to place all remaining nodes. We estimate a convergent cell's density as  $\min_{\alpha \in V} f(v_\alpha)$ . See Figure 16 for the resulting 2-cells.

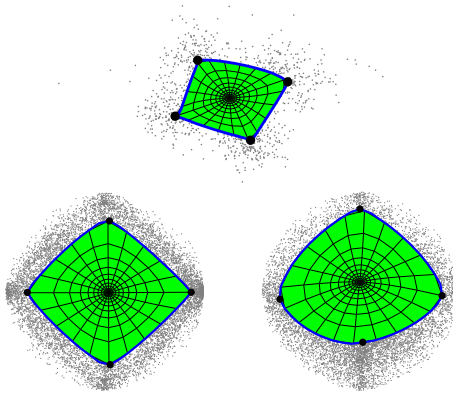


FIGURE 16. The 2-cells in the data sets. (*Top*) Social network data, projected to a plane using principal component analysis. (*Bottom left*) Range image patches, projected to the  $e_1e_5$  plane. (*Bottom right*) Optical flow patches, projected to the  $e_1^ue_2^u$  plane.

**Acknowledgements.** We would like to thank Tim Harrington and Andrew Tausz for their help with the social network data and Guillermo Sapiro for his help with the optical flow data. The first author was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation. The second author would like to thank the I.I. Rabi Science Scholars Program, Columbia University. This work is supported by Office of Naval Research Grant N00014-08-1-0931, Air Force Office of Scientific Research Grant FA9550-09-1-0143, Air Force Office of Scientific Research Grant FA9550-09-0-1-0531, and National Science Foundation Grant DMS-0905823.

#### REFERENCES

- [1] H. ADAMS AND G. CARLSSON, *On the nonlinear statistics of range image patches*, SIAM J. Imag. Sci. **2** (2009), 110–117.
- [2] M. BELKIN AND P. NIYOGI, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput. **15** (2003), 1373–1396.
- [3] P.T. BREMER, B. HAMANN, H. EDELSBRUNNER AND V. PASCUCCI, *A topological hierarchy for functions on triangulated surfaces*, IEEE Trans. on Visual Comput. Graphics **10** (2004), 385–396.



- [4] P. BUBENIK AND P.T. KIM, *A statistical approach to persistent homology*, Homology, Homotopy and Applications **9** (2007), 337–362.
- [5] G. CARLSSON, *Topology and data*, Bull. Amer. Math Soc. **46** (2009), 255–308.
- [6] G. CARLSSON, J. GORHAM, M. KAHLE AND J. MASON, *Computational topology for configuration spaces of hard disk*, Phys. Rev. E **85** (2012).
- [7] G. CARLSSON, T. ISHKHANOV, V. DE SILVA AND A. ZOMORODIAN, *On the local behavior of spaces of natural images*, Int. J. Comput. Vision **76** (2008), 1–12.
- [8] G. CARLSSON AND A. ZOMORODIAN, *The theory of multidimensional persistence*, Discrete Comput. Geom. **42** (2009), 71–93.
- [9] F. CHAZAL, L.J. GUIBAS, S.Y. OUDOT AND P. SKRABA, *Persistence-based clustering in Riemannian manifolds*, Proc. 27th Ann. ACM Sympos. on Comput. Geom., June, 2011, 97–106.
- [10] Y. CHENG, *Mean shift, mode seeking, and clustering*, IEEE Trans. Pattern Anal. **17** (1995), 790–799.
- [11] T.F. COX AND M.A.A. COX, *Multidimensional Scaling*, Chapman and Hall, London, 2001.
- [12] V. DE SILVA AND G. CARLSSON, *Topological estimation using witness complexes*, Eurographics Symposium on Point-Based Graphics, June, 2004.
- [13] H. EDELSBRUNNER AND J. HARER *Computational Topology: An Introduction*, Amer. Math. Soc., Providence, 2010.
- [14] H. EDELSBRUNNER, J. HARER AND A. ZOMORODIAN, *Hierarchical Morse–Smale complexes for piecewise linear 2-manifolds*, Discrete Comput. Geom. **30** (2003), 87–107.
- [15] H. EDELSBRUNNER, D. LETSCHER AND A. ZOMORODIAN, *Topological persistence and simplification*, Discrete Comput. Geom. **28** (2002), 511–533.
- [16] H. EDELSBRUNNER AND E. P. MÜCKE, *Three-dimensional alpha shapes*, ACM Trans. Graphics **13** (1994), 43–72.
- [17] B.T. FASY, F. LECCI, A. RINALDO, L. WASSERMAN, S. BALAKRISHNAN AND A. SINGH, *Statistical Inference For Persistent Homology: Confidence Sets For Persistence Diagrams* (2014), Preprint arXiv:1303.7117.
- [18] R. FORMAN, *A user’s guide to discrete Morse theory*, Séminaire Lotharinen de Combinatoire **48** (2002).
- [19] K. FUKUNAGA AND L.D. HOSTETLER, *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE Trans. Inform. Theory **21** (1975), 32–40.
- [20] S. GERBER, P.T. BREMER, V. PASCUCCI AND R. WHITAKER, *Visual exploration of high dimensional scalar functions*, IEEE Trans. on Visual Comput. Graphics **16** (2010), 1271–1280.
- [21] S. GERBER AND K. POTTER, *Data analysis with the Morse–Smale complex: The msr package for R*, J. Stat. Software **50** (2011), 1–22.
- [22] A. GYULASSY, P.T. BREMER, B. HAMANN AND V. PASCUCCI, *A practical approach to Morse–Smale complex computation: Scalability and generality*, IEEE Trans. Visual Comput. Graphics **14** (2008), 1619–1626.
- [23] A. GYULASSY, V. NATARAJAN, V. PASCUCCI AND B. HAMANN, *Efficient computation of Morse–Smale complexes for three-dimensional scalar functions*, IEEE Trans. Visual Comput. Graphics **13** (2007), 1440–1447.
- [24] J.A. HARTIGAN, *Clustering Algorithms*, Wiley, New York, 1975.
- [25] A. HATCHER, *Algebraic Topology*, Cambridge University Press, Cambridge, 2002.
- [26] G. HENKELMAN AND H.JÓNSSON, *Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points*, J. Chem. Phys. **113** (2000), 9978–9985.
- [27] G. HENKELMAN, B.P. UBERUAGA AND H. JONSSON, *A climbing image nudged elastic band method for finding saddle points and minimum energy paths*, J. Chem. Phys. **113** (2000), 9901–9904.
- [28] P. J. HUBER *Projection pursuit*, Ann. Stat. (1985), 435–475,

- [29] I.T. JOLLIFFE, *Principal Component Analysis*, Springer, 1986.
- [30] H. JÓNSSON, G. MILLS AND K.W. JACOBSEN, *Nudged elastic band method for finding minimum energy paths of transitions*, Classical and Quantum Dynamics in Condensed Phase Systems, (B.J. Berne, G. Ciccotti and D.F. Coker, eds.), World Scientific, Singapore, 1998, 385–404.
- [31] A.B. LEE, K.S. PEDERSEN AND D. MUMFORD, *The nonlinear statistics of high-contrast patches in natural images*, Int. J. Comput. Vision **54** (2003), 83–103. (2003,
- [32] J. MILNOR, *Morse Theory*, Princeton University Press, Princeton, 1965.
- [33] J. MOODY, *Race, school integration, and friendship segregation in America*, Am. J. Sociol. **107** (2001), 679–716.
- [34] J.A. PEREA AND G. CARLSSON, *A Klein-bottle-based dictionary for texture representation*, Int. J. Comput. Vision **107** (2014), 75–97.
- [35] S. ROTH AND M.J. BLACK, *On the spatial statistics of optical flow*, Int. J. Comput. Vision **74** (2007), 33–50.
- [36] S.T. ROWEIS AND L.K. SAUL, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290** (2000), 2323–2326.
- [37] D. SHEPPARD, R. TERRELL AND G. HENKELMAN, *Optimization methods for finding minimum energy paths*, J. Chem. Phys. **128** (2008).
- [38] B.W. SILVERMAN, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
- [39] W. STUETZLE AND R. NUGENT, *A generalized single linkage method for estimating the cluster tree of a density*, J. Comput. Graph. Stat., **19** (2010), 397–418.
- [40] *A global geometric framework for nonlinear dimensionality reduction*, J. B. TENENBAUM AND V. DE SILVA AND J. C. LANGFORD, Science, **290** (2000), 2319–2323.
- [41] J.H. VAN HATEREN AND A. VAN DER SCHAAF, *Independent component filters of natural images compared with simple cells in primary visual cortex*, Proc. R. Soc. Lond. Ser. B, **265** (1998), 359–366.
- [42] D. WISHART, *Mode analysis: a generalization of nearest neighbor which reduces chaining effects*, Numerical Taxonomy, (A.J. Cole, ed.), Academic Press, London, 1969, 282–311.
- [43] A. ZOMORODIAN AND G. CARLSSON, *Computing persistent homology*, Discrete Comput. Geom., **33** (2005), 249–274.

*Manuscript received December 31, 2013*

HENRY ADAMS  
 Department of Mathematics  
 Duke University  
 NC 27708, USA  
*E-mail address:* hadams@math.duke.edu

ATANAS ATANASOV  
 Department of Mathematics  
 Harvard University  
 Cambridge, MA 02138, USA  
*E-mail address:* nasko@math.harvard.edu

GUNNAR CARLSSON  
 Department of Mathematics  
 Stanford University  
 Stanford, CA 94305, USA  
*E-mail address:* gunnar@math.stanford.edu