

Stratified Space Learning: Reconstructing Embedded Graphs

Y. Bokor^a, D. Grixti-Cheng^a, M. Hegland^a, S. Roberts^a and K. Turner^a

^aMathematical Sciences Institute, Australian National University, Canberra, Australia
Email: yossi.bokor@anu.edu.au

Abstract: Many data-rich industries are interested in the efficient discovery and modelling of structures under-lying large data sets, as it allows for the fast triage and dimension reduction of large volumes of data embedded in high dimensional spaces. The modelling of these underlying structures is also beneficial for the creation of simulated data that better represents real data. In particular, for systems testing in cases where the use of real data streams might prove impractical or otherwise undesirable. We seek to discover and model the structure by combining methods from topological data analysis with numerical modelling. As a first step in combining these two areas, we examine the recovery a linearly embedded graph $|G|$ given only a noisy point cloud sample X of $|G|$.

An *abstract graph* G consists of two sets: a set of vertices V and a set of edges E . An *embedded graph* $|G|$ in n dimensions is a geometric realisation of an abstract graph obtained by assigning to each vertex a unique coordinate vector in \mathbb{R}^n and then each edge is the line segment between the coordinate vectors of the corresponding vertices. Given an embedded graph $|G| \subset \mathbb{R}^n$, a *point cloud* sample of $|G|$ consists of a finite collection of points in \mathbb{R}^n sampled from $|G|$, potentially with noisy perturbations. We will suppose that this sample has bounded noise of ϵ and is sufficiently sampled so every point in $|G|$ is within ϵ of some sample. Such a sample is called an ϵ -sample.

We can view this as a semiparametric model. Once the abstract graph is fixed we have a parametric model where the parameters are the locations of the vertices. In order to guarantee correctness of our algorithm, we will need to make some reasonable geometric assumptions on the embedded graph.

To learn the embedded graph $|G|$, we first learn the structure of the abstract graph G . We do this by assigning a dimension of either 0 or 1 to each $x \in X$ (depending on local topological structure) and then cluster points into groups representing embedded vertices or edges. Using local topological structure, we then assign to each abstract edge cluster a pair of abstract vertex clusters, to obtain the incidence relations of the abstract graph. Finally, we use nonlinear least squares regression to model the embedded graph $|G|$.

The approach presented in this paper relies on topological concepts, such as stratification and local homology, which will be used in future research that will generalise this approach to embeddings on more general structures, in particular *stratified spaces*.

A stratified space \mathbb{X} is a topological space with a partition into topological manifolds $\{X_i\}_{i \in I}$, called strata, such that for all i and j , $X_i \cap X_j = \emptyset$, and if $X_i \cap X_j \neq \emptyset$, then $X_i \subseteq \overline{X_j}$.

If we are interested in a specific topological structure, we can require that the strata satisfy additional criteria, so that this structure is constant across each strata. In the case presented in this paper, the vertices of a graph (abstract or embedded) are the 0-strata, and the (open) edges are 1-strata.

The authors are unaware of any algorithms which recover both the abstract graph, and model its embedding.

Keywords: *Stratified space learning, structure recovery, point clouds*

1 INTRODUCTION

We seek to uncover information about an embedded $|G| \subset \mathbb{R}^n$ from a set $X \subset \mathbb{R}^n$ of noisy sample point $x^{(i)}$ which are close to $|G|$. We consider $|G|$ to be topological space, and with the topology induced by \mathbb{R}^n . Here we consider a *stratified space* $|G|$ which consists of a partition $|G| = \bigcup_{i=1}^k |G|_i$ with strata $|G|_i$ overlapping at their boundaries.

Definition 1. *A topologically stratified space G is a space with a partition into topological manifolds $\{G_i\}_{i \in I}$, called strata, such that for all i and j , $G_i \cap G_j = \emptyset$, and if $G_i \cap \overline{G_j} \neq \emptyset$, then $G_i \subseteq \overline{G_j}$.*

One says that the strata G_i are *glued together* at their boundaries. The aim of this project is to recover the strata from the data points. Here the data points are an ϵ -sample $X \subset \mathbb{R}^n$ of $|G|$, which is a point cloud such that the *Hausdorff* distance between X and $|G|$ is bounded above by ϵ . Recall that the Hausdorff distance between two subsets $Y, Y' \subset \mathbb{R}^n$ is

$$d_H(Y, Y') := \max\left\{\sup_{y \in Y} \inf_{y' \in Y'} d(y, y'), \sup_{y' \in Y'} \inf_{y \in Y} d(y, y')\right\}, \quad (1)$$

with $d(y, y')$ the standard Euclidean distance on \mathbb{R}^n .

This is an inverse problem and the corresponding direct problem is generating data points from strata. The project consists of three logical steps:

1. the data is partitioned as

$$X = \bigcup_{j=1}^k X_j \quad (2)$$

where the partition X_j contain the points originating from stratum G_j ,

2. establish the dimension of the strata, and the glueing between them,
3. concise numerical descriptions of the strata are computed.

When viewed as a semiparametric modelling problem, the first two steps determine the non-parametric components. For this we use topological data analysis. In contrast the last step numerically fits the parametric model discovered in the earlier procedures.

In this paper we do not consider all stratified spaces and instead restrict to linear embeddings $|G|$ of abstract graphs G . An abstract graph G is a stratified space with only 0- and 1dimensional strata, and each 1 dimensional strata (1-stratum) has two 0 dimension strata (0-stratum) as its boundary. We thus consider the 1-stratum G_i^1 to be a tuple of 0-strata $(G_{i_1}^0, G_{i_2}^0)$. We often omit the superscript, in particular when writing 1-strata as a tuple.

To obtain a stratification of a graph, we will consider the *local homology* of each strata. Given a topological space G , the local homology at a point g describes the linear combinations of closed manifolds in a small region of G around g . Examples of stratifications using other local topological structures can be found in Hughes and Weinberger (2001), Nanda (2019), Brown and Wang (2018), and Bendich et al. (2012). Hughes and Weinberger (2001) provide an overview of different types of stratifications, Nanda uses information about the local cohomology of cells in a regular CW complex to obtain a canonical stratification, Brown and Wang (2018) apply sheaves to obtain stratifications of cell complexes, and Bendich et al. (2012) appeal to local homology to infer which points in a sample of a stratified space have been sampled from the same strata. Future research will expand the class of embedded stratified spaces we can recover from ϵ -samples.

2 GEOMETRIC AND TOPOLOGICAL INTUITION

The first part of this semi-parametric problem is to identify the abstract graph structure and partition the points with respect to this abstract graph. To do this we classify samples topologically by their local neighbourhoods. This can be defined rigorously through the notion of local homology. Although the theoretical development of modifications of local homology to work with point cloud motivates and informs the algorithms here, for the sake of clarity and brevity we will focus on the computational geometry intuition behind the methods. This

is possible as the local homology for graphs can be expressed in terms of connected components of appropriate subsets. For more complicated stratified spaces, more sophisticated machinery from algebraic topology will be required. A description relating this computational geometry to local homology is in the subsection below.

We begin with the idea that whether a point on a graph is on an edge or a vertex can be determined by looking at how many times the graph intersects a sphere around that point of a small enough radius. For a point on an edge, the sphere will intersect twice, and for a degree l vertex, it will intersect l times.

However, we do not have access to the underlying graph but instead only an ϵ -sample. Our sense of local has to be significantly larger than ϵ to be geometrically measurable. This has multiple flow on effects in terms of how we analyse local environments to determine the abstract structure and how we partition the points into the vertices and edges.

Since we have a sample of points that are not a subset of the underlying graph it is nonsensical to choose whether the samples lie *on* a vertex or an edge. Instead, we decide whether they are *near* a vertex or an edge. We will classify points by the local geometry. For this we will fix a distance of 10ϵ to be our sense of local. That is for each sample point we look at all the samples within 10ϵ of it and try to understand the local geometry at this 10ϵ scale. Does this geometric structure look that that of a neighbourhood of a vertex or of an edge? This is the purpose of the dimension function within the algorithm.

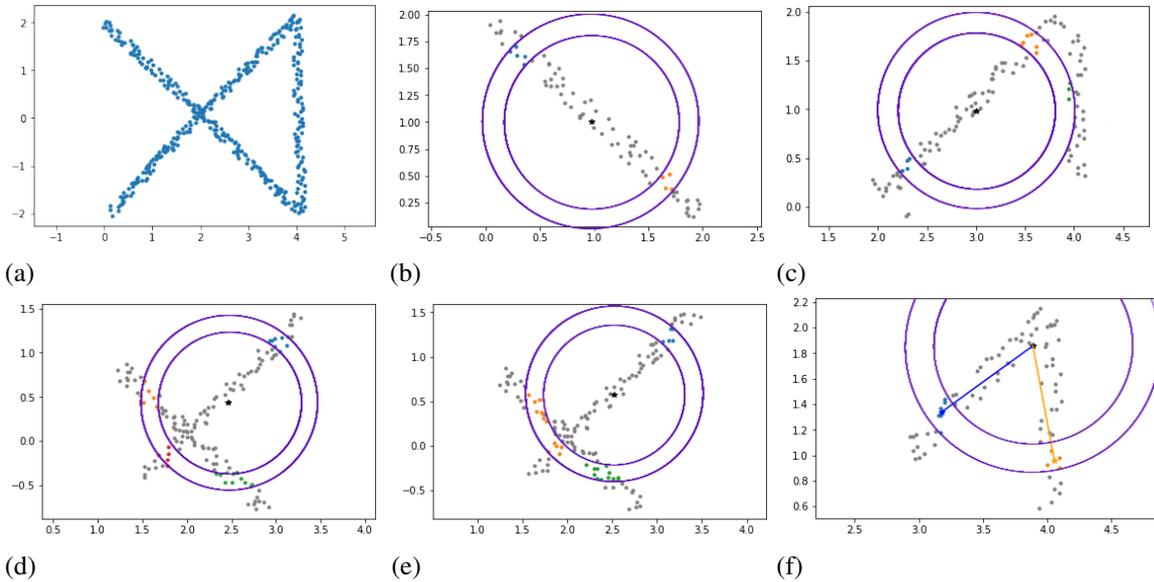


Figure 1. (a) contains the ϵ -sample. (b)-(f) illustrate different cases of local neighbourhood. The colours indicate different connected components in the annular region.

When we had the underlying graph we could measure the unique point on each nearby edge which intersected the sphere. With an ϵ -sample we have to use an annulus instead of a sphere in order to ensure that there exists a sample point for each intersection. By making the difference of the inner and outer radius to be 2ϵ (as in the region with distance between 8ϵ and 10ϵ from the sample) we can guarantee that a ball of radius ϵ around a location in the graph is contained in the annular region and hence there will be a sample point in the ϵ -sample. There can be multiple samples in this annular region for each edge and so we consider equivalence classes of points that would be samples of that part of the edge. This can be achieved by considering connected components where we connect two points when they are within 3ϵ .

There are two cases where we declare that a sample is near an edge (or rather that it is not near a vertex). The typical case is that the samples lying in an 10ϵ ball around sample q are noisy samples of a straight line that passes at most ϵ from q . This set of points are all connected at threshold 2ϵ and there are two connected components at threshold 3ϵ for the set of samples restricted to the annulus. Furthermore these connected components lie close to directly opposite each other. The angle between the centroids is at least $2 \arccos(1/4)$.

An example is shown in Figure 1 (b). The other case occurs when we are closer to a vertex where there is an acute angle between adjacent edges. Here the samples lying in an 10ϵ ball are not all connected at threshold 2ϵ . An example is shown in Figure 1 (c).

Given a lower bound of $\pi/6$ for the angle between adjacent vertices, we can guarantee one of these two cases will occur for every sample at least 15ϵ away from every vertex. Thus, each such sample will be classified as near an edge by the dimension function.

We also want to show that the dimension function declares all samples sufficiently close to a vertex are of dimension 0. Near a vertex of degree larger than 2, we will want the number of connected components to be at least 3 for all samples sufficiently close to the vertex. This will depend on a lower bound on the angle between adjacent edges and for our algorithm (with the inner radius set at 8ϵ) a sufficient condition is a lower bound of $\pi/6$. With such a bound, any sample q with $\|q - v\| < 2\epsilon$ the samples within an 10ϵ ball of q are all connected at a threshold of 2ϵ and each of the edges will contribute a different connected component within the annulus centred at q . An example is illustrated in Figure 1 (f).

Stronger statements can be made in terms of the existence of connected components representing each of a suitable subsets of edges near a sample point. This covers other cases such as in Figure 1 (d). We can guarantee that the samples declared to be of dimension 0 spread far enough to separate the points declared of dimension 1 on the different edges, so that they appear as separate clusters in $\dim^{-1}(1)$.

To detect vertices of degree 2 we have to introduce an angle condition. When the annulus has two connected components we measure the angle between the vectors pointing to the center of these two components. If we lie near an edge then the angle should be at least $2 \arccos(1/4)$. In contrast if the angle at the vertex is θ and q is within 4ϵ of both edges then the computed angle between the connected components is bounded from above by $\theta + \arcsin(1/4) + \arcsin(5/8)$. Since our algorithm declares the dimension is 0 once the angle is below $2 \arccos(1/4)$ we can guarantee identifying the abstract graph structure when $\theta \leq \pi/2$. This covers other cases such as in Figure 1 (e).

The dimension function effectively partitions the samples into those assigned to a vertex and those assigned to an edge we can then further partition the samples into the connected components with appropriate thresholds. We use 3ϵ as the threshold between samples in $\dim^{-1}(1)$ to say they are of the same edge. A low threshold is needed to ensure that we do not merge adjacent edges. We use the much larger 9ϵ when looking to see with samples in $\dim^{-1}(0)$ are connected. We can use this higher threshold because of the geometric assumptions about minimum distances between vertices.

Once we have a partition of the samples into the clusters for $\dim^{-1}(0)$ and $\dim^{-1}(1)$ we can read off the abstract graph. The $\dim^{-1}(0)$ and $\dim^{-1}(1)$ clusters are in bijection with the vertices and the edges respectively. Furthermore, for each edge $e \in E$ with corresponding $\dim^{-1}(1)$ cluster $c(e)$ there are exactly two $\dim^{-1}(0)$ clusters with points within 3ϵ of $c(e)$. These correspond to the vertices that bound e in the abstract graph.

We then wish to model the locations of the vertices as the vertices completely determine the embedding of the graph. We do this using a least squares estimator which will minimise the sums of distances from the samples to the embedded graph where this distance is the distance to a vertex if the sample is in $\dim^{-1}(0)$ and the distance is the distance to the corresponding edge if the sample is in $\dim^{-1}(1)$. It should be noted that this is naturally a biased estimator which is visible in the simulations. The bias is for vertices to drift in the direction where there are more edges. A future direction of research is to reduce, or estimate and correct for this bias.

2.1 Relationship to local homology

For Y a topological space, the homology groups $H_i(Y)$ are algebraic summaries of the topological structure of Y . We have a homology group for each dimension i : $H_0(Y)$ describing the connected components of Y and $H_1(Y)$ the space of loops. E.g. The homology of the wedge of N circles is $\text{rank}(H_1(\bigvee_N S^1)) = N$.

To measure the local topological structure at a point q we can construct a quotient space where everything far from q is quotiented to a single point. The local homology at q is defined at the limit of the homology of $Y/\{y \in Y : d(y, q) > r\}$ as r tends to 0. The case where the Y is an embedded graph then the local homology of a vertex with degree k is that of the wedge of $k - 1$ copies of S^1 . The local homology of a point along an edge is that of $\bigvee_{k-1} S^1$.

Since we do not have access to the underlying graph but instead an ϵ -sample it is not meaningful to take the limit as r goes to 0. Instead the practice in topological data analysis is to use the sample point cloud to estimate the R -local homology which uses a fixed radius $r = R$ in the construction of the relevant quotient spaces. In

the case of an embedded graph Y with all vertices separated by at least $2R$ and q a point near X , this R -local homology is that of the wedge of $k - 1$ spheres where k is the number of points in $Y \cap \partial B(q, R)$. Note that if an edge intersects twice if and only if it is its own separate connected components in $X \cap B(q, R)$. By counting the connected components in $Y \cap B(q, R)$ and $Y \cap \partial B(q, R)$ we can completely determine the topological structure of embedded graph Y near q .

3 ALGORITHM

The steps 2 through 5 of the algorithm use topological data analysis while the last step is a nonlinear least-squares fit.

Step 1: Construct $\mathfrak{G}_{3\epsilon}(X)$. This step constructs an embedded graph $\mathfrak{G}_{3\epsilon}(X)$ on the sample X , by connecting points $p, q \in X$ if $d(p, q) \leq 3\epsilon$.

Step 2: Dimension Function. Given a point $q \in X$, we construct the subgraph of $\mathfrak{G}_{3\epsilon}(X)$ generated by the points within 10ϵ of q , and restrict the edges. If this graph is disconnected, the function outputs dimension 1. Otherwise, we consider the subcomplex generated by the points in the annulus of radii 8ϵ and 10ϵ around q . If the number of connected components is not 2, we assign dimension 0, if it is 2, we apply the angle test.

1. Take the subgraph \mathfrak{G}_q of $\mathfrak{G}_{3\epsilon}(X)$ generated by points $p \in X$ such that $d(p, q) \leq 10\epsilon$.
2. Remove edges between points p, p' if $d(p, p') > 2\epsilon$.
3. For p, p' vertices in \mathfrak{G}_q , add an edge between p and p' if $d(p, p') \leq 2\epsilon$.
4. If the number of connected components in \mathfrak{G}_q is not 1, output dimension 1.
5. Else, consider subgraph $\mathfrak{G}_q^{(1)}$ of $\mathfrak{G}_{3\epsilon}(X)$ generated by $p \in \mathfrak{G}_q$ such that $d(p, q) \geq 8\epsilon$.
6. If the number of connected components in $\mathfrak{G}_q^{(1)}$ is not 2, return dimension 0.
7. Else, do Step 2a.

Step 2a: Angle test. For points p with 2 connected components in the annulus with radii $(8\epsilon, 10\epsilon)$, we calculate the angle between the components by taking the mid-point of each cluster, and calculating the angle between the lines joining p and each mid-point. If this angle is less than $2 \arccos(1/4)$, the dimension function outputs 0, otherwise it outputs 1.

1. Find average of coordinates of points in the two connected components.
2. Calculate angle between the line segments from averages to q .
3. If angle is less than $2 \arccos(1/4)$ return dimension 0.
4. Else return dimension 1.

Step 3: Finding the vertices. Having determined the dimension of each point in the sample, we consider the subcomplex generated by the points we have assigned dimension 0. Considering the connected components of this subcomplex, we obtain the clusters of points surrounding vertices.

1. Generate subgraph of $\mathfrak{G}_{9\epsilon}(X)$ on the points with dimension function 0.
2. Map each connected component of the subgraph to a vertex of an abstract graph.

Step 4: Finding the edges. Having determined the dimension of each point in the sample, we consider the subcomplex generated by the points we have assigned dimension 1. Considering the connected components of this subcomplex, we obtain the clusters of points surrounding the edges.

1. Generate subgraph of $\mathfrak{G}_{3\epsilon}(X)$ on the points with dimension function 1.
2. Map each connected component of the subgraph to a vertex of an abstract graph.

Step 5: Assigning boundary vertices to edges. For each edge cluster, we find the two vertex clusters which have points connected to those in the edge cluster in the full complex.

1. For each connected component G_i^1 in the preimage of 1 under the dimension function, we add in points in the preimage of 0 which are within 3ϵ of G_i^1 .
2. Add an edge to the abstract graph (i, j) if points from the i^{th} and j^{th} connected components in preimage of 0 are added above.

Step 6: Nonlinear least squares fitting. This is the final step in recovering the embedded graph $|G|$. Thus far, we have the abstract graph (V_0, E_0) of G . We also determined for each data point $x^{(i)}$ the stratum G_j which it is closest to. We will now determine the geometric information used to embed X in \mathbb{R}^d . For the 0-dimensional strata X_j this is just one point $x_i \in \mathbb{R}^n$. The set of 0 dimensional strata is just $\{x_j \mid j \in V_0\}$. The 1-dimensional strata are of the form $X_j = \{\theta x_{j_1} + (1 - \theta)x_{j_2} \mid \theta \in (0, 1)\}$ and the set of 1-dimensional strata is $\{\theta x_{j_1} + (1 - \theta)x_{j_2} \mid (j_1, j_2) \in E, \theta \in (0, 1)\}$. Thus we need to determine the x_j for $j \in V_0$. We will use a least-squares approach where x_j are such that they minimise the sum of Euclidean distances squared between data points and their associated strata. For the 0 dimensional strata we introduce a partial objective as

$$\phi_i(x_1, \dots, x_{k_v}, \theta_i) = \|p^{(i)} - x_{j(i)}\|^2. \quad (3)$$

Here $\theta_i = 0$. For 1-dimensional strata we include a parameter θ_i which models the local coordinate of the data point in X_j and get a partial objective

$$\phi_i(x_1, \dots, x_{k_v}, \theta_i) = \|p^{(i)} - \theta_i x_{j_1(i)} - (1 - \theta_i)x_{j_2(i)}\|^2. \quad (4)$$

We then get an optimisation problem with objective

$$\Phi(x_1, \dots, x_{k_v}, \theta_1, \dots, \theta_n) = \sum_{i=1}^n \phi_i(x_1, \dots, x_{k_v}, \theta_i). \quad (5)$$

We minimise this Φ and have the following conditions for θ_i :

- $\theta_i \in [0, 1]$
- $\theta_i = 0$ if $j(i) \in V_0$.

This optimisation problem is then solved using a code from `scipy.optimize`. It implements the trust-region reflexive method and uses a Gauss-Newton approach. For more details and references see Jones et al. (2001). As the problem is sparse and converges within a finite number of steps in practice, the complexity of this approach is approximately $O(n)$.

4 SIMULATIONS

In this section, let G be the abstract graph with vertex set $\{0, 1, 2, 3, 4\}$ and edge set $\{(0, 1), (1, 2), (2, 3), (3, 1)\}$ and $\epsilon = 0.1$. We consider two embeddings of G : $|G|_2 \subset \mathbb{R}^2$ and $|G|_3 \subset \mathbb{R}^3$. We will follow the same procedure for both. Taking an ϵ -sample X , we will partition X , and then use these partitions to model the embedding. We used Hunter (2007) for generating the images. See Figure 2 and Figure 3. We provide these figures for illustrative purposes, and the graphs were chosen to display the cases discussed earlier. We are unaware of any algorithms which both model the embedding $|G|$ and recover the abstract structure of G from an ϵ -sample X .

5 CONCLUSIONS

We have provided an algorithm that can learn embedded graphs under reasonable geometric constraints. This semiparametric modelling algorithm requires as input an ϵ -sample of the embeded graph, and consists of two main components. First, it learns the abstract graph structure (the non-parametric component), and then using non-linear least squares regression, it finds an optimal embedding of vertices (the parameters of the model). Future directions include reduction, estimation and correction of the bias, expanding the algorithm to piecewise linear embedded structures, and modifications to deal with semi-algebraic sets.

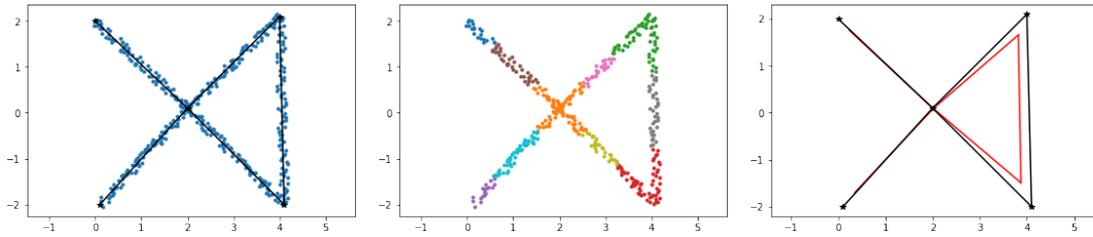


Figure 2. Left: An embedding $|G|_2$ of G in \mathbb{R}^2 (black) and an ϵ -sample X_2 of $|G|_2$ (red). Middle: partitioning of X_2 obtained from algorithm. Right: The embedded graph $|G|_2$ (black) and the model $|\tilde{G}|_2$ (red).

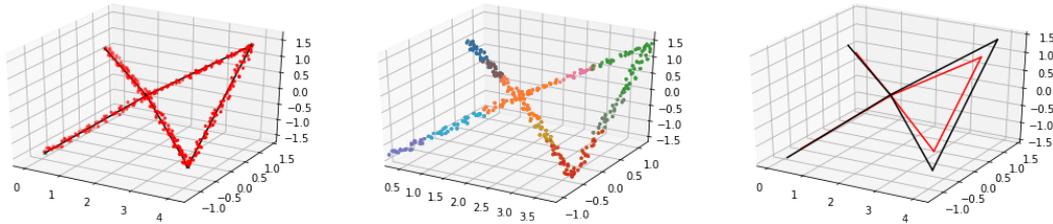


Figure 3. Left: An embedding $|G|_3$ of G in \mathbb{R}^3 (black) and an ϵ -sample X_3 of $|G|_3$ (red). Middle: partitioning of X_3 obtained from algorithm. Right: The embedded graph $|G|_3$ (black) and the model $|\tilde{G}|_3$ (red).

ACKNOWLEDGEMENT

The first author was supported by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship through the Australian National University. The first author would also like to thank James Morgan for various helpful discussions.

REFERENCES

- Bendich, P., B. Wang, and S. Mukherjee (2012). Local homology transfer and stratification learning. In *Proceedings 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1355–1370.
- Brown, A. and B. Wang (2018). Sheaf-Theoretic Stratification Learning. In B. Speckmann and C. D. Tóth (Eds.), *34th International Symposium on Computational Geometry (SoCG 2018)*, Volume 99 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, pp. 14:1–14:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
- Hughes, B. and S. Weinberger (2001). Surgery and stratified spaces. *Surveys on surgery theory*, 319–352.
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 9(3), 90–95.
- Jones, E., T. Oliphant, P. Peterson, et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed 19. July 2019].
- Nanda, V. (2019). Local cohomology and stratification. *Foundations of Computational Mathematics*.