Manifold Modelling with Minimum Spanning Trees

D. M. Bot[1] P. Huo[2] A. Arleo[3] F. Paulovich[3] and J. Aerts[1,2,*]

[1]Data Science Institute (DSI), Hasselt University, Belgium
[2]Biosystems Department, KU Leuven, Belgium
[3]Department of Mathematics and Computer Science, TU Eindhoven, The Netherlands
[*]Corresponding author: jan.aerts@kuleuven.be

Abstract

Recent dimensionality reduction algorithms operate on a manifold assumption and expect data to be uniformly sampled from that underlying manifold. While some algorithms attempt to be robust for non-uniform sampling, their reliance on k-nearest neighbours to approximate manifolds limits how well they can span sampling gaps without introducing shortcuts. We present a minimum-spanning-tree-based manifold approximation approach that overcomes this problem and demonstrate it crosses sampling-gaps without introducing shortcuts while creating networks with few edges. A python package implementing our algorithm is available at `https://github.com/vda-lab/multi_mst`.

CCS Concepts

• ***Computing methodologies*** → *Dimensionality reduction and manifold learning;*

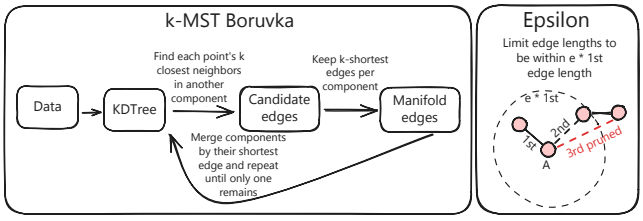
1. Introduction

Dimensionality reduction techniques are commonly used to visualise and explore multidimensional data. Classical techniques—such as PCA and MDS—operate globally by attempting to preserve variance or all pairwise distances. More recent techniques operate on a manifold assumption, preserving locality and neighbourhoods instead (e.g., [RS00, Ten00, vdMH08]). These algorithms first approximate the manifold using an undirected graph and then compute a layout that preserves the graph's structure.

Generally, such dimensionality reduction algorithms assume data is uniformly sampled from an underlying manifold (e.g., [BN03]), meaning the complete manifold is observed and sampled without gaps. While UMAP is designed to be robust against this assumption [MHM18], there are scenarios in which more than a k-nearest neighbour network (k-NN) is needed to approximate a manifold (f.i., Figures 2a and 2b). This poster presents a k-nearest Minimum Spanning Tree (k-MST) manifold approximation approach that can deal with such non-uniform sampling. Our main research question is: How do k-NNs and k-MSTs compare in modelling non-uniformly sampled manifolds?

2. k-nearest Minimum Spanning Tree (k-MST)

The k-nearest Minimum Spanning Tree (k-MST) is inspired by Pathfinder networks that, with a specific parameter selection, yield the union set of all possible MSTs in a network (e.g., [QCGB*08, AKM17]). We generalise k-nearest neighbour networks to minimum spanning trees, to make MST unions work for distances



(a) (b)

Figure 1: ***(a)*** *The Boruvka algorithm adapted to find the k-shortest edges between connected components. (**b**) A distance threshold* ε *provides an upper distance limit for the* 2*-to-k additional edges.*

which may have few identically weighted connections. We believe using more than only the MST (as in, f.i., [DTS*20]) aids the preservation and interpretation of local structure.

Our implementation adapts fast HDBSCAN's version of the Boruvka Algorithm [MC23] (Figure 1a). It relies on a KDTree to find each point's k-nearest neighbours that are not in the same connected component. Then, the shortest k candidates per component are added to k-MST. Only the shortest edge is used to update the connected component to ensure the algorithm finds all edges included in a normal MST. This process repeats until one connected component remains. Since data points start as distinct components, all k-NN edges are included in the k-MST.

A distance threshold ε can be applied to avoid creating shortcuts or sparsify the manifold (Figure 1b). The parameter is specified as a fraction of the shortest edge between components and provides

DOI: 10.2312/evp.20241088

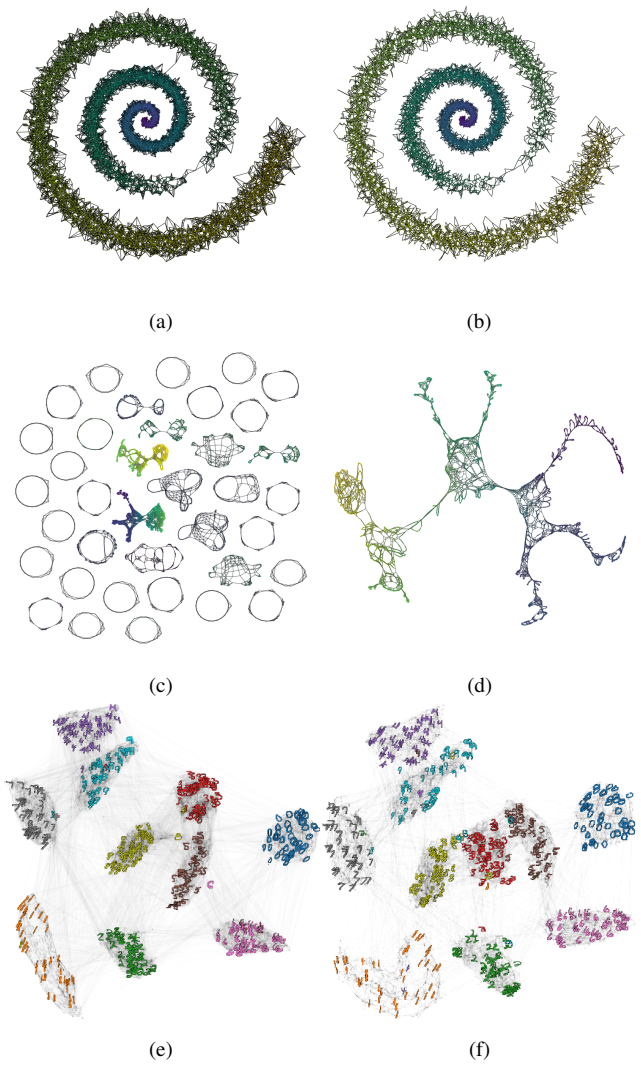


Figure 2: *Top-down views of a non-uniformly sampled 3D Swiss roll for a (**a**) 5-NN and (**b**) 2-MST. Force directed layouts [Hu05] of a horse-shaped mesh reconstruction dataset [SP04] for a (**c**) 5-NN and (**d**) 3-MST. UMAP's cross-entropy optimised layouts [MHM18] (spectral initialisation, 200 epochs with repulsion strength* 0.1*, 300 epochs with repulsion strength* 1.0*) of MNIST [LBBH98] for a (**e**) 5-NN and (**f**) 2-MST with* $\varepsilon = 1.1$*. All edges are drawn using Datashader [BCT*23] with 45 random observations per digits drawn as the graph's nodes.*

an upper distance limit for the 2-to-k alternative edges. A similar threshold for k-NNs also avoids shortcuts, but prevents crossing observation gaps.

3. Demonstration

We present three cases to compare k-MSTs with k-NNs (Figure 2). k-NNs were computed as UMAP graphs, using NNDescent to find approximate nearest neighbours [DML11]. Compute times on an 8-core 3.8 GHz CPU are listed in each case.

The first case demonstrates the methods' ability to cross sampling gaps on a non-uniformly sampled Swiss roll (22.196 observations, 3 features). The data was constructed from lengths l and depths d: $x = sl^2\cos(l)$, $y = sl^2\sin(l)$, $z = d$, $s = 0.03$. Gaussian noise was added to all coordinates scaled by the length: $\sigma = 0.0395l$. Specific depth—length samples were removed between $l = 16.2$ and $l = 19.8$ to create a small observation gap along the manifold. The 5-NN was computed in 124 ms and crosses the gap but also introduces a shortcut (Figure 2a). The 2-MST required 105 ms to compute and recovers a small manifold approximation graph without shortcuts (Figure 2b).

The second case quantifies the methods' quality as dimensionality reduction approach by the Sortedness [PSNCP23] of their force directed layouts [Hu05] on a horse-shaped mesh reconstruction dataset [SP04] (8.431 observations, 3 features). The 5-NN was computed in 68 ms and does not recover a single connected manifold (Figure 2c). It has a Sortedness of 0.72. At 10 neighbours, the k-NN's Sortedness improves to 0.89, but the manifold remains disconnected and details in local structures start to obscure. The 3-MST required 40 ms to compute and recovers a small, connected manifold with a high Sortedness: 0.90 (Figure 2d).

The third case demonstrates the methods' behaviours and graph sizes on a dataset with clusters: MNIST [LBBH98] (70.000 observations, 784 features). The 5-NN required 1.91 s to compute and recovers a fully connected graph containing 427.046 edges (Figure 2e). The 2-MST with $\varepsilon = 1.1$ was computed in 647 s and recovers the data's structure using only 201.032 edges (Figure 2f).

4. Discussion and Future Work

Our method's main benefit over regular k-nearest neighbour networks is its improved handling of non-uniformly sampled manifolds. The k-MST can span sparse regions without introducing shortcuts. This property is also helpful to create models of datasets that are better interpreted as multiple distinct manifolds. In that case, our approach captures the distance and direction between these separate entities.

Using minimum spanning trees as a basis makes the k-MST discover connectivity at all (relevant) distance scales. Capturing this longer-range connectivity at low values of k balances the local and global structure—which is desirable in dimensionality reduction (e.g., [MHM18, MvDW*17])—and results in small models that are cheap to lay out. We speculate it also reduces the method's sensitivity to scale compared to k-NN based approaches such as UMAP and t-SNE, which would simplify the parameter tuning stage.

A limitation of our implementation is the compute cost associated with KDTrees on some data sets. A NNdescent-based approximation for MSTs could make our techniques computationally competitive for larger datasets.

5. Acknowledgements

This work was supported by Hasselt University BOF grant [BOF20OWB33] and KU Leuven grant ITP-E5160-STG/23/040.

References

[AKM17] ARLEO A., KWON O. H., MA K. L.: GraphRay: Distributed pathfinder network scaling. *2017 IEEE 7th Symp. Large Data Anal. Vis. LDAV 2017 2017-Decem* (2017), 74–83. doi:10.1109/LDAV.2017.8231853. 1

[BCT*23] BEDNAR J. A., CRAIL J., THOMAS I., CRIST-HARIF J., RUDIGER P., BRENER G., B C., MEASE J., SIGNELL J., LIQUET M., STEVENS J.-L., COLLINS B., HANSEN S. H., THUYDOTM, THORVE A., ESC, KBOWEN, ABDENNUR N., SMIRNOV O., MAIHDE, HAWLEY A., ORIEKHOV A., AHMADIA A., JR B. A. B., BRANDT C. H., TOLBOOM C., G. E., WELCH E., BOURBEAU J., SCHMIDT J. J.: holoviz/datashader: Version 0.16.0, Oct. 2023. URL: https://doi.org/10.5281/zenodo.10044690, doi:10.5281/zenodo.10044690. 2

[BN03] BELKIN M., NIYOGI P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput. 15*, 6 (2003), 1373–1396. doi:10.1162/089976603321780317. 1

[DML11] DONG W., MOSES C., LI K.: Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proc. 20th Int. Conf. World wide web* (New York, NY, USA, mar 2011), ACM, pp. 577–586. URL: https://dl.acm.org/doi/10.1145/1963405.1963487, doi:10.1145/1963405.1963487. 2

[DTS*20] DORAISWAMY H., TIERNY J., SILVA P. J. S., NONATO L. G., SILVA C.: TopoMap: A 0-dimensional Homology Preserving Projection of High-Dimensional Data. *IEEE Trans. Vis. Comput. Graph.* (2020), 1–1. URL: https://ieeexplore.ieee.org/document/9222271/, arXiv:2009.01512, doi:10.1109/TVCG.2020.3030441. 1

[Hu05] HU Y. W. R. I.: Efficient and High Quality Force-Directed Graph Drawing. *Math. J. 10*, 1 (2005), 37–71. 2

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proc. IEEE 86*, 11 (1998), 2278–2324. URL: http://ieeexplore.ieee.org/document/726791/, doi:10.1109/5.726791. 2

[MC23] MCINNES L., CONTRIBUTORS: Fast HDBSCAN (version 0.1.2), 2023. URL: https://github.com/TutteInstitute/fast_hdbscan/releases. 1

[MHM18] MCINNES L., HEALY J., MELVILLE J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL: http://arxiv.org/abs/1802.03426, arXiv:1802.03426. 1, 2

[MvDW*17] MOON K. R., VAN DIJK D., WANG Z., GIGANTE S., BURKHARDT D. B., CHEN W. S., YIM K., VAN DEN ELZEN A., HIRN M. J., COIFMAN R. R., IVANOVA N. B., WOLF G., KRISHNASWAMY S.: Visualizing Structure and Transitions for Biological Data Exploration. *bioRxiv* (2017), 1–92. doi:10.1101/120378. 2

[PSNCP23] PEREIRA-SANTOS D., NEVES T. T. A. T., CARVALHO A. C. P. L. F. D., PAULOVICH F. V.: Nonparametric Dimensionality Reduction Quality Assessment based on Sortedness of Unrestricted Neighborhood. In *EuroVis Workshop on Visual Analytics (EuroVA)* (2023), Angelini M., El-Assady M., (Eds.), The Eurographics Association. doi:10.2312/eurova.20231093. 2

[QCGB*08] QUIRIN A., CORDÓN O., GUERRERO-BOTE V. P., VARGAS-QUESADA B., MOYA-ANEGÓN F.: A quick MST-based algorithm to obtain Pathfinder networks $(\infty, n-1)$. *Journal of the American Society for Information Science and Technology 59*, 12 (2008), 1912–1924. doi:https://doi.org/10.1002/asi.20904. 1

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science (80-.). 290*, 5500 (dec 2000), 2323–2326. URL: https://www.science.org/doi/10.1126/science.290.5500.2323, doi:10.1126/science.290.5500.2323. 1

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. *ACM Trans. Graph. 23*, 3 (aug 2004), 399–405. URL: https://dl.acm.org/doi/10.1145/1015706.1015736, doi:10.1145/1015706.1015736. 2

[Ten00] TENENBAUM J. B.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science (80-.). 290*, 5500 (dec 2000), 2319–2323. doi:10.1126/science.290.5500.2319. 1

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *J. Mach. Learn. Res. 9* (nov 2008), 2579–2625. 1