Hotspot identification for Mapper graphs

Ciara F. Loughrey[1], Nick Orr[2], Paweł Dłotko[3], Anna Jurek-Loughrey[1],

[1]School of Electronics, Electrical Engineering and Computer Science, [2]Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Northern Ireland

[3] Dioscuri Centre in Topological Data Analysis, Mathematical Institute, Polish Academy of Sciences Warsaw, Poland

Research Contribution

- Mapper graphs provide a handy way to detect anomalous regions within a dataset. There are however two issues;

1. Manual analysis of large graphs is prohibitive
2. Selecting appropriate parameters for Mapper construction that reveal the hotspots in data is nontrivial. Numerous lens functions may need to be considered

- To address this challenge we propose a new technique for automatic detection of hotspots in the Mapper graph.
- Using a real-world breast cancer dataset, we demonstrate how the proposed algorithm can be used to automatically select a lens function based on its ability to discriminate subgroups of patients that present increased survival outcomes.

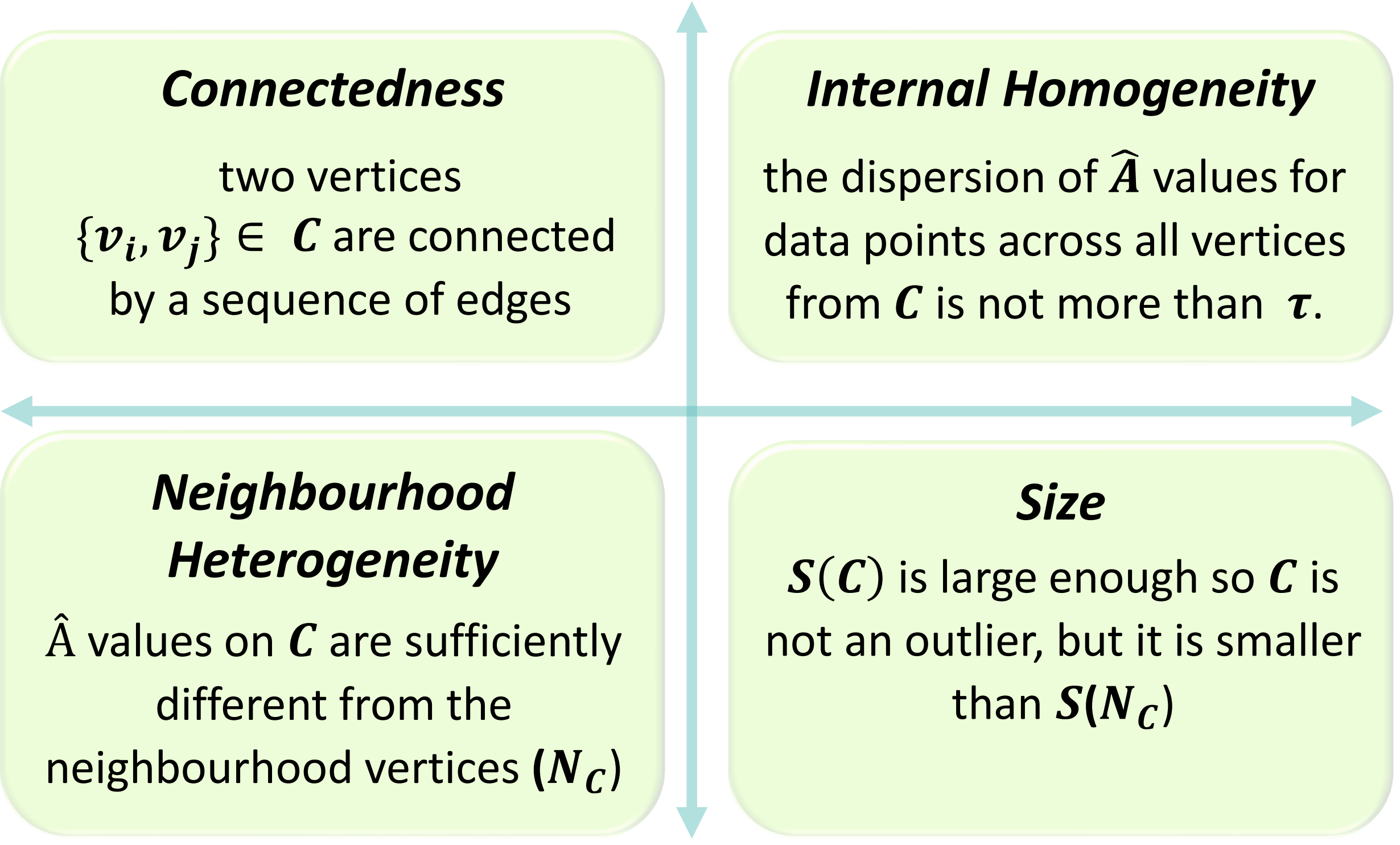
Problem Statement

We address the problem of detecting regions within a Mapper graph that are structurally coherent and homogeneous on a value attribute of interest, while also differing sufficiently from its neighbourhood within graph.

We search for the presence of hotspots that may exist within larger graph communities. This is relevant to the task of patient stratification in biomedicine, where the identification of small patient groups that show contrasting survival patterns within larger disease subtypes can support improved diagnosis and treatment outcomes.

Point cloud	$X = \{x_1, \ \dots, x_k\}, \ x_i \in R^n$
Attribute (e.g. survival)	$A: X \to R$
Mapper Graph	$G = < V, E >$
Induced attribute function	$\hat{A}: X \to R$

$\boldsymbol{C} \subset \boldsymbol{V}$ will be called a hotspot within $\boldsymbol{G}$ with respect to $\hat{\boldsymbol{A}}$ if the following conditions holds:



Connectedness

two vertices $\{v_i, v_j\} \in \boldsymbol{C}$ are connected by a sequence of edges

Internal Homogeneity

the dispersion of $\hat{\boldsymbol{A}}$ values for data points across all vertices from $\boldsymbol{C}$ is not more than $\boldsymbol{\tau}$.

Neighbourhood Heterogeneity

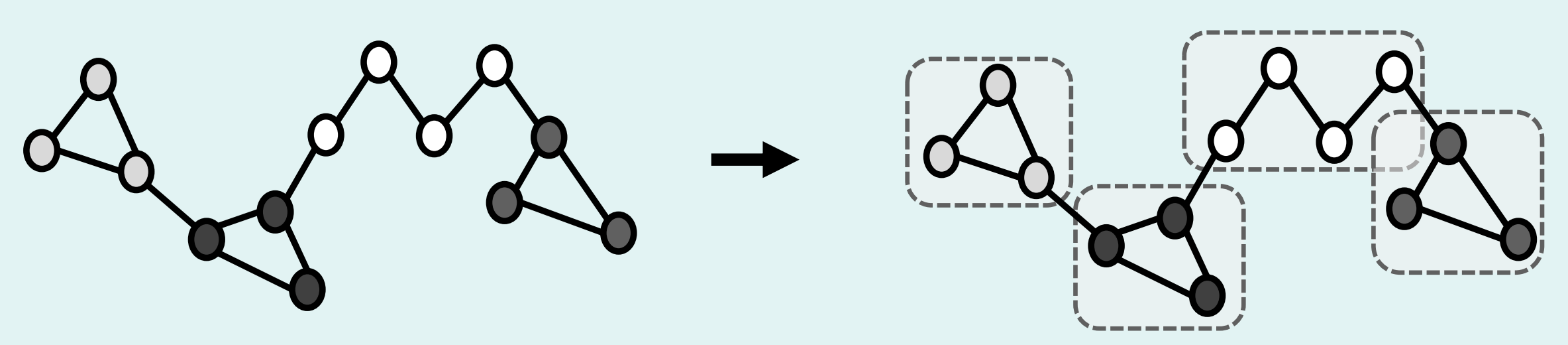
$\hat{A}$ values on $\boldsymbol{C}$ are sufficiently different from the neighbourhood vertices $(\boldsymbol{N_C})$

Size

$\boldsymbol{S(C)}$ is large enough so $\boldsymbol{C}$ is not an outlier, but it is smaller than $\boldsymbol{S(N_C)}$

Step 1: Cluster Detection

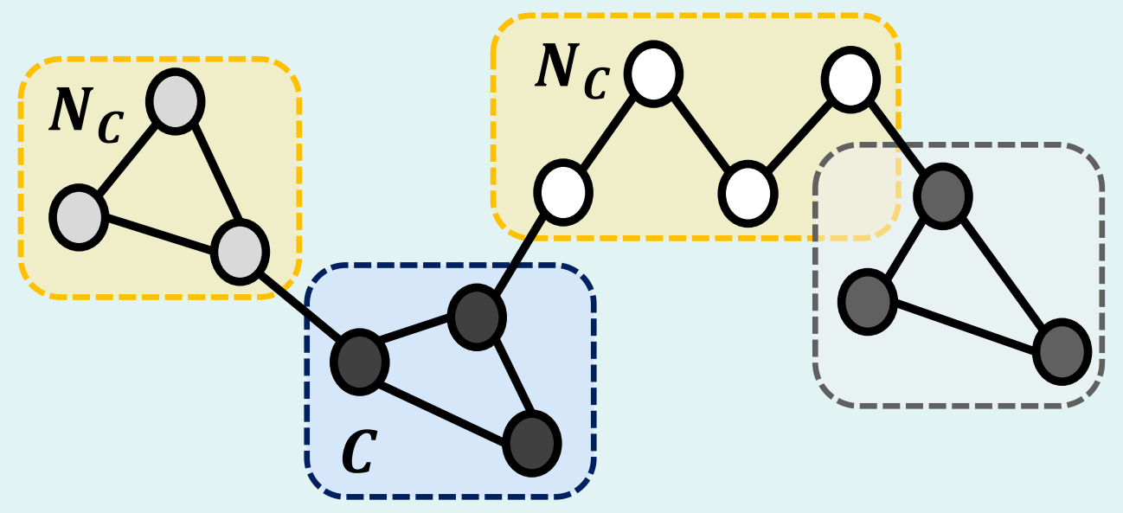
Non-intersecting connected components of $\boldsymbol{G}$ that are homogeneous with respect to $\hat{A}$ are chosen.



1. Define the edge weight according to the absolute difference in $\hat{A}$ values of vertices
2. Perform single linkage on the vertices
3. Identify all connected components $\boldsymbol{C_1}, \dots, \boldsymbol{C_n}$ that are connected in the dendrogram below the level $\boldsymbol{\tau}$.

Step 2: Cluster Classification

Each component is classified as either hotspot or non-hotspot.



1. For each $\boldsymbol{C}$ calculate size $\boldsymbol{S(C)}$ of the neighbourhood $\boldsymbol{N_C}$
2. If $\boldsymbol{S(C)} < \ \boldsymbol{\sigma_1}$ or $| \boldsymbol{S(N_C)} - \boldsymbol{S(C)}| \ < \ \boldsymbol{\sigma_2}$ then classify $\boldsymbol{C}$ as a non-hotspot. $\boldsymbol{S}$ can consider nodes or samples.
3. For each $\boldsymbol{C}$ calculate $\hat{A}(\boldsymbol{N_C})$ as the mean value of $\hat{A}$ across all vertices within $\boldsymbol{N_C}$
4. If $\left|\hat{A}(\boldsymbol{C}) \ - \ \hat{A}(\boldsymbol{N_C})\right| \ > \ \boldsymbol{\epsilon}$ then $\boldsymbol{C}$ is considered as a hotspot.

- We assume that $\boldsymbol{\tau}$,$\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma_1}$ should be set by the user as they strongly depend on the domain and $\hat{A}$.
- We propose for $\boldsymbol{\sigma_2}$ to be calculated as one median absolute deviation of $\{\boldsymbol{S(C_i)}\dots\boldsymbol{S(C_n)}\}$.

Datasets

2-D two circle	A Mapper graph is built with a lens function of L2norm, 7 intervals, 20% overlap and agglomerative clustering using wards linkage and 6 clusters per interval
Complex 2-D graph ***Complex 3-D graph***	A sufficiently dense 2-D or 3-D grid of points was selected. All neighbouring grid elements were connected and a smaller number of connections between random vertices was added
TCGA Breast Cancer	• Gene expression dataset for 1027 patients • DSGA transformed[2] to 1146 genes

Results

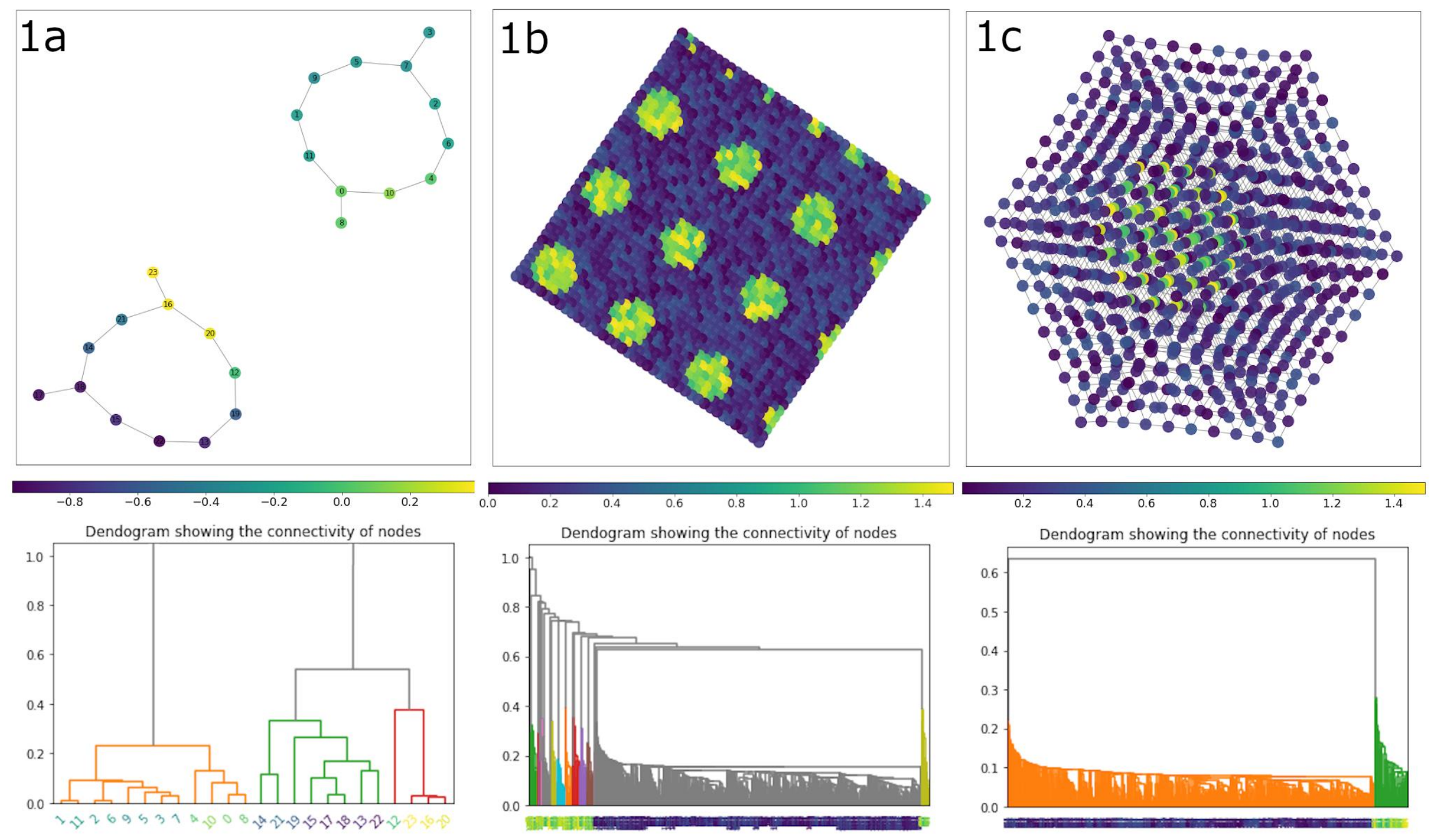


Figure 1: Three artificial graphs and the corresponding dendrograms **1a)** Two circle d**1b)** 2-D dataset **1c)** 3-D dataset

To build the $\hat{A}$ values for each artificial dataset:

Minimum value for each vector in the two-circle dataset
In the 2-D instance a correction of 1 is added to all the grid points (x, y) for which $\sin x + \ \sin y \ \geq 1$.
In the 3-D case the correction is added to all grid points (x, y, z) for which $x^2 + y^2 + zz \leq 1$.
The TCGA graph was coloured by survival outcome of patients

- In the two circles, out of three connected components found, the smaller yellow region of high values was deemed a hotspot
- In the 2-D graph, 9 hotspots were identified.
- We found the algorithm was sensitive to the $\boldsymbol{\sigma_1}$ parameter. Setting too low a value resulted in larger numbers of potential outliers.
- A single hotspot was detected in the 3-D graph.
- Within the artificial graphs, no false positives or false negatives were detected by the hotspot analysis.
- The $\boldsymbol{\epsilon}$ parameter required some exploration to ensure we chose a stringent threshold at which to consider a candidate a hotspot.

We built a Mapper graph on the TCGA data that revealed a hotspot of improved survival. Using the same Mapper parameters, we reconstructed the graph according to many different lens functions based on random feature combination, trying to find the same (or better) hotspot using our proposed algorithm. We ran this search 1000 times.

Table 1: Results and the parameter settings for the hotspot algorithm hotspot detection on the artificial datasets.

Dataset	σ_1 (nodes)	ϵ	Hotspot Count
Two circles	2	0.1	1
2-D graph	15	0.01	9
3-D graph	10	0.01	1

Results

As a criteria for hotspot detection in the TCGA dataset, we set $\boldsymbol{\sigma_1}$ at 5 for nodes, $\boldsymbol{\sigma_1}$ at 10 for samples and $\boldsymbol{\epsilon}$ at 0.15

- For every type of lens function, we were able to find at least one graph with a hotspot region of increased survival.
- A filter function based on linear combination of a subset of features revealed interesting results (Figure 2).
- 95% percent of the patients within this hotspot (n=20) were found to have an ER+ status, and consisted solely of Her2, LumA, and LumB subtypes. This indicates an unusual cluster of patients similar to that found in Nicolau et al. (2011)[2].

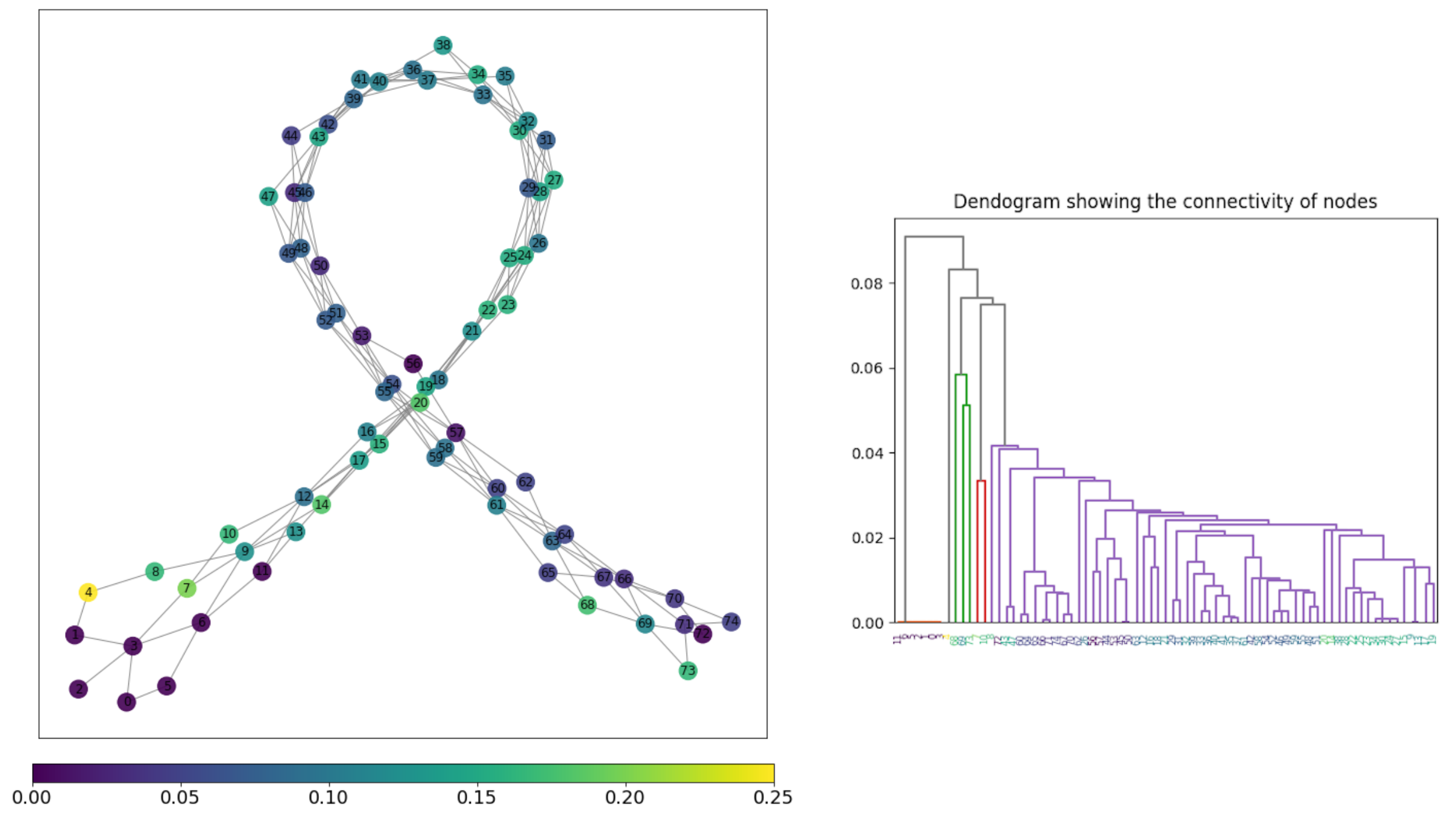


Figure 2: Mapper graph constructed from the TCGA dataset using the proposed algorithm, with the corresponding dendrogram of node connectivity based on edge weights

Future Work

To further evaluate the algorithm, we will:

- Biologically validate the quality of the retrieved hotspots
- Investigate the presence of these hotspots in a secondary female breast cancer dataset.
- Explore the problem of overfitting while sampling from a space of lenses with the objective of hotspot detection.
- Investigate how variation in Mapper parameters affects results

Acknowledgements: This work was sponsored by a PhD studentship from the Northern Ireland Department for the Economy. PD acknowledge the support of Dioscuri program initiated by the Max Planck Society, jointly managed with the National Science Centre (Poland), and mutually funded by the Polish Ministry of Science and Higher Education and the German Federal Ministry of Education and Research.

References:

1. The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490(7418), 61.
2. Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences, 108(17), 7265–7270.