UNIVERSITY OF Southampton

Jet Clustering with Spectral Clustering

Henry Day-Hall[1]
Supervisors: Prof. Claire Shepherd-Themistocleous[1,2], Prof. Stefano Moretti[1], Prof. Srinandan Dasmahapatra[1], Dr. Emmanuel Olaiya[2]

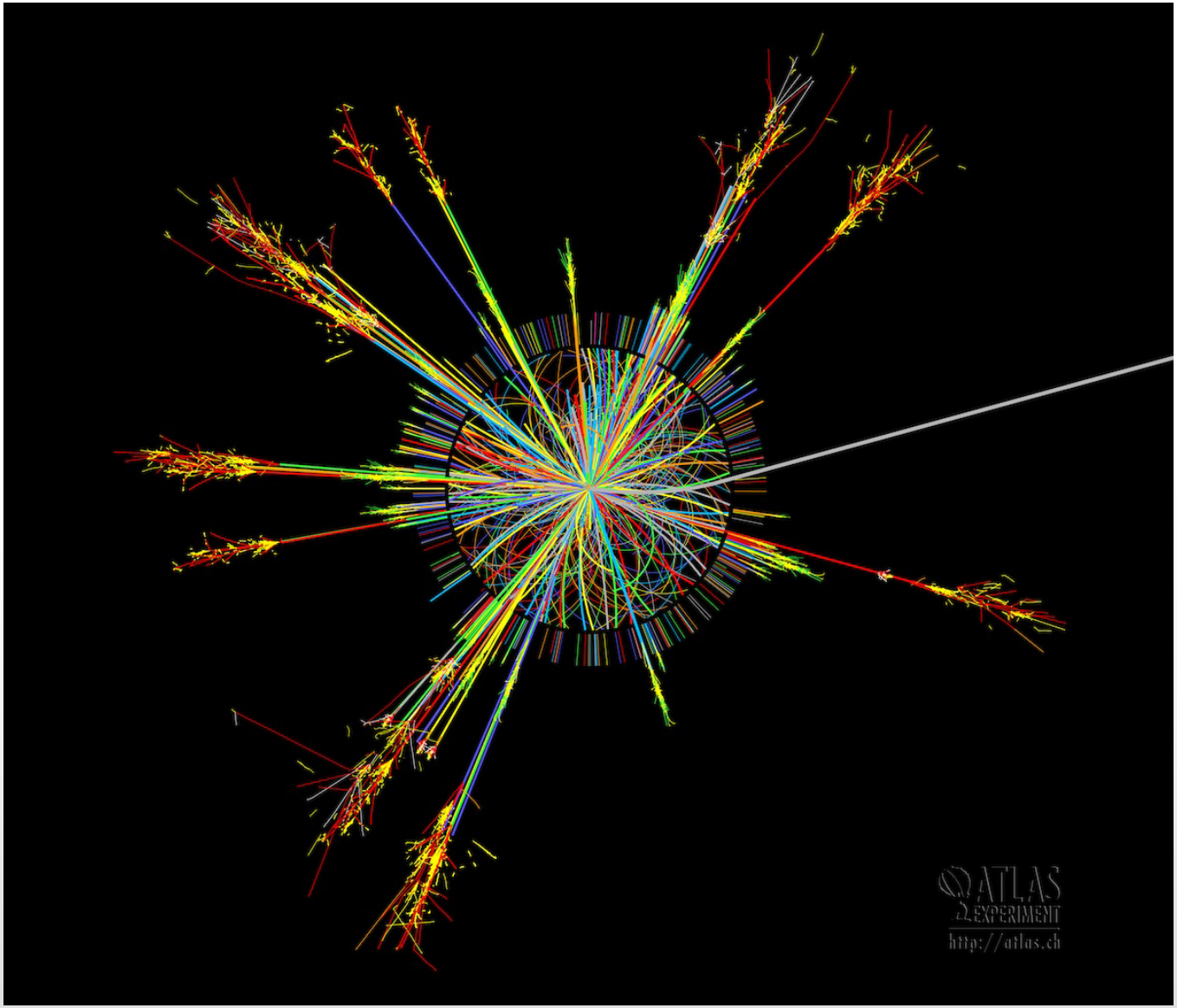
[1]University of Southampton, UK
[2]Rutherford Appleton Laboratory, UK

April 29, 2020

Table of Contents

UNIVERSITY OF Southampton

Jets



Physics Objective

UNIVERSITY OF Southampton

- A good jet clustering algorithm will accurately match the kinematics of the partons chosen as tags.

Physics Objective

UNIVERSITY OF Southampton

- A good jet clustering algorithm will accurately match the kinematics of the partons chosen as tags.
- This accuracy should vary smoothly with the cut-off parameter.

Physics Objective

UNIVERSITY OF Southampton

- A good jet clustering algorithm will accurately match the kinematics of the partons chosen as tags.
- This accuracy should vary smoothly with the cut-off parameter.
- The jets formed should replicate the mass of the partons in the hard interaction.

Clustering in ML

UNIVERSITY OF Southampton

Many attempts have been made to write a 'good' clustering algorithm. Most of them are not hierarchical, they are based on fitting a predefined model. This poses a challenge for jet clustering, we do not have a predefined number of clusters.

Clustering comparison

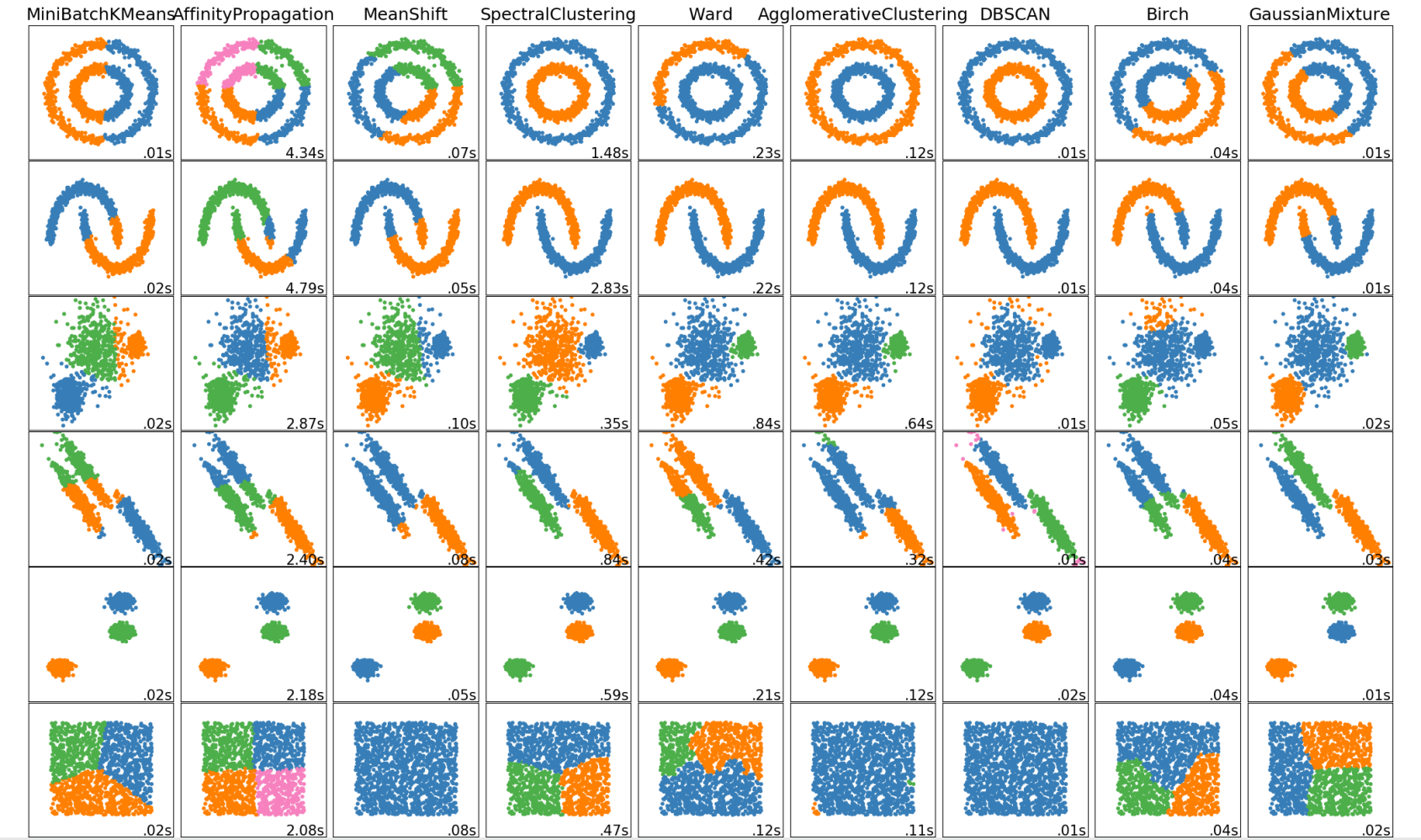
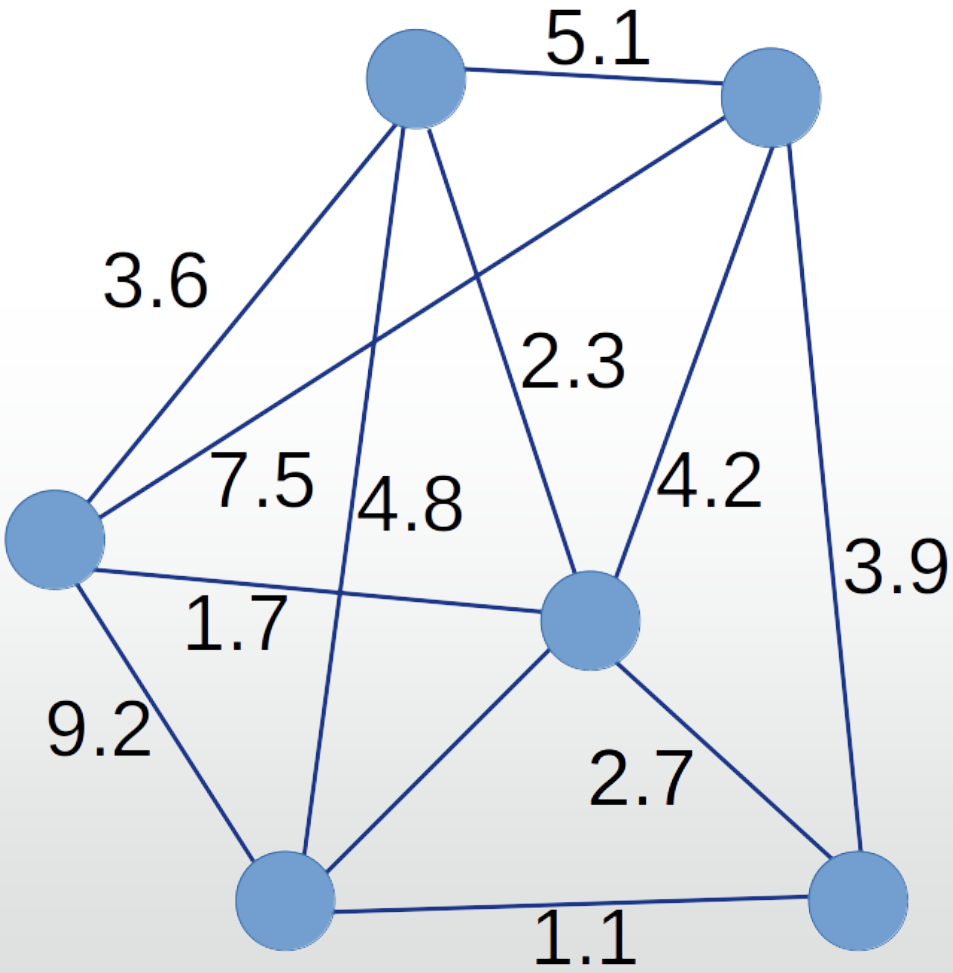


Figure: Taken from
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-nee

Aim of clustering

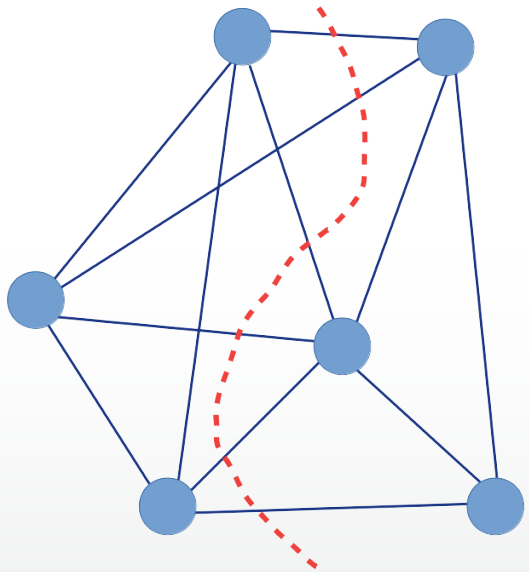
Let our points be nodes of a graph and the vertices carry a measure of the affinity, $a_{i,j}$.



Aim of clustering

UNIVERSITY OF Southampton

We wish to split the points such that the severed affinities are minimised.



Often the optimum split by this metric will isolate one point. To avoid this small clusters are penalised.

Aim of clustering

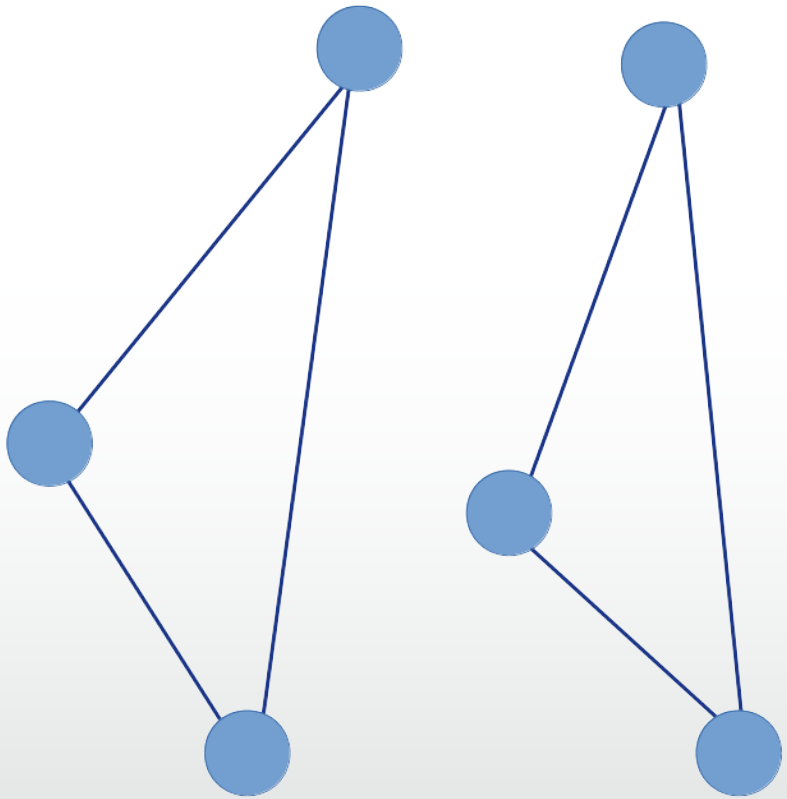
UNIVERSITY OF Southampton

These criteria result in RatioCut. If $W(A, B) = \sum_{i \in A, j \in B} a_{i,j}$ is the sum of the affinities that cross from A to B, and $|A|$ is the number of nodes in A;

$$\text{RatioCut}(A_1, A_2, \ldots A_n) \equiv \frac{1}{2} \sum_{i=1}^{n} \frac{W(A_i, \bar{A}_i)}{|A_i|}$$

In the case of disconnected components (with zero affinity between clusters) this can be solved for with the eigenvalues of the matrix known as the graph Laplacien.

Let us imagine a graph, disconnected in n clusters.



Let us imagine a graph, disconnected in n clusters. Membership of cluster k is determined by the indicator vector h_k;

$$h_{i,k} = \begin{cases} 1/\sqrt{|A_k|}, & \text{if point } i \in A_k \\ 0, & \text{otherwise} \end{cases}$$

The graph is represented by the graph Laplacien;

$$L = \begin{bmatrix} \sum a_{1,i} & -a_{1,2} & -a_{1,3} & \dots \\ -a_{1,2} & \sum a_{2,i} & -a_{2,3} & \\ -a_{1,3} & -a_{2,3} & \sum a_{3,i} & \\ \vdots & & & \ddots \end{bmatrix}$$

Then

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

Let us imagine a graph, disconnected in n clusters. Membership of cluster k is determined by the indicator vector h_k;

$$h_{i,k} = \begin{cases} 1/\sqrt{|A_k|}, & \text{if point } i \in A_k \\ 0, & \text{otherwise} \end{cases}$$

The graph is represented by the graph Laplacien;

$$L = \begin{bmatrix} \sum a_{1,i} & -a_{1,2} & -a_{1,3} & \dots \\ -a_{1,2} & \sum a_{2,i} & -a_{2,3} & \\ -a_{1,3} & -a_{2,3} & \sum a_{3,i} & \\ \vdots & & & \ddots \end{bmatrix}$$

Then

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

Let us imagine a graph, disconnected in n clusters. Membership of cluster k is determined by the indicator vector h_k;

$$h_{i,k} = \begin{cases} 1/\sqrt{|A_k|}, & \text{if point } i \in A_k \\ 0, & \text{otherwise} \end{cases}$$

The graph is represented by the graph Laplacien;

$$L = \begin{bmatrix} \sum a_{1,i} & -a_{1,2} & -a_{1,3} & \dots \\ -a_{1,2} & \sum a_{2,i} & -a_{2,3} & \\ -a_{1,3} & -a_{2,3} & \sum a_{3,i} & \\ \vdots & & & \ddots \end{bmatrix}$$

Then

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

$$h'_k L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

Then stack the of all clusters together

$$h'_k L h_k = (H' L H)_{kk}$$

and the RatioCut aim described earlier is the trace;

$$\text{RatioCut}(A_1, A_2, \ldots A_n) \equiv \frac{1}{2} \sum_{i=1}^{n} \frac{W(A_i, \bar{A}_i)}{|A_i|} = \text{Tr}(H' L H)$$

Where $H'H = I$. Trace minimsation in this form is done by finding the eigenvectors of L with smallest eigenvalues.
Generalising this to a graph that is not disconnected is just relaxing the requirements on the form of the indicator vectors; h_k.

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

Then stack the of all clusters together

$$h_k' L h_k = (H' L H)_{kk}$$

and the RatioCut aim discribed earlier is the trace;

$$\text{RatioCut}(A_1, A_2, \ldots A_n) \equiv \frac{1}{2} \sum_{i=1}^{n} \frac{W(A_i, \bar{A}_i)}{|A_i|} = \text{Tr}(H' L H)$$

Where $H'H = I$. Trace minimsation in this form is done by finding the eigenvectors of L with smallest eigenvalues.
Generalising this to a graph that is not disconnected is just relaxing the requirements on the form of the indicator vectors; h_k.

$$h_k' L h_k = \frac{1}{|A_k|} \sum_{i \in A_k, j \in A_k} \left(\delta_{i,j} \sum_l a_{l,i} - a_{i,j} \right) = \frac{W(A_k, \bar{A}_k)}{|A_k|}$$

Then stack the of all clusters together

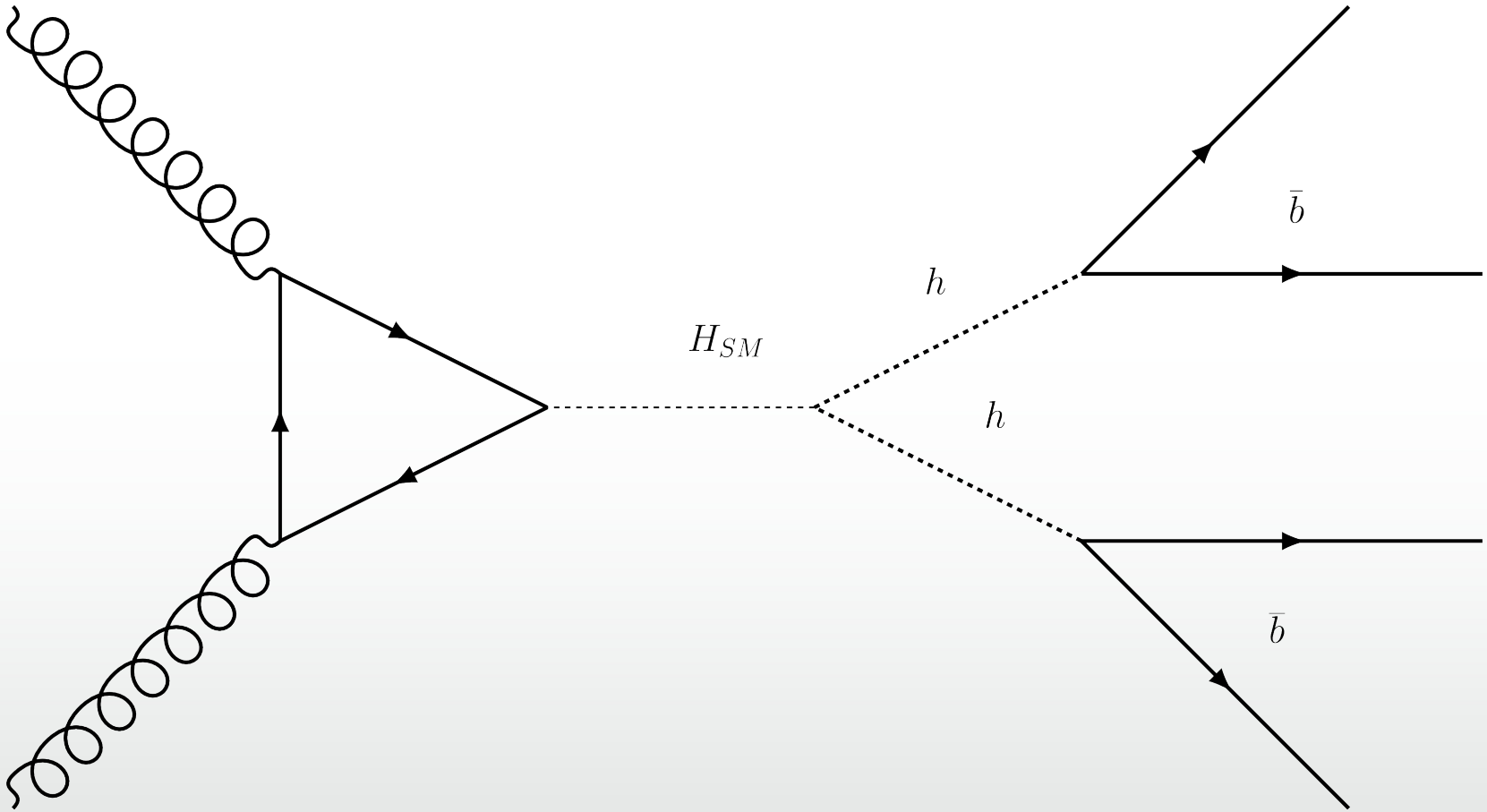
$$h_k' L h_k = (H' L H)_{kk}$$

and the RatioCut aim discribed earlier is the trace;

$$\text{RatioCut}(A_1, A_2, \ldots A_n) \equiv \frac{1}{2} \sum_{i=1}^{n} \frac{W(A_i, \bar{A}_i)}{|A_i|} = \text{Tr}(H' L H)$$

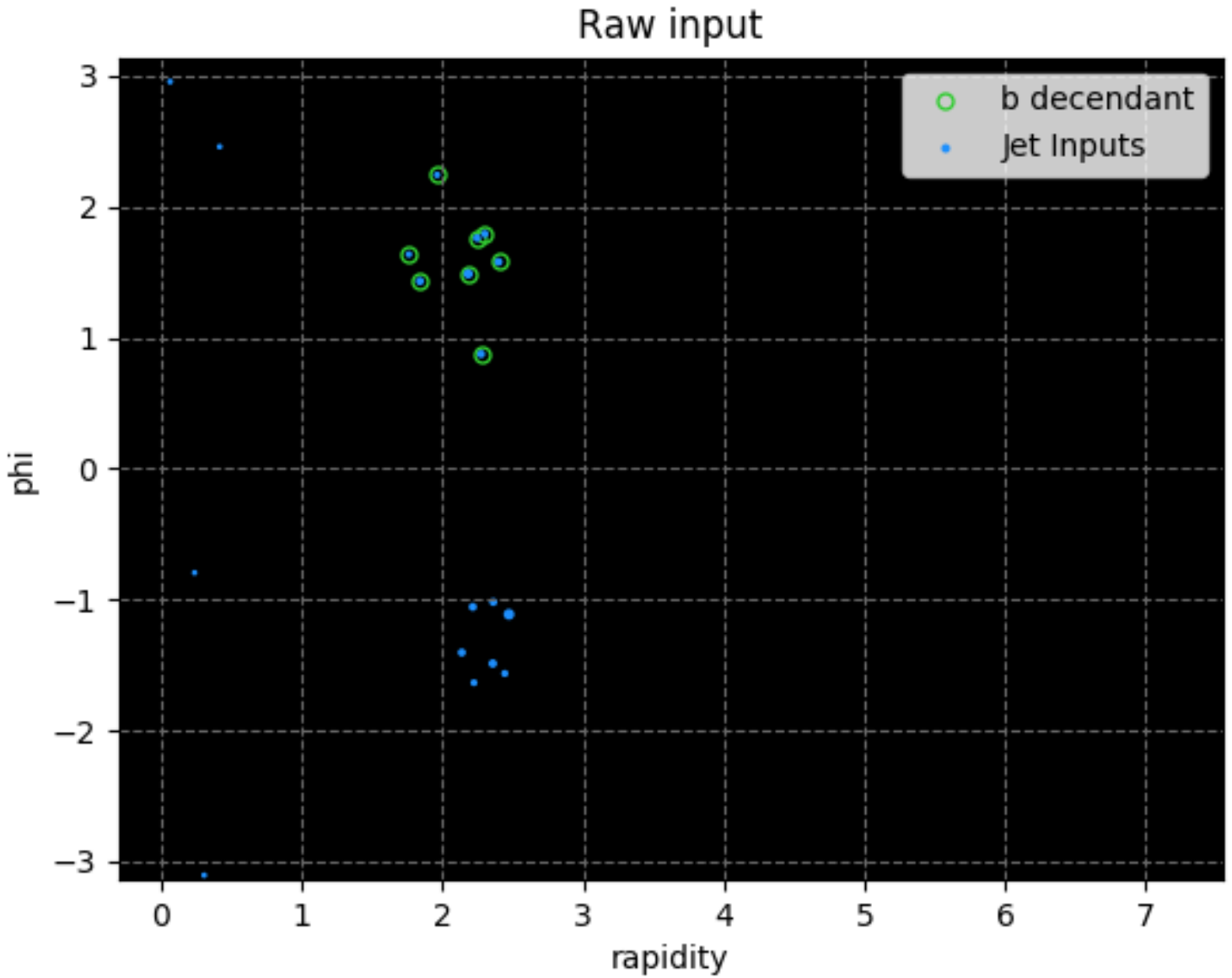
Where $H'H = I$. Trace minimsation in this form is done by finding the eigenvectors of L with smallest eigenvalues.
Generalising this to a graph that is not disconnected is just relaxing the requirements on the form of the indicator vectors; h_k.

Practice



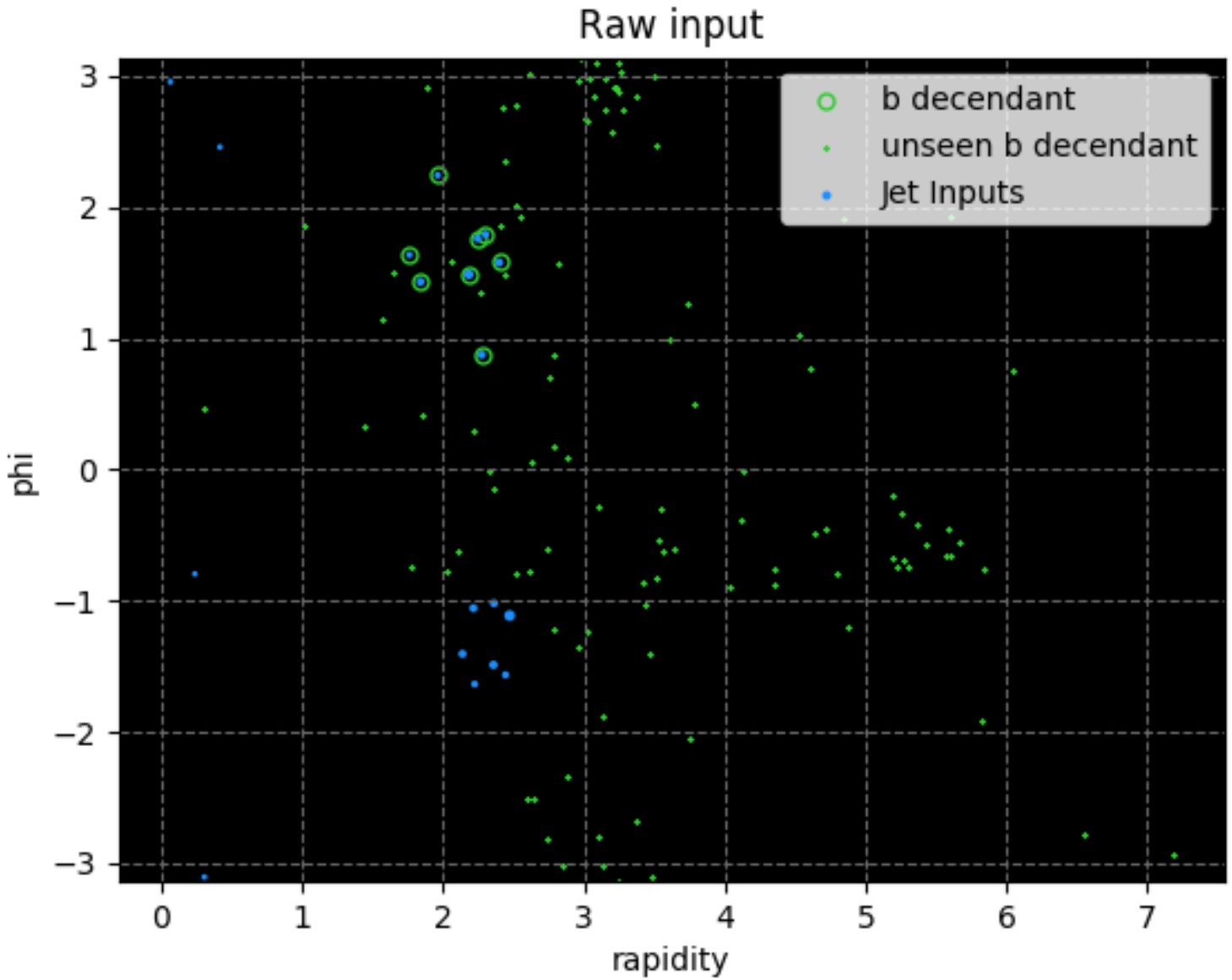
Practice

UNIVERSITY OF Southampton

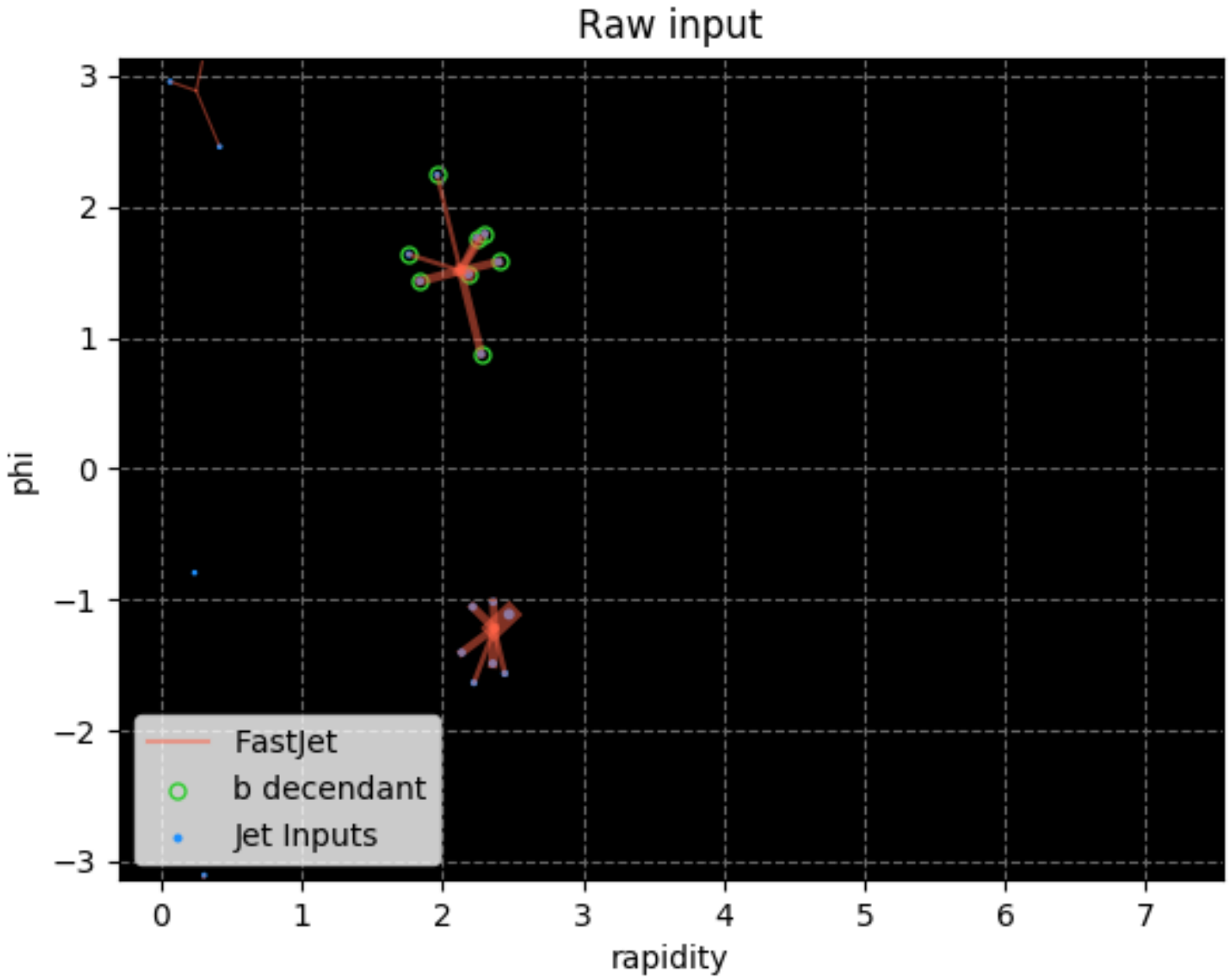


Practice

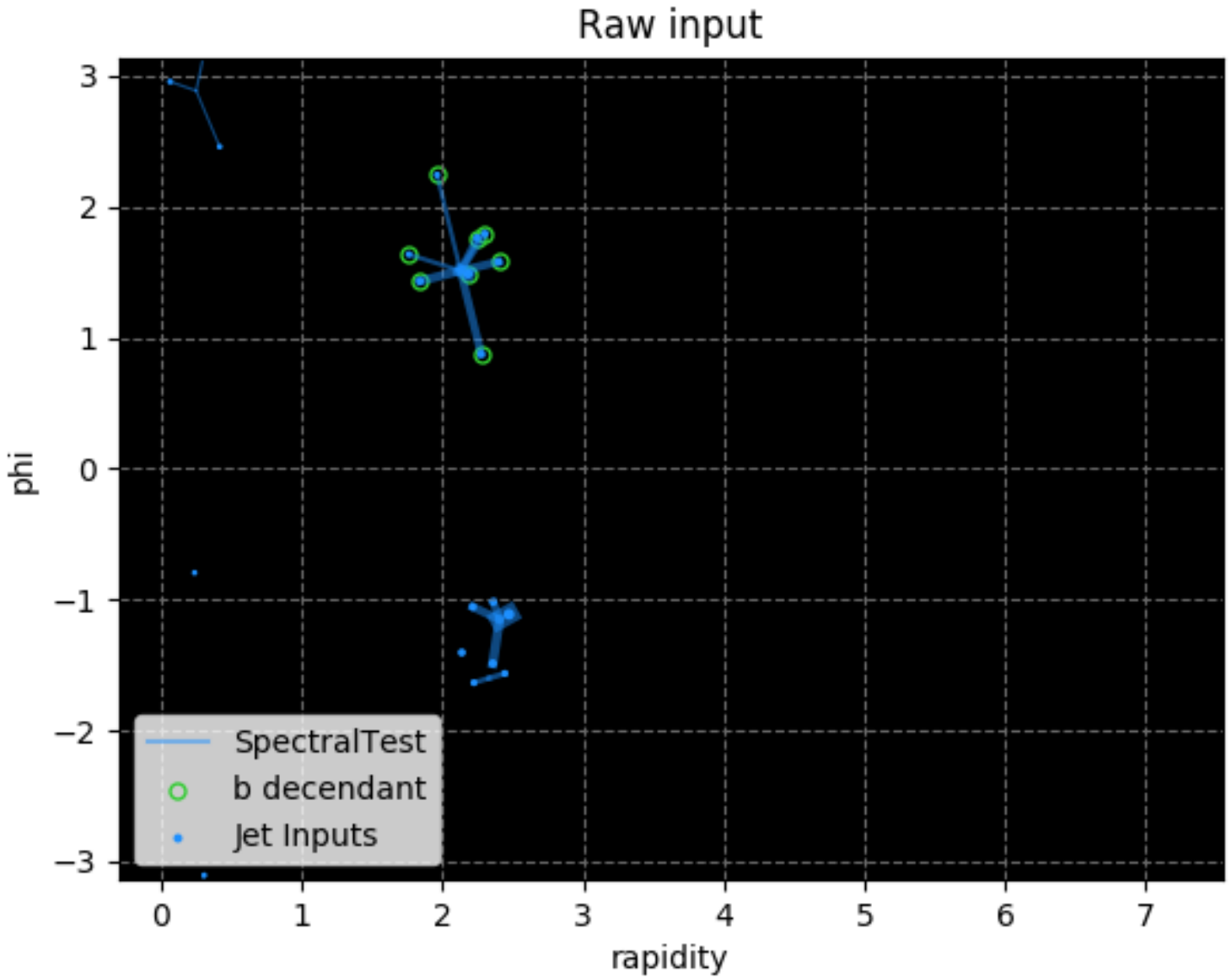
UNIVERSITY OF Southampton



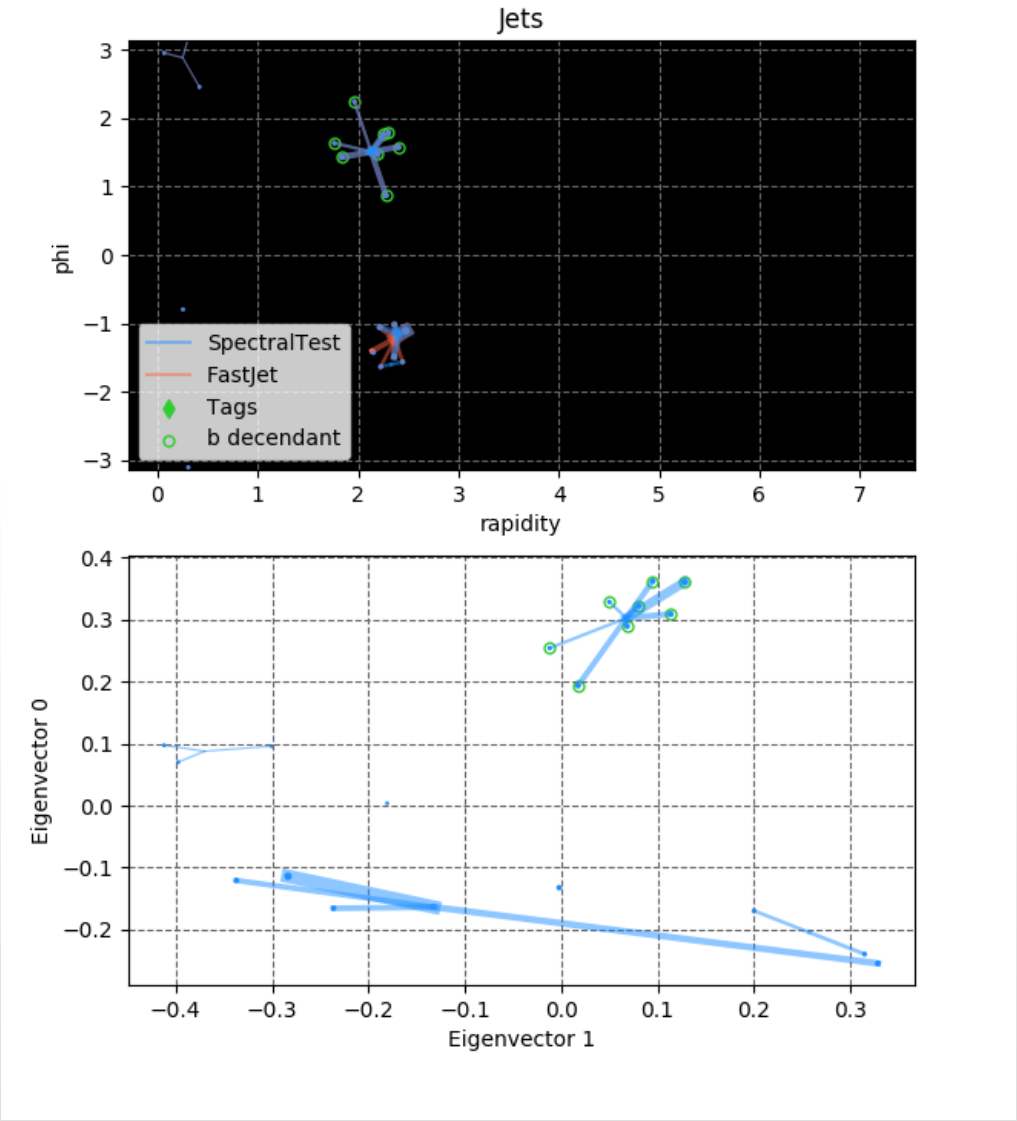
Practice


Raw input
phi
rapidity
FastJet
b decendant
Jet Inputs
3
2
1
0
−1
−2
−3
0
1
2
3
4
5
6
7

Practice

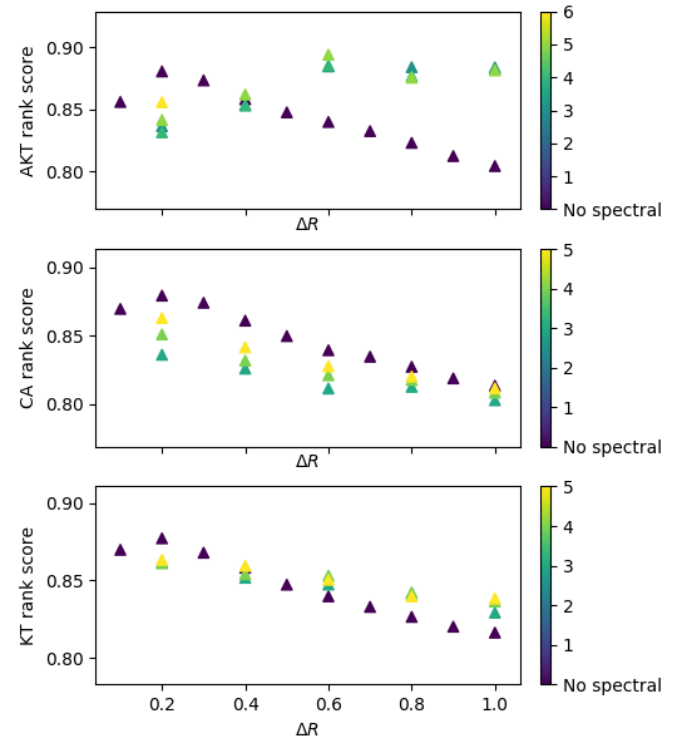

Raw input
phi
rapidity
3
2
1
0
−1
−2
−3
0
1
2
3
4
5
6
7
SpectralTest
b decendant
Jet Inputs

Practice


Jets
phi
rapidity
SpectralTest
FastJet
Tags
b decendant
Eigenvector 0
Eigenvector 1

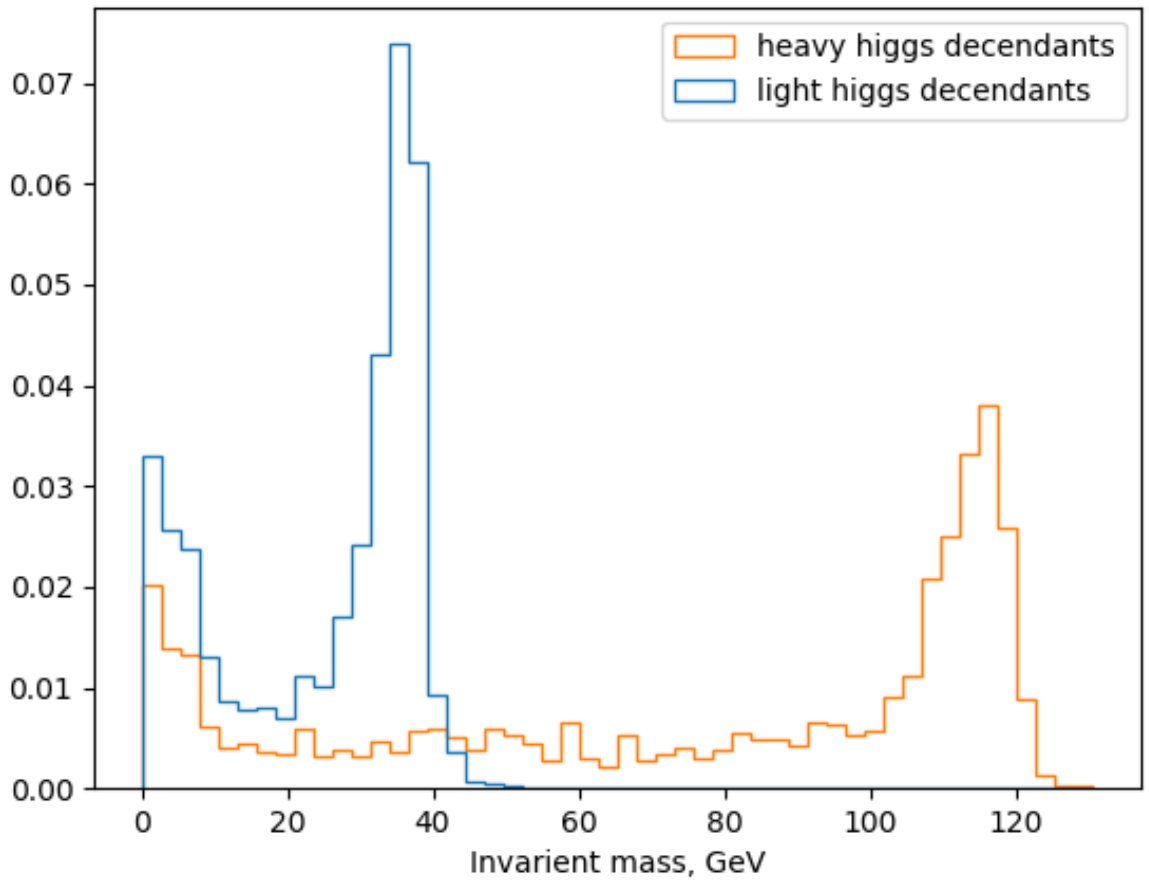
Results

A jet clustering algorithm that matches the kinematics of the tagging partons is good. Spectral clustering can do this quite well, for example here is a comparison of the p_T of the jets compared to the tagging partons.



Results

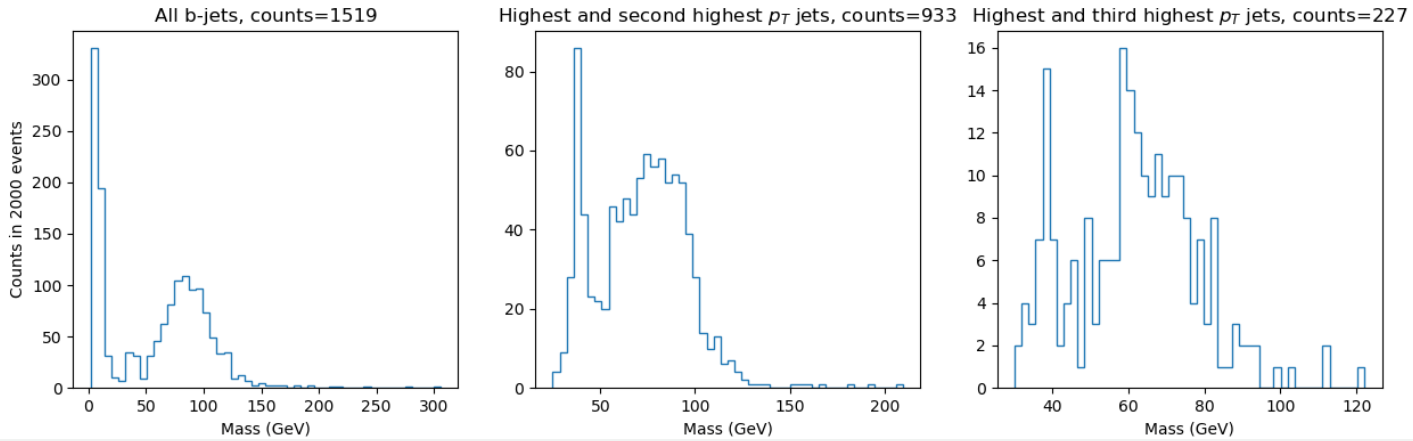
The data was generated from a $125GeV$ Higgs decaying into 2 $40GeV$ higgs. Cuts on p_T and η shift the mass peaks that could be reconstructed.



Results

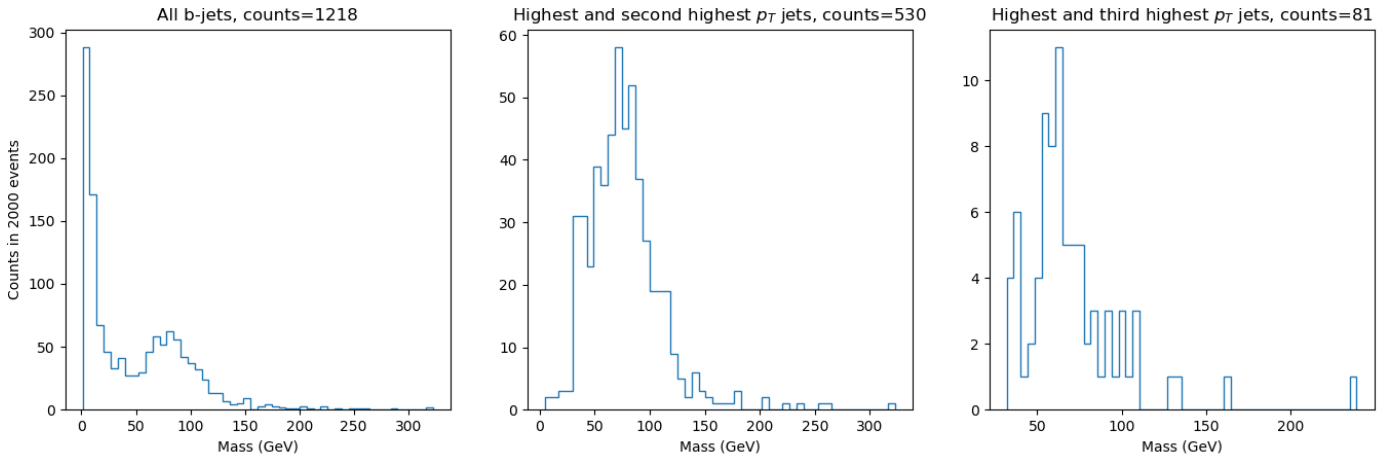
UNIVERSITY OF Southampton

The Cambridge Aachen algorithms recreates the first peak (from one light higgs) very well, and the second peak from the heavy Higgs a little;



Results

Spectral clustering is not producing such recognisable peaks.



I am not sure why this is failing to reconstruct peaks.

Conclusions

UNIVERSITY OF Southampton

This clustering method is interesting in theory and well motivated, and it can accurately reconstruct the kinematics of the partons creating the shower. However its failure to reconstruct mass peaks found by Cambridge on Monte Carlo data is frustrating and it more investigation is needed to see why it fails on this point.

Thank you for listening.