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]

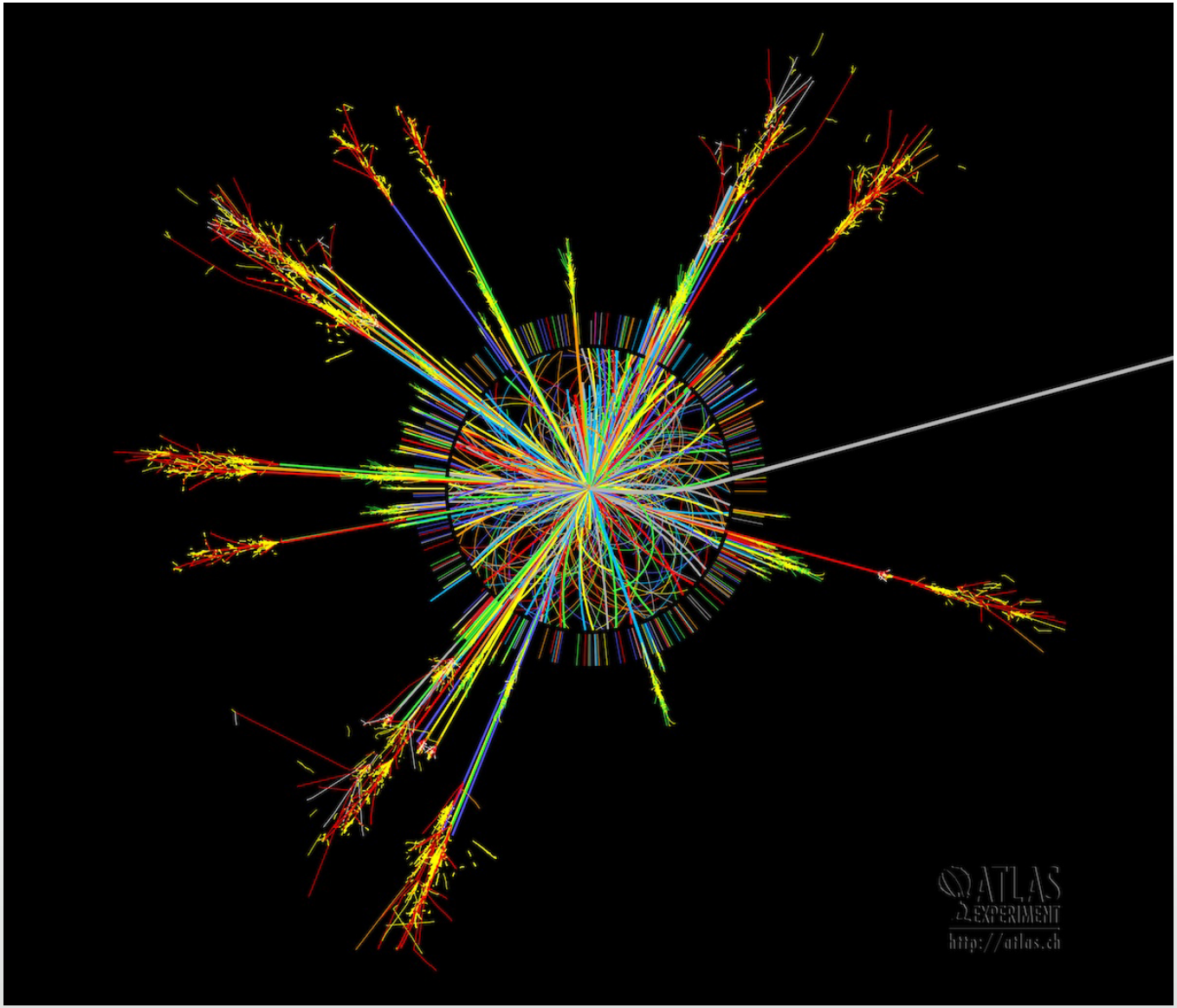
[1]University of Southampton, UK
[2]Rutherford Appleton Laboratory, UK

January 6, 2020

Table of Contents

UNIVERSITY OF Southampton

Jets



Physics Objective

UNIVERSITY OF Southampton

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

Physics Objective

UNIVERSITY OF Southampton

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

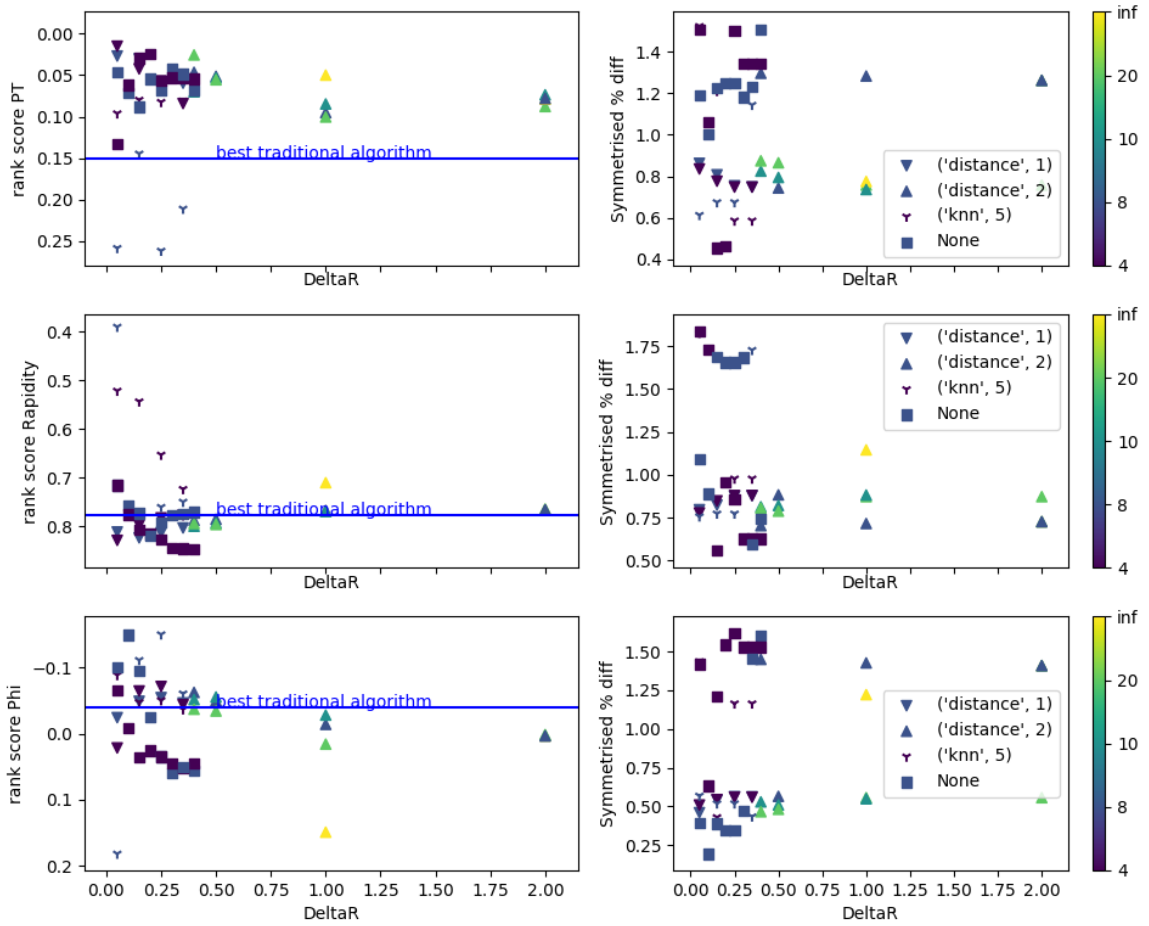
Physics Objective

UNIVERSITY OF Southampton

- A good jet cluttering 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 higher level shape variables.

Results

jet class = Spectral
0.0 < mean num jets < 20.0
Laplacien = unnormalised
AffinityType = linear
WithLaplacienScaling = False
ExponentMultiplier = 1



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

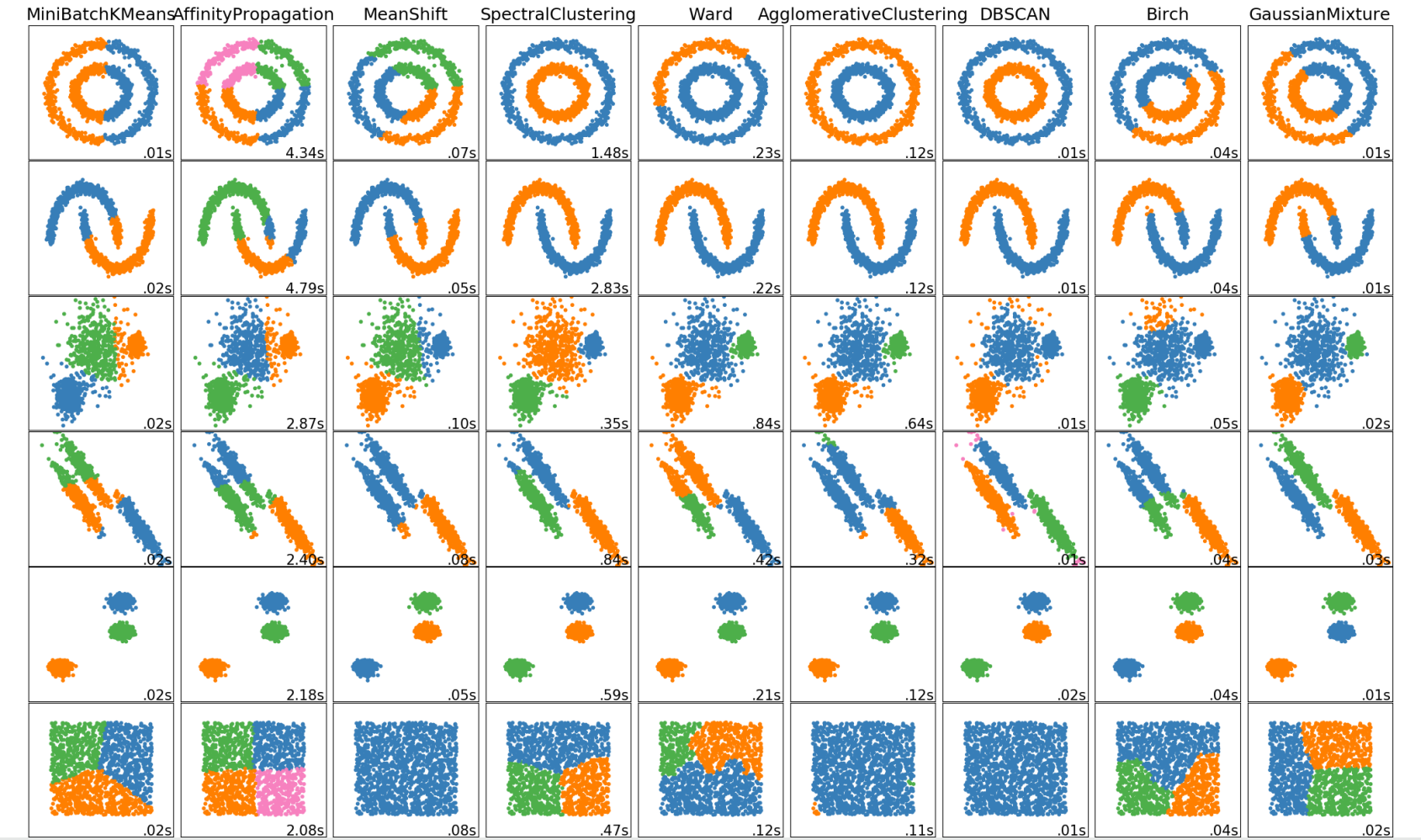
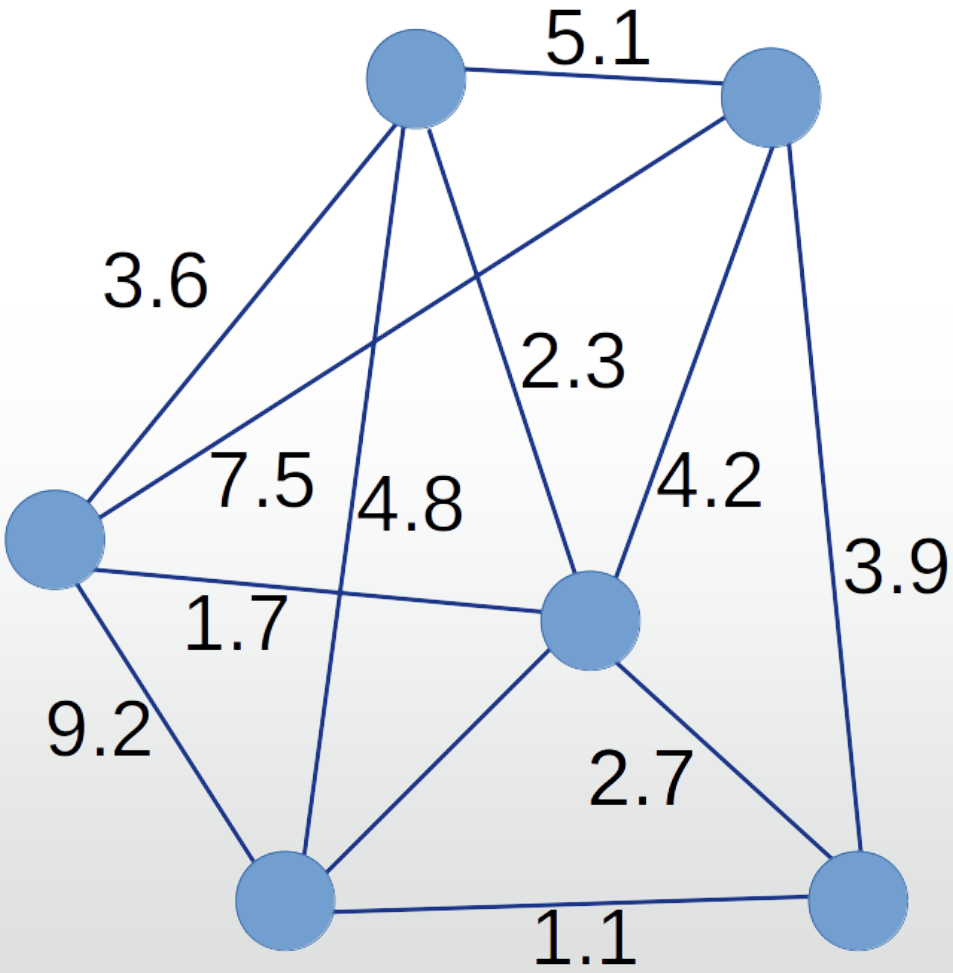


Figure: Taken from
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-nee

Aim of clustering

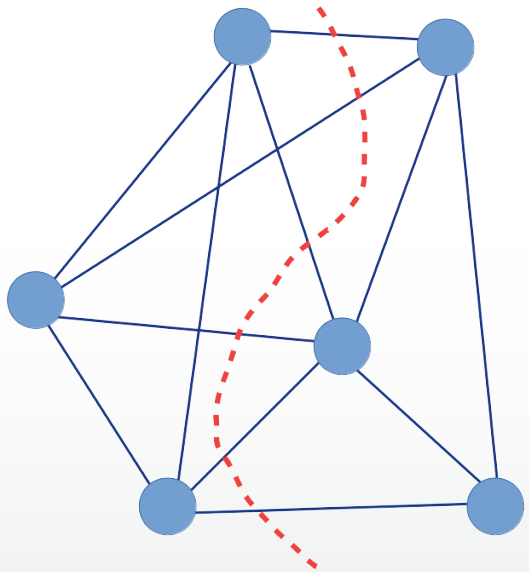
Let our points be nodes of a graph and the vertices carry a measure of the affinity, $a_{i,j}$.



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

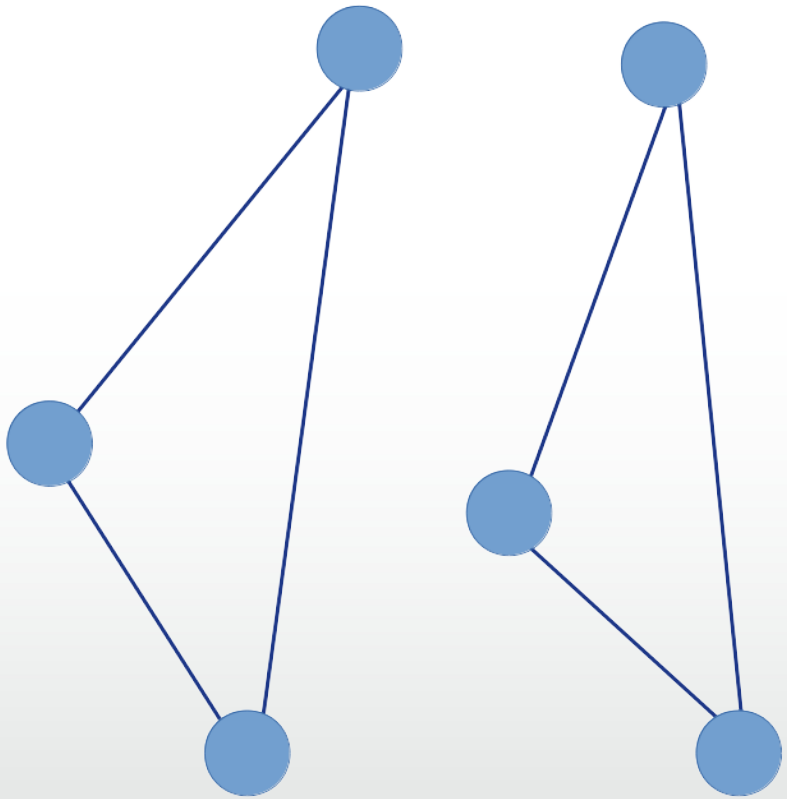
These criteria result in RatioCut. If $W(A, B) = \sum_{i \in A, j \in B} a_{i,j}$ is the sum of the affinities that cross from A to B, and $|A|$ is the number of nodes in A;

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

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.

Ideal case

Let us imagine a graph, disconnected in n clusters.



Ideal case

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

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

Ideal case

UNIVERSITY OF Southampton

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

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

Ideal case

UNIVERSITY OF Southampton

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

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

Ideal case

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

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.

Ideal case

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

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.

Ideal case

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

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.

Process

UNIVERSITY OF Southampton

To find n clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.

Process

UNIVERSITY OF Southampton

To find n clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.
2. Construct the graph Laplacien;

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

Process

UNIVERSITY OF Southampton

To find n clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.
2. Construct the graph Laplacien;

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

3. Calculate the eigenvectors v of L corresponding to the $n+1$ smallest eigenvalues.

Process

UNIVERSITY OF Southampton

To find n clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.
2. Construct the graph Laplacien;

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

3. Calculate the eigenvectors v of L corresponding to the $n+1$ smallest eigenvalues.
4. Stack the eigenvectors (aside from the first) v into a matrix E that is n by m. Call E the eigenspace, each point in the original dataset is represented by one row.

Process

UNIVERSITY OF Southampton

To find n clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.
2. Construct the graph Laplacien;

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

3. Calculate the eigenvectors v of L corresponding to the $n+1$ smallest eigenvalues.
4. Stack the eigenvectors (aside from the first) v into a matrix E that is n by m. Call E the eigenspace, each point in the original dataset is represented by one row.
5. Cluster in the eigenspace, E, using knn.

Physics Process

UNIVERSITY OF Southampton

To find ? clusters from m points;

1. Identify affinities between all points; $a_{i,j}$.
2. Construct the graph Laplacien;

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

3. Calculate the eigenvectors v of L corresponding to the $q+1$ smallest eigenvalues.
4. Stack the eigenvectors (aside from the first) v into a matrix E that is q by m. Call E the eigenspace, each point in the original dataset is represented by one row.
5. Cluster in the eigenspace, E, using with a hierarchical method.

Conclusions

UNIVERSITY OF Southampton

This is a well motivated clustering method.

- The best hyperparameters need to be identified.
- It should be tested for IRC safety.
- It's replication of event shape variables should be tested.

These hurdles aside, the method shows potential when compared to traditional jet clustering algorithms.

Thank you for listening.