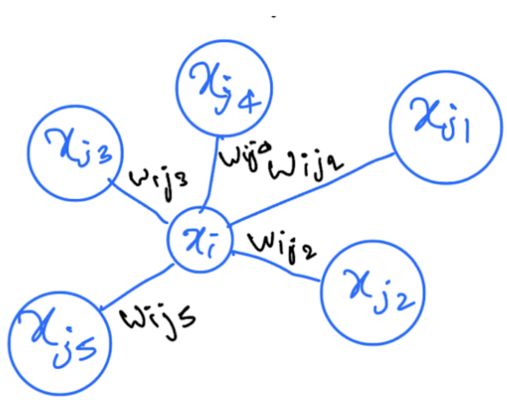
UMAP

- Stands for Uniform Manifold Approximation and Projection
- Uses the underlying concepts of algebraic topology and topological data analysis.

Steps for UMAP Algo:

1. Finalize the number of neighbors to consider for point x_i -> hyperparameter
2. Create a weighted graph for x_i and its nearest neighbor
    a. Where each neighbor edge is given weights inversely proportional to distance for x_i



$$w_{ij} = \frac{1}{dist(x_i, x_j)}$$

Goal: When we move from a higher dim to a lower dimension we want

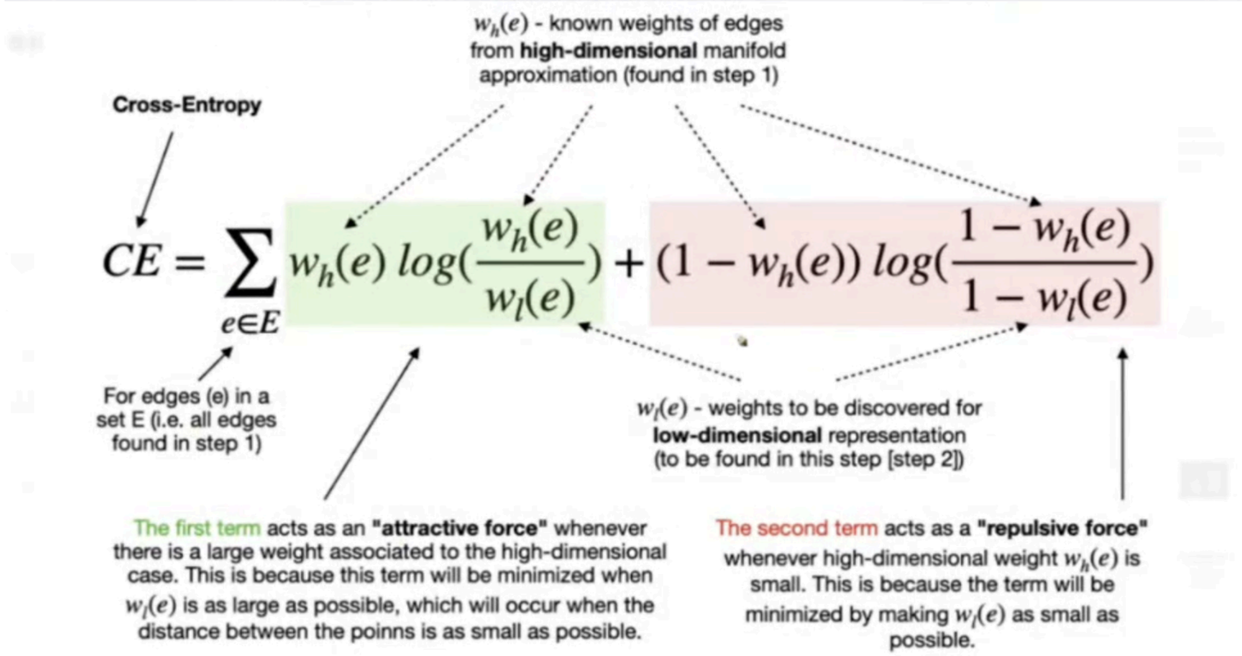
- Graph(x_i) should be similar to Graph(x_i')
- Where x_i is a point in high dim
- And x_i' is the point in low dimensional space

Why similar graphs?

- Helps in preserving neighborhood information.

How do we optimize UMAP algo?

Using the following loss:



Terms:

- $\sum_{e \in E}$ represents summation for each edge e among the set of edges E in the d-dimensional graph

- $w_h(e)$ is the weight of an edge in **high dimensional space** and $w_l e$ is the weight of an edge in **low dimensional space**

Goal: Reduce the loss

First half of the equation:

$$w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right)$$

- If $w_h e$ is large, then make sure that $w_h e$ and $w_l e$ are very close to each other.
- The ratio of $w_h e$ and $w_l e$ has to be 1, which can happen if they are very close to each other.

Second half:

$$(1 - w_h(e))\ log(\frac{1 - w_h(e)}{1 - w_l(e)})$$

- If $w_h e$ is small, then the term $1\ -\ w_h e$ will be large
    - Makes 1st half insignificant but 2nd half significant

- even if $w_h e$ is small, we want $w_h e$ and $w_l e$ to be close to each other, which is represented by the ratio in the second half of the equation