Connections Between Graph Spectral Clustering and PDEs

Catherine Huang[1] and Chloe Makdad[2], with faculty mentor Dr. Akil Narayan

Summer@ICERM 2020

What is Graph Spectral Clustering?

Graph spectral clustering is a method of partitioning data using spectral properties of its Laplacian matrix.

Algorithm

1. From data, construct a graph using a similarity metric.
2. Construct the Laplacian matrix $L := D - A$ where D is the degree matrix and A is the adjacency matrix.
3. Populate columns of U with the smallest k eigenvectors of L, where k is the number of clusters.
4. Run a clustering algorithm (k-means) on the rows of U.
5. Since each row of U corresponds to a data point, we get our clusters.

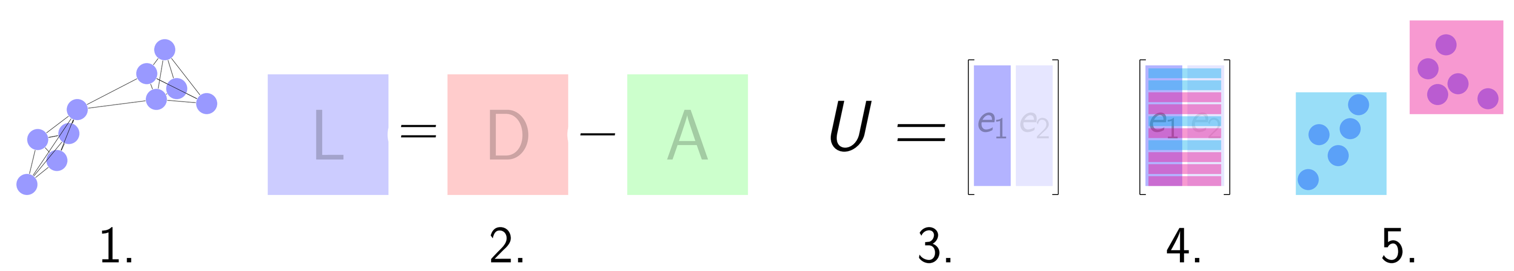
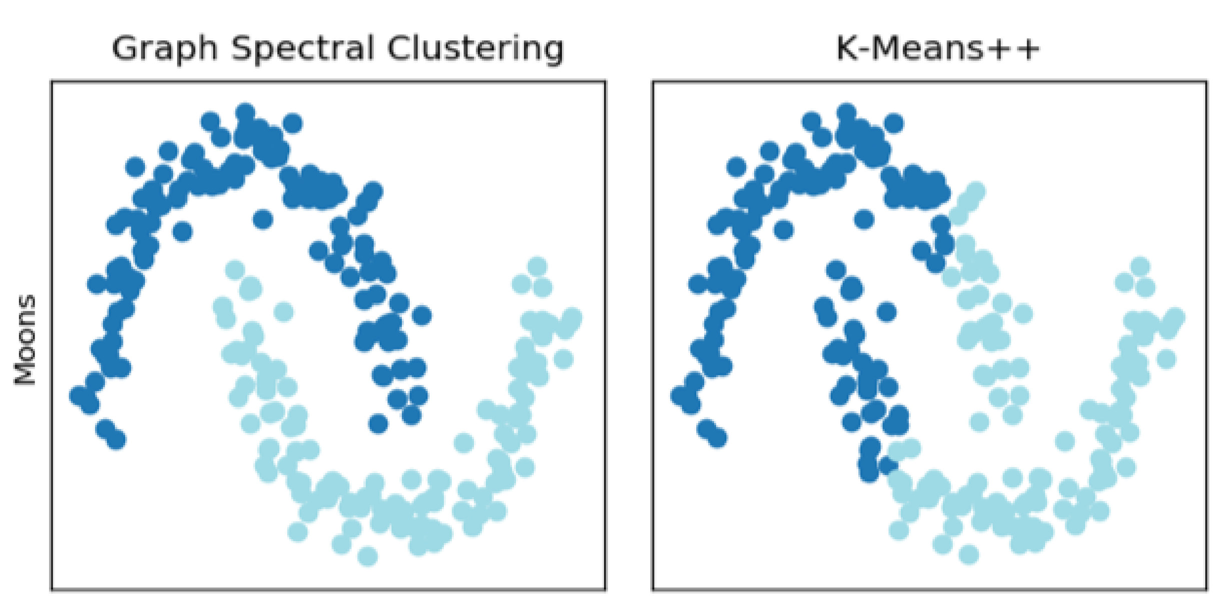


Figure: Steps of the graph spectral clustering algorithm

Why Graph Spectral Clustering?

Graph spectral clustering is computationally expensive. Why should we use it?



- Popular clustering algorithms like k-means fail to appropriately cluster data sets like the one above, where Euclidean distance isn't the best metric for clustering.
- Graph spectral clustering effectively clusters the data by grouping points that quickly diffuse **heat** to each other, but not other points.

The Heat Equation

Definition (The Heat Equation)

We define the heat equation as $\frac{\partial u}{\partial t} = \Delta u$, where $u(x,t)$ outputs the temperature at position x and time t.

With *boundary conditions* $u(0,t) = u(1,t) = 0$ and the *initial condition* $u(x,0) = f(x)$, solutions are of the form

$$u(x,t) = \sum_{n=-\infty}^{\infty} A_n \cdot \underbrace{e^{n^2\pi^2 t}}_{\substack{\text{frequency}\\ \text{component}}} \cdot \underbrace{e^{in\pi x}}_{\substack{\text{operator,}\\ \text{Fourier basis}}}$$

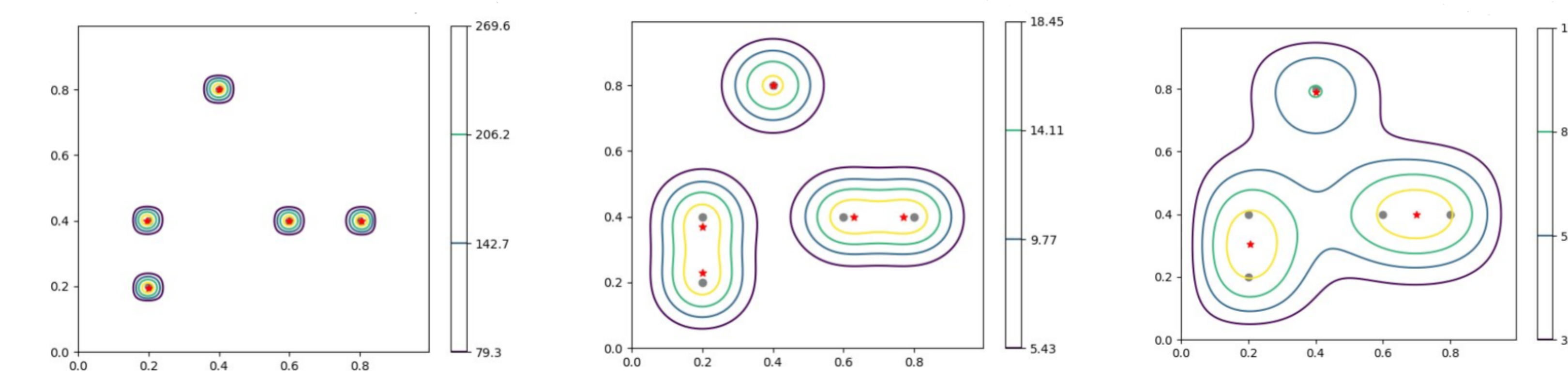
$$\text{where } A_n = \int_0^1 f(x) e^{in\pi x} dx.$$

Connecting the Heat Equation and GSC

- The more similar two points are, the more influence they have on each other with respect to temperature change.
- Let $f_{i,t}$ be the temperature of data point i at time t, then
$$\frac{\partial f_{i,t}}{\partial t} = \sum_{j:(i,j)\in E} (f_{j,t} - f_{i,t}) w_{ij}$$
- Combining these equations for all datapoints, we get
$$\frac{\partial \mathbf{f}_t}{\partial t} = -L\mathbf{f}_t \implies \mathbf{f}_{t+1} - \mathbf{f}_t = -L\mathbf{f}_t \implies \mathbf{f}_t = (I - L)^t \mathbf{f}_0$$
- Let v_i be eigenvectors and λ_i be the eigenvalues of L. These form a basis, so $\mathbf{f}_t$ can be rewritten as
$$\mathbf{f}_t = \sum_{i=1}^{n} \langle \mathbf{f}_0, v_i \rangle \cdot \underbrace{(1-\lambda_i)^t}_{\text{frequency}} \cdot \underbrace{v_i}_{\text{operator}}$$
- The smallest eigenvalues of L represent the slowest diffusing heat distributions. This picks out clusters with high intracluster similarity and intercluster disimilarity.

Diffusion and Clustering

Below, we cluster some points according to the diffusion process modeled by the heat equation:



(a) $u(x, t=0)$ (b) $u(x, t=0.0041)$ (c) $u(x, t=0.0091)$

Figure: Contour maps of u. Red stars indicate local maxima.

(a) This graph represents our initial data points, each with different random initial temperatures.

(b) The contours pick out three distinct clusters.

(c) As time goes on, boundary conditions enforce that the heat of the entire region decays to zero, eventually giving us a single cluster.

Further Questions

- In [3], Sahai introduces a distributed clustering method using the wave equation.
- Thus, it is natural to wonder which other PDEs lend themselves to clustering.

References

- Belkin, Mikhail and Partha Niyogi, "Laplacian Eigenmaps for Dimesionality Reduction and Data Representation". In: *Neural Computation* 15.6 (2003), pp. 1373-1396. doi:10.1162/089976603321780317.
- von Luxburg, Ulrike. "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4 (Dec. 2007). arXiv: 0711.0189[cs.DS], pp. 395–416.doi:10.1007/s11222-007-9033-z.
- Sahai, Tuhin, Speranzon, Alberto, and Andre Banazuk. "Hearing the clusters in a graph: A distributed algoithm." arXiv:0911.4729[physics], Apr 2011. arXiv: 0911.4729.

Joint work with Max Daniels and Shubham Makharia

[1]UC Berkeley, thecatherinehuang@berkeley.edu ; [2]Butler University, cmakdad@butler.edu