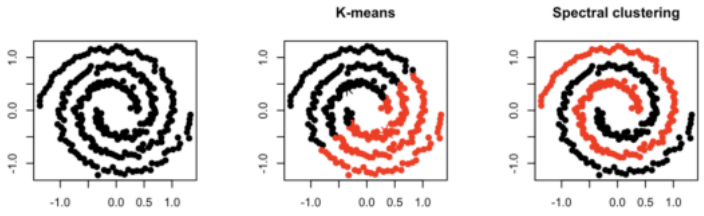
Dawei Wang
Mentor: Vydhourie Thiyageswaran
SPA-DRP
Autumn 2021

Graph Clustering

I am fortunate to participate in UW Statistics and Probability Association's Directed Reading Program this quarter and explore the statistical concept of graph clustering with my mentor Vydhourie Thiyageswaran.

First, I was introduced to the idea of networks. Although it is a common concept, the weekly assigned reading lets me view it in a greater depth and a statistical background in mind. For example, in a network setting, one should not evaluate their actions in isolation. Instead, cause and effect relationships can be very subtle. A small change with the assumption that everything else will remain static might create incentives that shift behavior across the network in ways that were initially unintended in reality. Furthermore, the dynamics of aggregate behavior could let us observe the significant impact made by such subtle effects. It was also interesting to see that Game theory, which I learned in microeconomics, is also being discussed in related topics of networks.

To talk about graph clustering, the key concept of a graph is that it is connected if for every pair of nodes, there is a path between them. And if every node in a subset has a path to every other, this subset (group) is called a connected component. For clustering, we seek to partition observations into distinct groups so that the observations within each group are similar, while observations in different groups are different. We mainly focused on learning k-means clustering and spectral clustering in our project. In short, k-means clustering is partitioning a data set into K distinct, non-overlapping clusters. And each observation belongs to the cluster with the nearest mean. Whereas spectral clustering makes use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. Below is an intuitive simple illustration of the two clustering methods.



I find it most interesting to relate such concepts with real world examples and how they can analytically generalize many phenomena. Moreover, using idealized models and equations could effectively help us discover more out of limited data. With this in mind, our case study was about the 381 paintings of Bob Ross. His paintings mainly consist of natural elements such as trees, mountains, clouds, lakes and snow. The article we referenced used clustering analysis to cluster similar paintings based on the contained elements, and come up with results such as “a cluster of 28 paintings each with an oval white-space frame” and “a cluster of 35 paintings of ocean scenes.” I tried to reproduce this analysis by using R Studio and k-means clustering. The process gave me the valuable experience of implementing k-means clustering in the real world, and I have gained a lot about data analysis, problem solving and research in general.