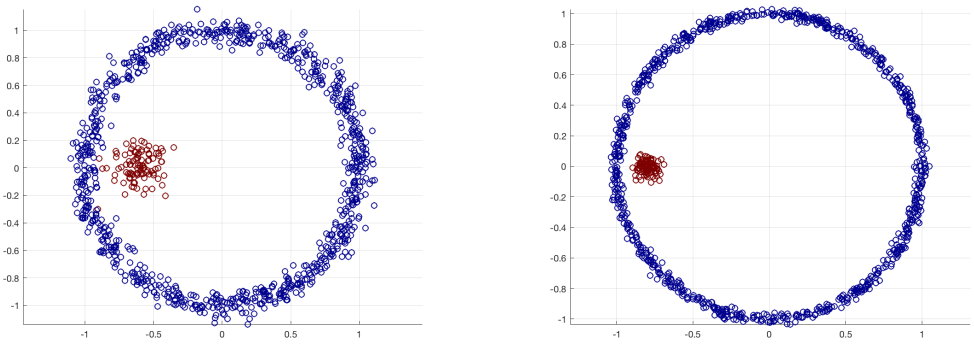
# Math 690 F2017: Topics in Data Analysis and Computation
# Homework 4

Xiuyuan Cheng

1. (Circle-and-dot two clusters) Consider the kernelized spectral clustering method applied to the following example of two clusters: the first cluster (blue) consists of points uniformly sampled on a circle plus noise $\mathcal{N}(0, \sigma^2 I_2)$, and the second cluster (red) points sampled from the normal distribution $\mathcal{N}(0, (2\sigma)^2 I_2)$, where the two clusters are separated by $\delta > 0$. What happens to the eigenvectors of graph laplacian when both $\sigma$ and $\delta$ decreases? You may consider effects of cluster size (proportion of points in cluster 1 and 2), the relative magnitude of $\sigma$ and $\delta$, the construction of graph laplacian (including self-tuning of width of gaussian kernel) and so on.



2. (Stochastic block model) Consider the SBM where the graph has $n = 100$ nodes, $k = 4$ equal-size clusters, and $p_{ij} = 0.8$ for $i, j$ in same cluster, and $p_{ij} = 0.2$ otherwise. A typical realization of the adjacency matrix $A$ is shown in the left of Figure 1, where the matrix of $P$ is shown on the right, $A_{ij} \sim \text{Bern}(p_{ij})$.

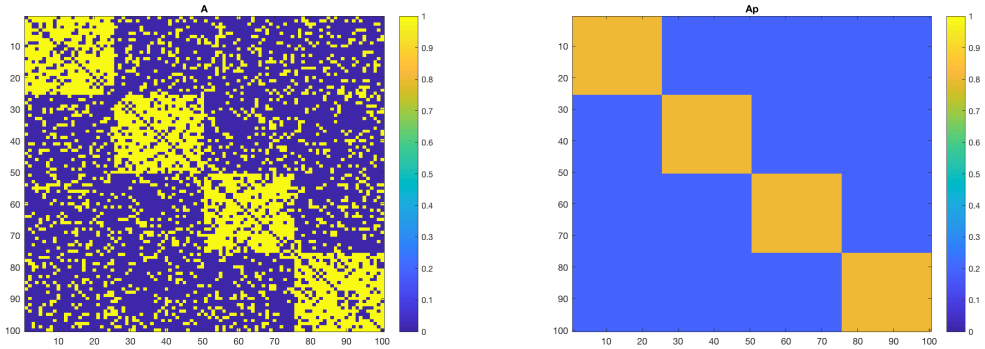



Figure 1: SBM with 4 blocks.

(1) Effect of $n$ and $k$: what happens to the eigenvectors and eigenvalues of $A$ as $n$ increases? What if $k$ also increases?

(2) Effect of the balance of cluster size: what if cluster size is unequal?

(3) Effect of $B_{kl}$: what if $B_{kl} = p_{ij}$ for $i \in C_k$ and $j \in C_l$ are not necessarily constant across $k$ and $l$?

(4) Effect of normalization: any difference in using $L_{\text{rw}}$, $L_{\text{sym}}$ or $L_{\text{un}}$?