Statistical Decision Theoretical Problem Statement for Clustering Graph Vertices

(i) Sample Space: $\mathcal{G}_n = (\mathcal{A}, \mathcal{Y})$, where $\mathcal{A} \in \{0,1\}^{n \times n}, \mathcal{Y} \in \{0,1\}^n$, n = number of vertices; Ex: $n = 4$.

(ii) Model: $SBM_n^k(\vec{\rho}, \vec{\beta})$, where $\vec{\rho}$ are regions and $\vec{\beta}$ are Bernoulli p's associated with each region k: $\vec{\rho} \in \Delta_k$, $\vec{\beta} \in (0,1)^{k \times k}$; Ex: $SBM_4^2(\vec{\rho}, \vec{\beta})$, $\vec{\rho} \in \Delta_2$, $\vec{\beta} \in (0,1)^{2 \times 2}$.

(iii) Action Space: $\mathcal{A} = \{y \in \{0,1\}^n\}$; Ex: $\mathcal{A} = \{y \in \{0,1\}^4\}$, i.e., indexing a 4-tuple by the vertices.

(iv) Decision Rule: $\phi = \phi(\mathcal{G}_n) \to \mathcal{A}$; Ex: ϕ is the k-means clustering function in Matlab.

(v) Loss function: $l : \mathcal{G}_n \times \mathcal{A} \to \mathbb{R}_+$, where l is the cost of each decision, e.g., a weighted probability of that decision; Ex: for k-means clustering, the loss l is the length of distances when putting a point in one cluster versus another cluster. Adjusted Rand Index (ARI) is another cluster validation analysis approach, measuring similarity between two partitions.

(vi) Risk function: $\mathcal{P} \times \mathcal{L} \to \mathbb{R}_+$; Ex: $E[l]$.