

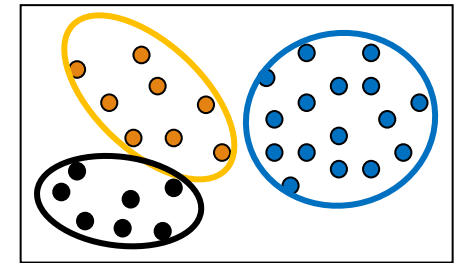
The background features a complex network of thin, light-colored lines forming a web-like structure. Scattered throughout are small, colored dots in shades of green, blue, and orange. A prominent, darker, reddish-brown geometric shape, resembling a stylized letter 'A' or a cluster of points, is visible in the center. The overall aesthetic is technical and data-driven.

Internal Measures for Clustering Validation



Internal Measures (I): BetaCV Measure

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering $C = \{C_1, \dots, C_k\}$ with k clusters, cluster C_i containing $n_i = |C_i|$ points
 - Let $W(S, R)$ be sum of weights on all edges with one vertex in S and the other in R
 - The sum of all the intra-cluster weights over all clusters: $W_{in} = \frac{1}{2} \sum_{i=1}^k W(C_i, C_i)$
 - The sum of all the inter-cluster weights: $W_{out} = \frac{1}{2} \sum_{i=1}^k W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1} \sum_{j>i}^k W(C_i, C_j)$
 - The number of distinct intra-cluster edges: $N_{in} = \sum_{i=1}^k \binom{n_i}{2}$
 - The number of distinct inter-cluster edges: $N_{out} = \sum_{i=1}^{k-1} \sum_{j=i+1}^k n_i n_j$
- **Beta-CV measure:** $BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$
 - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
 - The smaller, the better the clustering

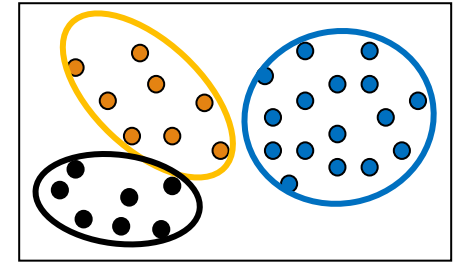


Internal Measures (II): Normalized Cut and Modularity

□ **Normalized cut:**
$$NC = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{vol(C_i)} = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{W(C_i, V)} = \sum_{i=1}^k \frac{W(C_i, \bar{C}_i)}{W(C_i, C_i) + W(C_i, \bar{C}_i)} = \sum_{i=1}^k \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \bar{C}_i)} + 1}$$

where $vol(C_i) = W(C_i, V)$ is the volume of cluster C_i

- The higher normalized cut value, the better the clustering



□ **Modularity** (for graph clustering)
$$Q = \sum_{i=1}^k \left(\frac{W(C_i, C_i)}{W(V, V)} - \left(\frac{W(C_i, V)}{W(V, V)} \right)^2 \right)$$

- Modularity Q is defined as

where
$$W(V, V) = \sum_{i=1}^k W(C_i, V) = \sum_{i=1}^k W(C_i, C_i) + \sum_{i=1}^k W(C_i, \bar{C}_i) = 2(W_{in} + W_{out})$$

- Modularity measures the difference between the observed and expected fraction of weights on edges within the clusters.
- The smaller the value, the better the clustering—the intra-cluster distances are lower than expected