Clustering Aggregation

-Terminology:

- Clustering: A group of clusters output by a clustering algorithm
- Cluster: A group of points

-Goals:

- Compare clusterings
- Combine the information from multiple clusterings to create a new clustering

-Comparing clusterings:

- Clusterings can be the same even if the assignments / labels are inconsistent.
- If many points were assigned to the same clusters in both clustering C and clustering P, then C and P should have a small distance.

-Disagreement Distance:

$$D(P, C) = \sum_{x,y} \mathbb{I}_{P,C}(x, y)$$

$$\mathbb{I}_{P,C}(x, y) = \begin{cases} 1 & \text{if P \& C disagree on which clusters x \& y belong to} \\ 0 & \end{cases}$$

-Aggregate clustering:

-Goal: From a set of clusterings $C_1, \ldots, C_m$, generate a clustering C^* that minimizes:

$$\sum_{i=1}^{m} D(C^*, C_i)$$

-Pros:

- Can identify the best number of clusters
- Can handle / detect outliers (points where there is no consensus)

- Improve robustness of the clustering algorithms since combining clusterings can produce a better result
- Privacy preserving clustering (can compute aggregate clustering with only sharing the assignments)

-Cons:

-NP hard question

-Majority rule only works if it produces a clustering