Lecture 8 Clustering Aggregation

- Clustering : a group of clusters output by some clustering algorithm
- Cluster: A group of points
- Goals of clustering Aggregation
  - Compare clusterings
  - Combine the information from multiple clusterings to make a new clustering
- Disagreement Distance
  - Given two clustering's P and C (For two partitions P and C of the dataset)
  - $D(P, C) = \sum_{x,y} Ip, c(x, y)$ *where*
  - $Ip, c(x, y) =$
    1 *if* P *and* C *disagree which cluster* x *and* y *belong to and* 0 *otherwise*
  - Formally,

$$I_{C,P}(x, y) = \begin{cases} 1 & \text{if } C(x) = C(y) \text{ and } P(x) \neq P(y) \\ & \text{OR} \\ & \text{if } C(x) \neq C(y) \text{ AND } P(x) = P(y) \\ 0 & \text{otherwise} \end{cases}$$

  -
  - Disagreement distance is a measure of how different two clusterings of a set of data points are , counting the number of disagreeing object pairs between
  - P(x) is the cluster index in partition P that contains object x
  - C(x) is the cluster index in partition C that contains object x
  - Comparing rows in table not cols
  - N choose k pairs
    - $\frac{n}{k} = \frac{n!}{k!(n-k)!}$

Object	Cluster in C	Cluster in P
x1	1	1
x2	1	2
x3	2	1
x4	3	3
x5	3	4

- **(x1, x2)**:
  - P: same ($1 \neq 2$) → different
  - C: same ($1 = 1$) → same → **Disagree → 1**
- **(x1, x3)**:
  - P: same ($1 = 1$)
  - C: different ($1 \neq 2$) → **Disagree → 1**
- **(x1, x4)**:
  - P: different
  - C: different → **Agree → 0**

- **(x1, x5)**:
  - P: different
  - C: different → **Agree → 0**
- **(x2, x3)**:
  - P: different (2 ≠ 1)
  - C: different (1 ≠ 2) → **Agree → 0**
- **(x2, x4)**:
  - P: different
  - C: different → **Agree → 0**
- **(x2, x5)**:
  - P: different
  - C: different → **Agree → 0**
- **(x3, x4)**:
  - P: different
  - C: different → **Agree → 0**
- **(x3, x5)**:
  - P: different
  - C: different → **Agree → 0**
- **(x4, x5)**:
  - P: different (3 ≠ 4)
  - C: same (3 = 3) → **Disagree → 1**

So the total disagreement distance:

  - $D(P,C)=1+1+1=3$

- Aggregate Clustering
  - Can identify best number of clusters
  - Can handle and detect outliers
  - Combining clusterings can produce better results
  - Preserves privacy
  - Problem is in NP Hard