

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle, light gray geometric pattern. To the left of the banner, there is a small, rectangular inset image showing a dense cluster of orange and red dots on a light background. The main background is a dark, reddish-brown color, overlaid with a network of thin, light-colored lines that form a complex, interconnected web. Scattered throughout this network are numerous small, green dots. The overall aesthetic is technical and data-driven, suggesting a focus on network analysis or data visualization.

External Measures III: Pairwise Measures

Pairwise Measures: Four Possibilities for Truth Assignment

Four possibilities based on the agreement between cluster label and partition label

TP: true positive—Two points \mathbf{x}_i and \mathbf{x}_j belong to the same partition T , and they also in the same cluster C

$$TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$$

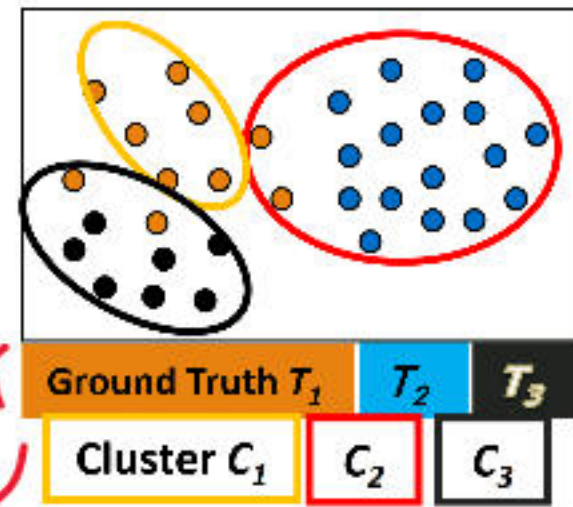
True Positive

where y_i : the true partition label, and \hat{y}_i : the cluster label for point \mathbf{x}_i

FN: false negative: $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$

FP: false positive: $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i = \hat{y}_j\}|$

TN: true negative: $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j \text{ and } \hat{y}_i \neq \hat{y}_j\}|$



Calculate the four measures:

$$N = \binom{n}{2}$$

Total # of pairs of points

$$TP = \sum_{i=1}^r \sum_{j=1}^k \binom{n_{ij}}{2} = \frac{1}{2} \left(\sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right) - n$$

$$FN = \sum_{j=1}^k \binom{m_j}{2} - TP$$

$$FP = \sum_{i=1}^r \binom{n_i}{2} - TP \quad TN = N - (TP + FN + FP) = \frac{1}{2} \left(n^2 - \sum_{i=1}^r n_i^2 - \sum_{j=1}^k m_j^2 + \sum_{i=1}^r \sum_{j=1}^k n_{ij}^2 \right)$$

Pairwise Measures: Jaccard Coefficient and Rand Statistic

- **Jaccard coefficient:** Fraction of true positive point pairs, but after ignoring the true negatives (thus asymmetric)

- $Jaccard = TP / (TP + FN + FP)$ [i.e., denominator ignores TN]

- Perfect clustering: $Jaccard = 1$

- **Rand Statistic:**

- $Rand = (TP + TN) / N$

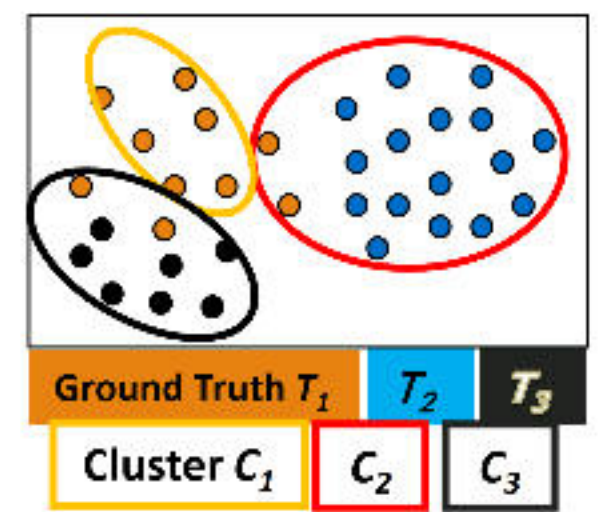
- Symmetric; perfect clustering: $Rand = 1$

- **Fowlkes-Mallow Measure:**

- Geometric mean of precision and recall

$$FM = \sqrt{prec \times recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

- Using the above formulas, one can calculate all the measures for the green table (leave as an exercise)



$C \backslash T$	T_1	T_2	T_3	Sum
C_1	0	20	30	50
C_2	0	20	5	25
C_3	25	0	0	25
m_j	25	40	35	100