Clustering fundamentals

Data Base and Data Mining Group of Politecnico di Torino

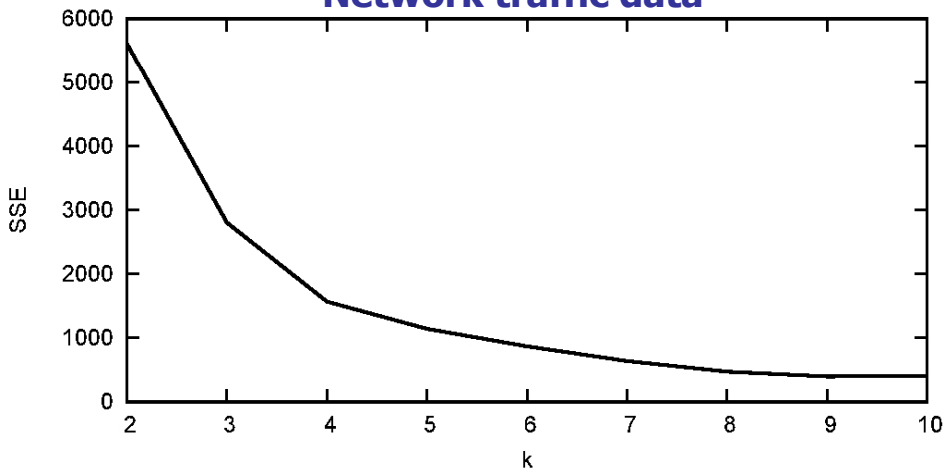
Elena Baralis, Tania Cerquitelli

Politecnico di Torino

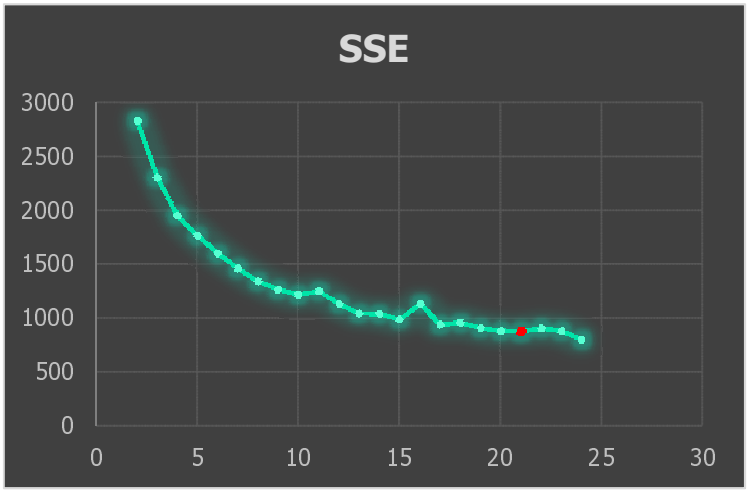
K-means parameter setting

- Elbow graph (Knee approach)
  - Plotting the quality measure trend (e.g., SSE) against K
  - Choosing the value of K
    - the gain from adding a centroid is negligible
    - The reduction of the quality measure is not interesting anymore

Network traffic data



Medical records



Evaluating cluster quality: Silhouette

- To ease the interpretation and validation of consistency within clusters of data
  - a succinct measure to evaluate how well each object lies within its cluster
- For each object *i*
  - *a(i):* the average dissimilarity of *i* with all other data within the same cluster (the smaller the value, the better the assignment)
  - *b(i):* is the lowest average dissimilarity of *i* to any other cluster, of which *i* is not a member

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & a(i) < b(i) \\ 0, & a(i) = b(i) \\ b(i)/a(i) - 1 & a(i) > b(i) \end{cases}$$

  - The average s(i) over all data of the dataset measures how appropriately the data has been clustered
  - The average s(i) over all data of a cluster measures how tightly grouped all the data in the cluster are