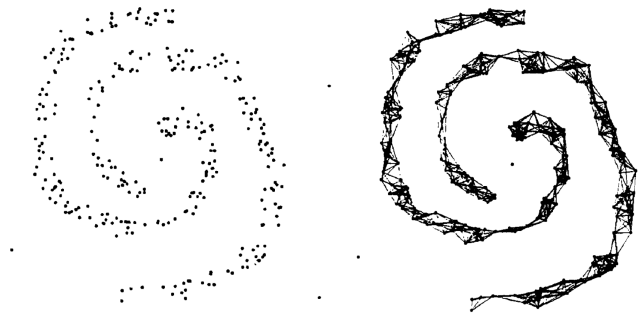
Mutual k-nearest neighbor graph for data analysis: Application to metric space clustering

Proposal: Broaden the applicability of a theoretically sound and simple clustering algorithm

Build the Mutual Mutual k-Nearest Neighbor graph of the data, for certain k

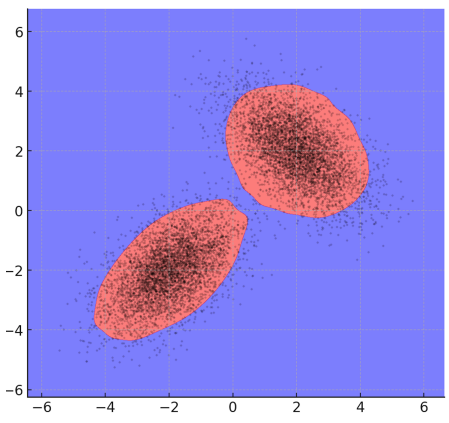
- Large connected components are clusters
- Small components are outliers



Old algorithm requires a bounded-away from zero distribution

Idea: Shave the distribution!

- Filter out low density regions
- The remaining is bounded-away from zero by construction



The correct k can be found iteratively

- We can use dendograms

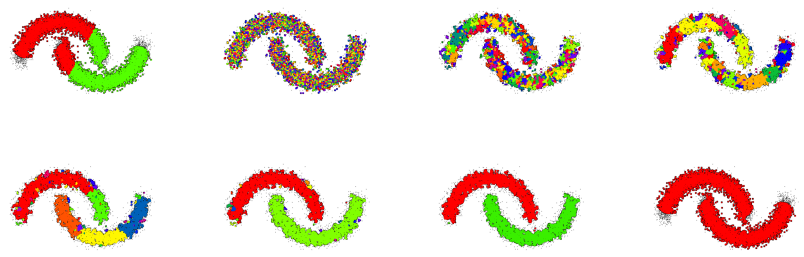


Figure: Two Moons partitions. [top left] k-means partition. [others] our procedure fixing $\tau_{\mathrm{LOF}} = 1.1$ and evolving $k \in \{1, 4, 5, 6, 7, 10, 1000\}$ in reading order

Without shaving

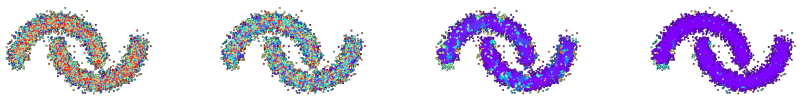


Figure: Two Moons partitions without prior filtering and evolving $k \in \{1, 2, 4, 5\}$ in reading order

Compared to DBSCAN

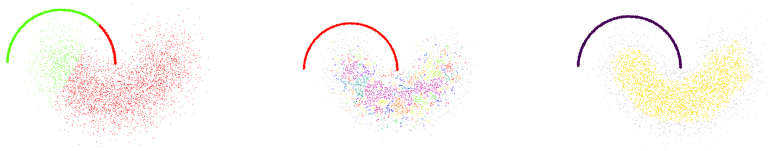


Figure: Adapted Two Moons dataset with varying densities. [left] k-means partition. [center] DBSCAN partition. [right] m-based partition

Take out

- The filtering step, to be useful, should retain most data from the original dataset
- Finding the proper filtering value is a problem in itself, it should be related to *fords* detection

Edgar Chavez, Stephane Marchand-Maillet and Adolfo Quiroz

elchavez@cicese.edu.mx stephane.marchand-maillet@unige.ch aj.quiroz1079@uniandes.edu.co

CICESE, Mexico & University of Geneva, Switzerland & Universidad de los Andes, Colombia

sisap