Chameleon 2

A human-like hierarchical clustering algorithm

T. Barton
tomas.barton@img.cas.cz

T. Bruna

P. Kordik

P. Bartunek

CZ OPENSCREEN

IMG
Institute of Molecular Genetics of the ASCR, v. v. i.

Clustering algorithms

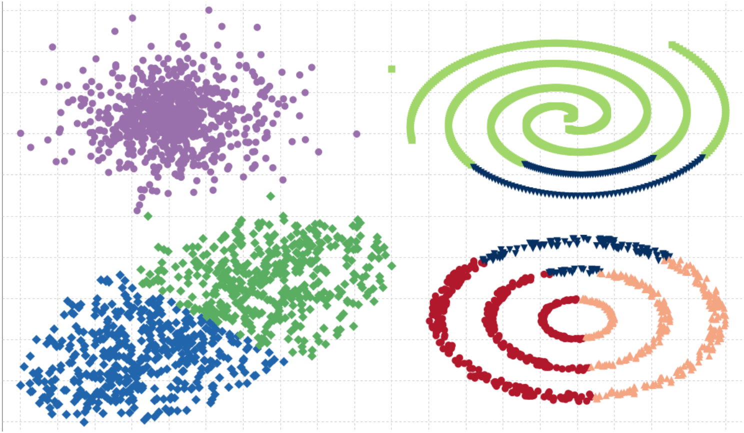
There are many clustering algorithms:

- *k*-means
- Hierarchical clustering
- DBSCAN
- CLARANS
- Markov clustering
- Affinity propagation
- x-means
- Spectral clustering
- Self Organizing Maps
- Fanny
- Transitivity clustering
- CLUTO
- clusterdp
- Chinese Whispers
- Fast Community
- ... and many others

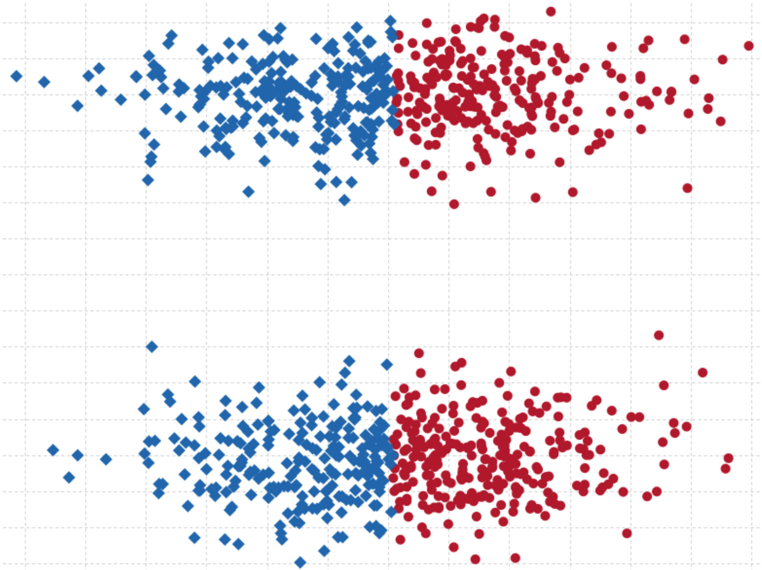
Why introduce a new one?

k-means clustering

- most algorithms optimize single objective
- e.g. minimize square distance inside a cluster
- fast, but inaccurate

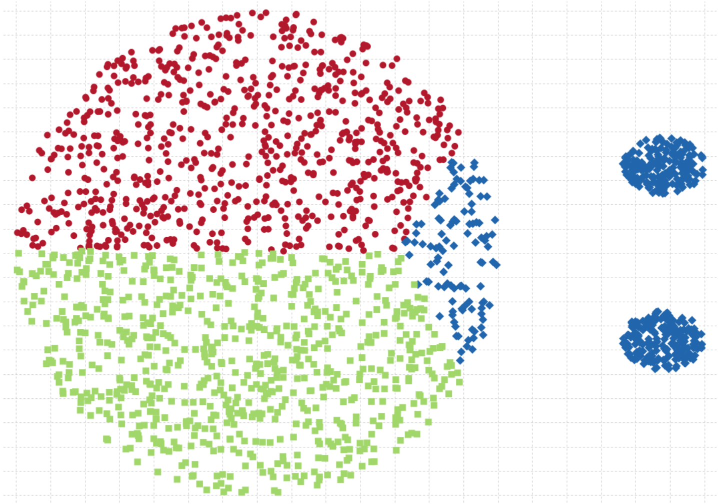


Limitation of k-means



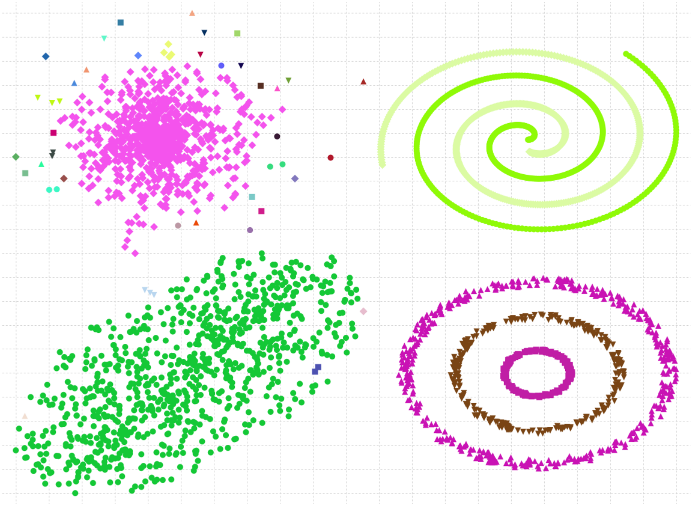
k-means ($k = 3$)

- need to specify number of clusters
- sensitivity to initialization



Single-Link clustering

- capable of discovering arbitrary shaped clusters
- but too sensitive to noise



Algorithms are hard to configure

- configuration requires either expert or sophisticated heuristic
- some methods are too sensitive to parameter changes
- same configuration being applied to different problems

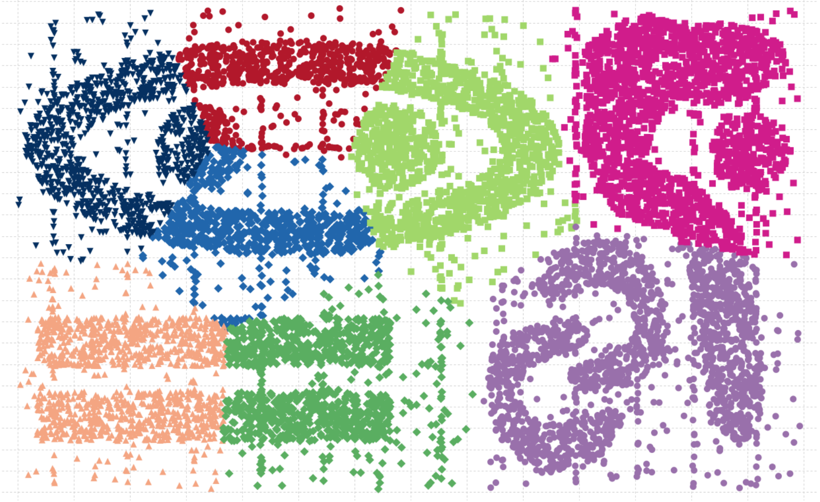
Chameleon

1999 Karypis, George, Eui-Hong Han, and Vipin Kumar. "Chameleon: Hierarchical clustering using dynamic modeling." Computer 32.8

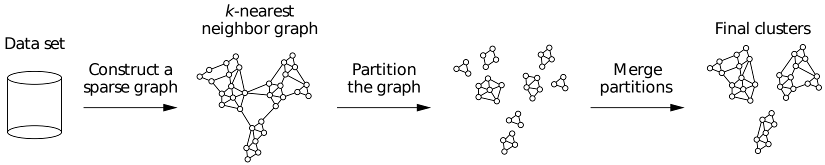
Problems?

- nobody ever reproduced that result
- no implementation of that algorithm exists
- core is a black-box graph partitioning algorithm

k-means



Chameleon algorithm

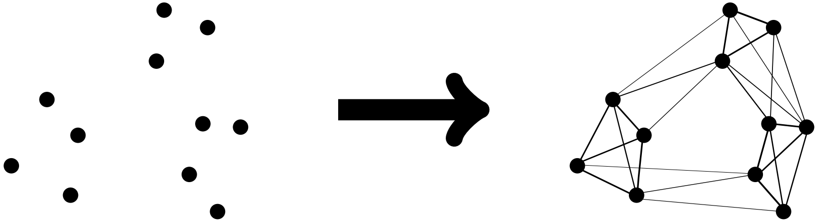


1. create k-nearest neighbor graph
2. partion the graph
3. merge partitions

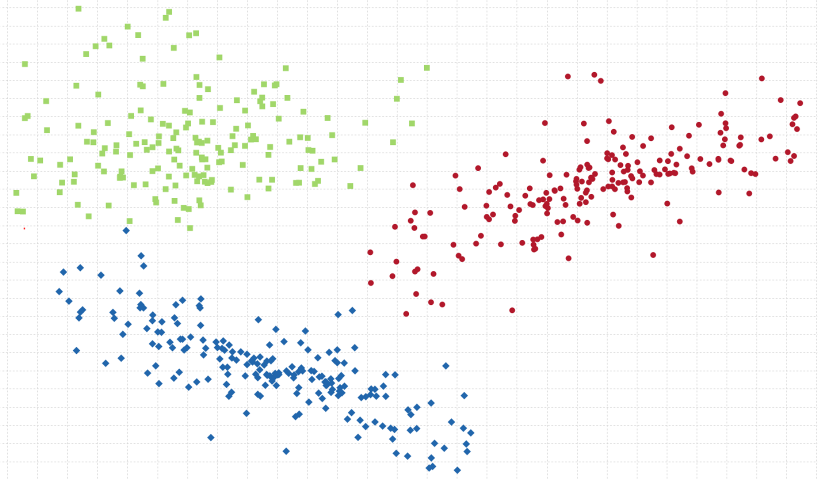
Chameleon

1. k-nn

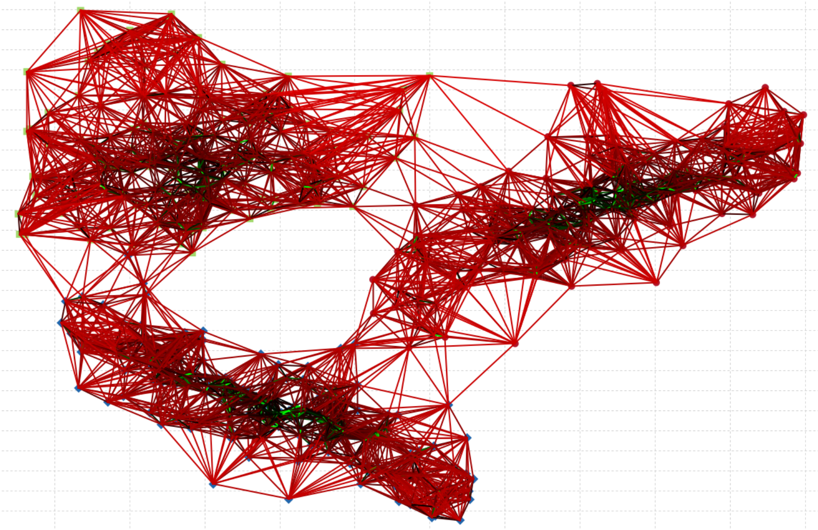
- dataset represented as a graph



Example: DS-577



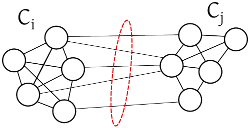
Example: DS-577



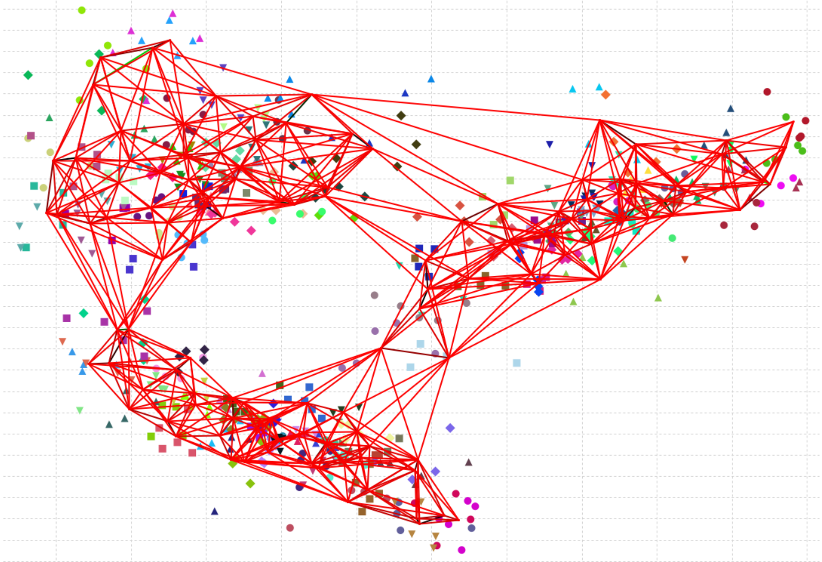
Chameleon

2. partitioning

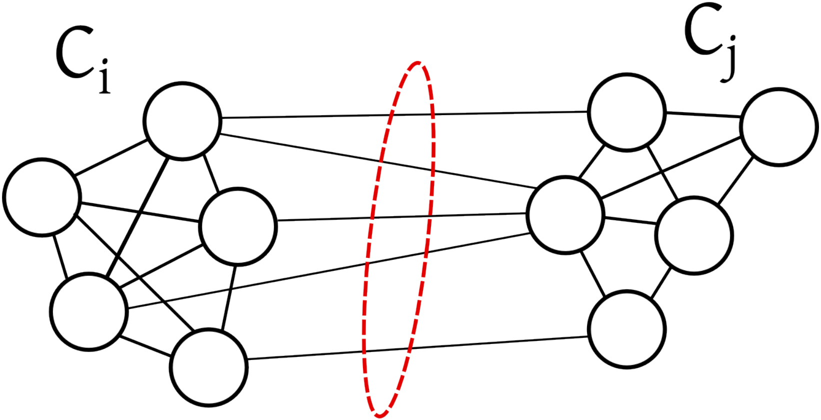
- split graph into many components
- minimalize egdes cut
- optimal bisection is NP-complete, thus we use approximation:
    - Kerighan-Lin $\mathcal{O}(n^3)$
    - Spectral bisection $\mathcal{O}(n^3)$
    - Fiduccia-Matheyses $\mathcal{O}(|E|)$
    - METIS / hMETIS



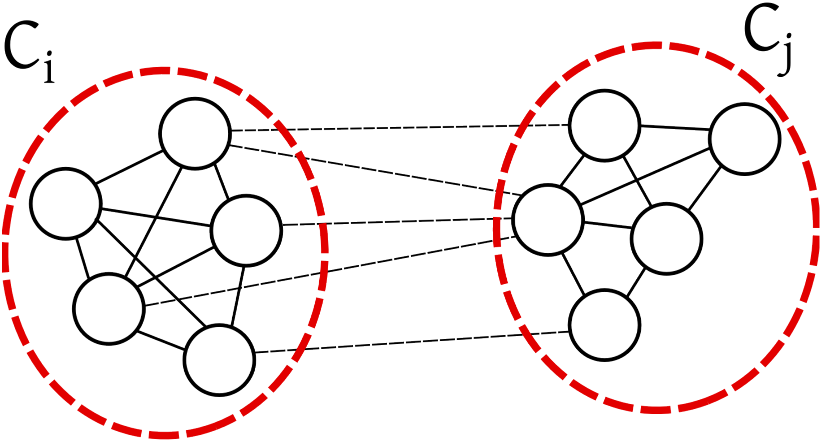
Example: DS-577



$\bar{s}(C_i, C_j)$



$\bar{\phi}(C_i), \bar{\phi}(C_j)$



Chameleon

3. merging

- find best candidate for merging

$$R_{IC}(C_i, C_j) = \frac{s(C_i, C_j)}{\frac{\phi(C_i) + \phi(C_j)}{2}}$$

$$R_{RC}(C_i, C_j) = \frac{\bar{s}(C_i, C_j)}{\frac{|C_i|}{|C_i| + |C_j|}\bar{\phi}(C_i) + \frac{|C_j|}{|C_i| + |C_j|}\bar{\phi}(C_j)}$$

Chameleon

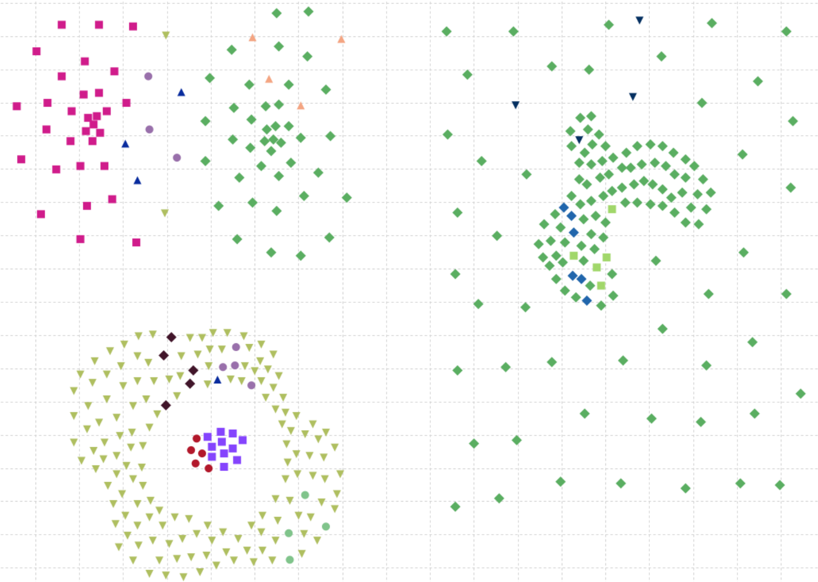
Combination of objectives

$$Sim(C_i, C_j) = R_{CL}(C_i, C_j)^{\alpha} \cdot R_{IC}(C_i, C_j)^{\beta}$$

- α, β – user defined priorities
- $\alpha = 2$
- $\beta = 1$

Chameleon 1

Standard similarity



Chameleon 2

Chameleon 2

essential modifications

1. partition fill step
2. inverse edges weights
3. replaced randomized bisections by deterministic approach
4. improved similarity measure
5. elimination of tiny clusters:

$$Sim(C_i, C_j) = \begin{cases} |E_{C_i}| \vee |E_{C_j}| = 0 & R_{\text{CLS}}(C_i, C_j) \cdot m_{\text{fact}} \\ |E_{C_i}| \wedge |E_{C_j}| > 0 & Sim_{\text{shat}}(C_i, C_j) \end{cases}$$

Chameleon 2

improved similarity measure

$$Sim_{\text{shat}}(C_i, C_j) = R_{\text{CLS}}(C_i, C_j)^{\alpha} \cdot R_{\text{ICS}}(C_i, C_j)^{\beta} \cdot \gamma(C_i, C_j)$$

$$R_{CLS}(C_i, C_j) = \frac{\bar{s}(C_i, C_j)}{\frac{|E_{C_i}|}{|E_{C_i}|+|E_{C_j}|}\bar{s}(C_i) + \frac{|E_{C_j}|}{|E_{C_i}|+|E_{C_j}|}\bar{s}(C_j)}$$

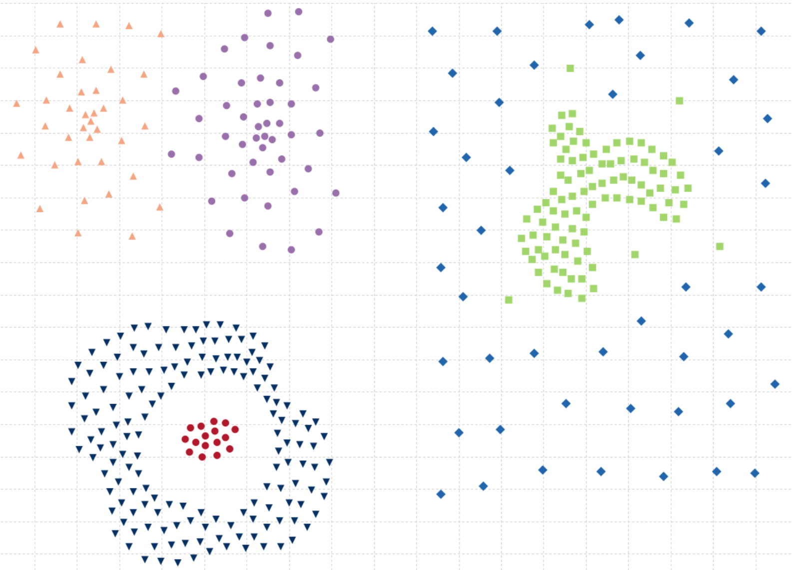
$$R_{\text{ICS}}(C_i, C_j) = \frac{\min\{\bar{s}(C_i), \bar{s}(C_j)\}}{\max\{\bar{s}(C_i), \bar{s}(C_j)\}}$$

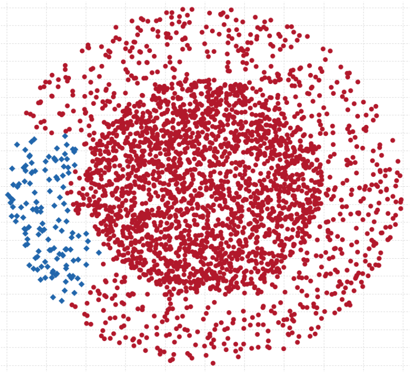
$$\gamma(C_i, C_j) = \frac{|E_{C_{i,j}}|}{\min(|E_{C_i}|, |E_{C_j}|)}$$

Where $\bar{s}(C_i)$ is defined as sum of edges' weights in a cluster

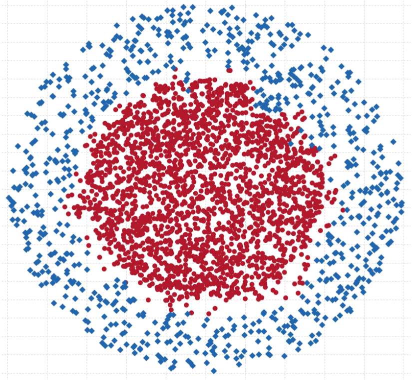
Chameleon 2

compound dataset





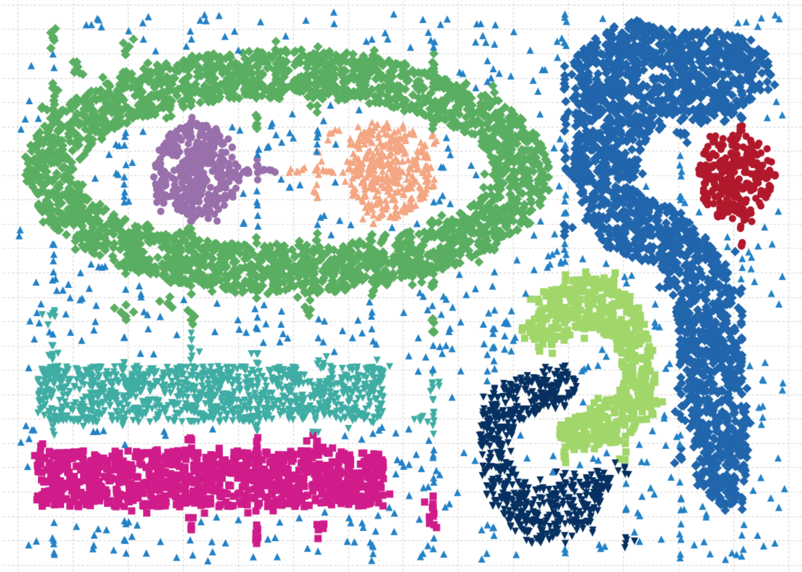
(a) CLUTO, NMI = 0.18



(b) Ch2, NMI = 0.76

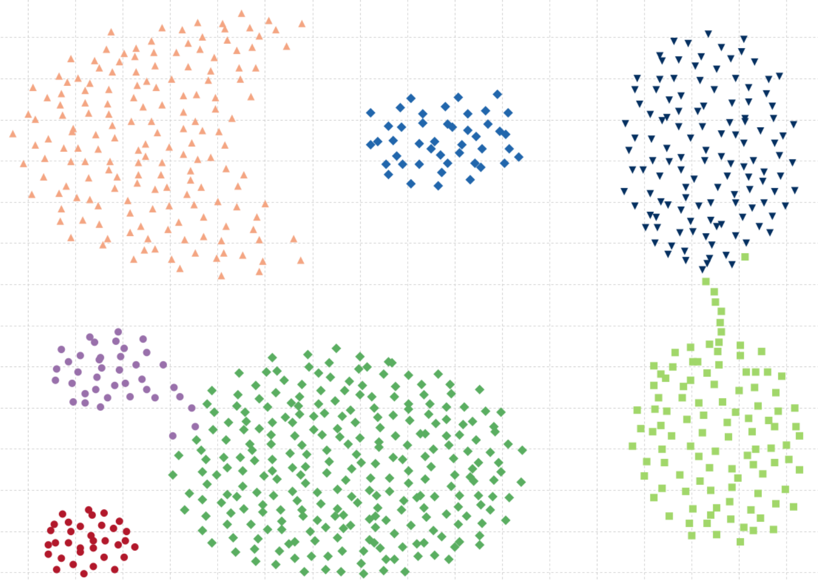
Chameleon 2

cluto-t7.10k



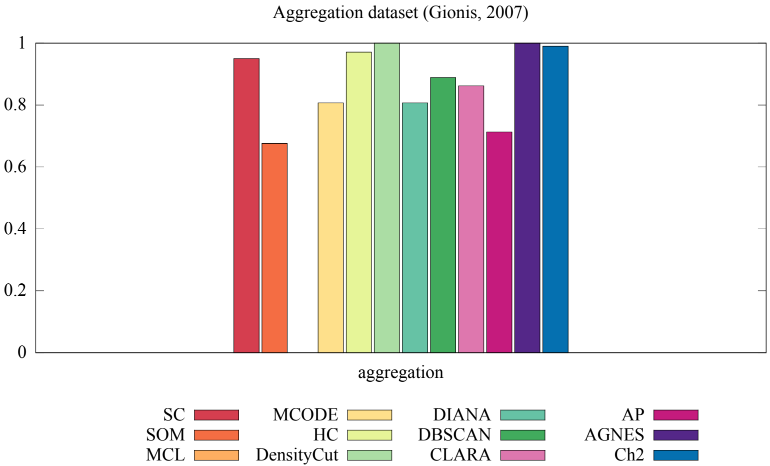
Chameleon 2

aggregation



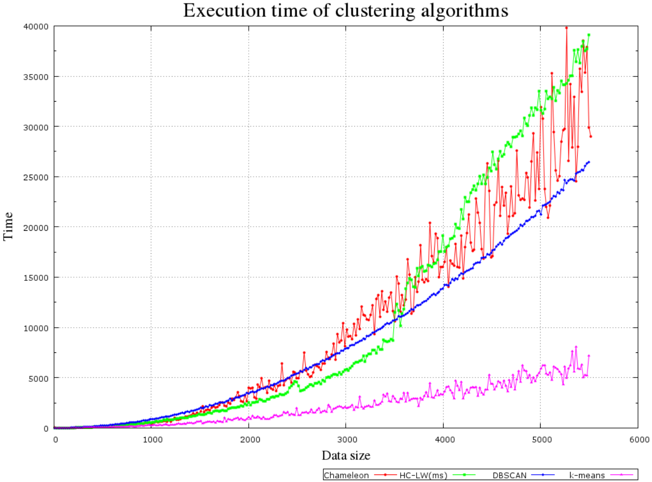
Chameleon 2

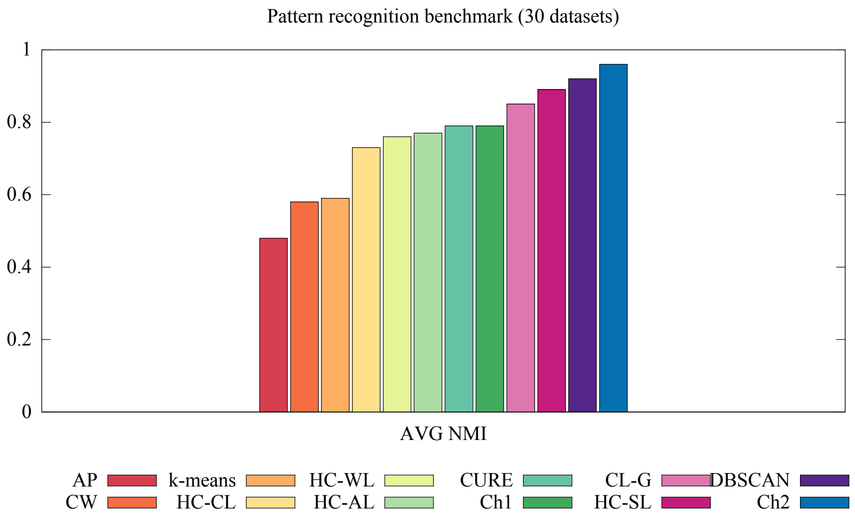
aggregation



Complexity

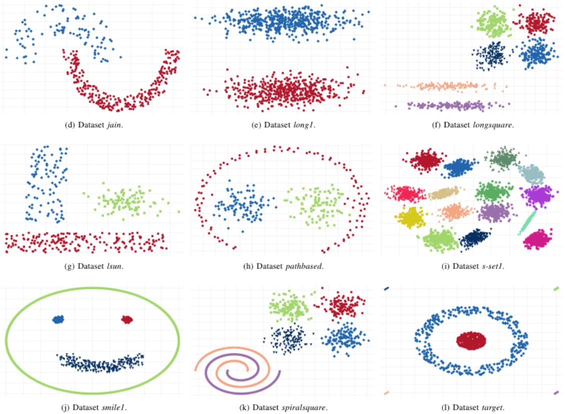
$$O\left(dn^2 + n + (n + m^2)\log(m)\right)$$




Pattern recognition benchmark (30 datasets)
1
0.8
0.6
0.4
0.2
0
AVG NMI
AP
CW
k-means
HC-CL
HC-WL
HC-AL
CURE
Ch1
CL-G
HC-SL
DBSCAN
Ch2

Benchmark

- new clustering algorithms are validated on artificial datasets



Summary

Chameleon 2

- based on stable methods
- no parameter tuning necessary
- hierarchical result

Questions?

Thank you for your attention

`tomas.barton@img.cas.cz`

Chameleon 2: an improved graph based clustering algorithm for finding structure in complex data was submitted to IEEE Transactions on Knowledge and Data Engineering