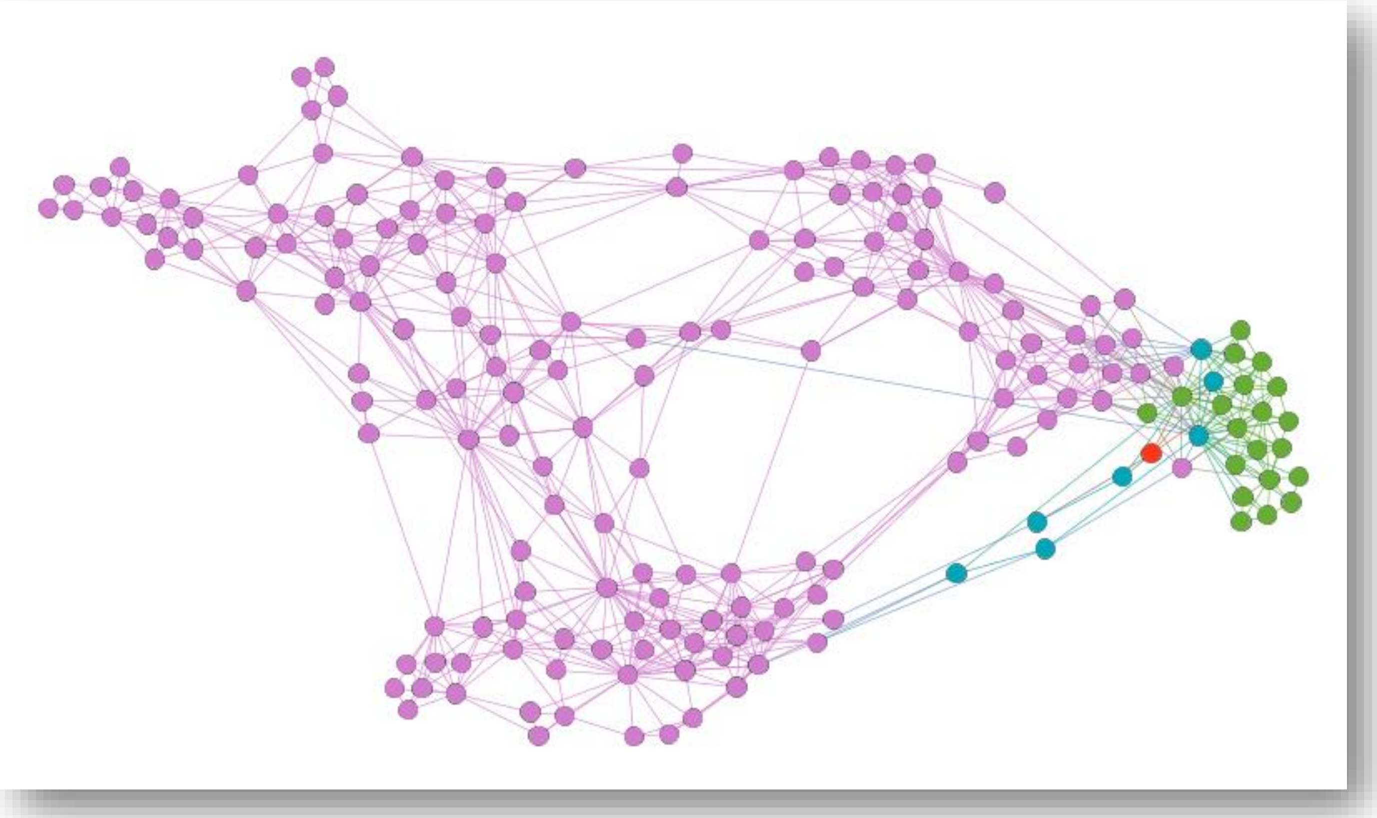
DATA ANALYSIS II

Network Construction

2021/22

Motivation

- It is always a good idea to perform exploratory data analysis, such as plotting the data, before applying a machine learning method…



Problem statement

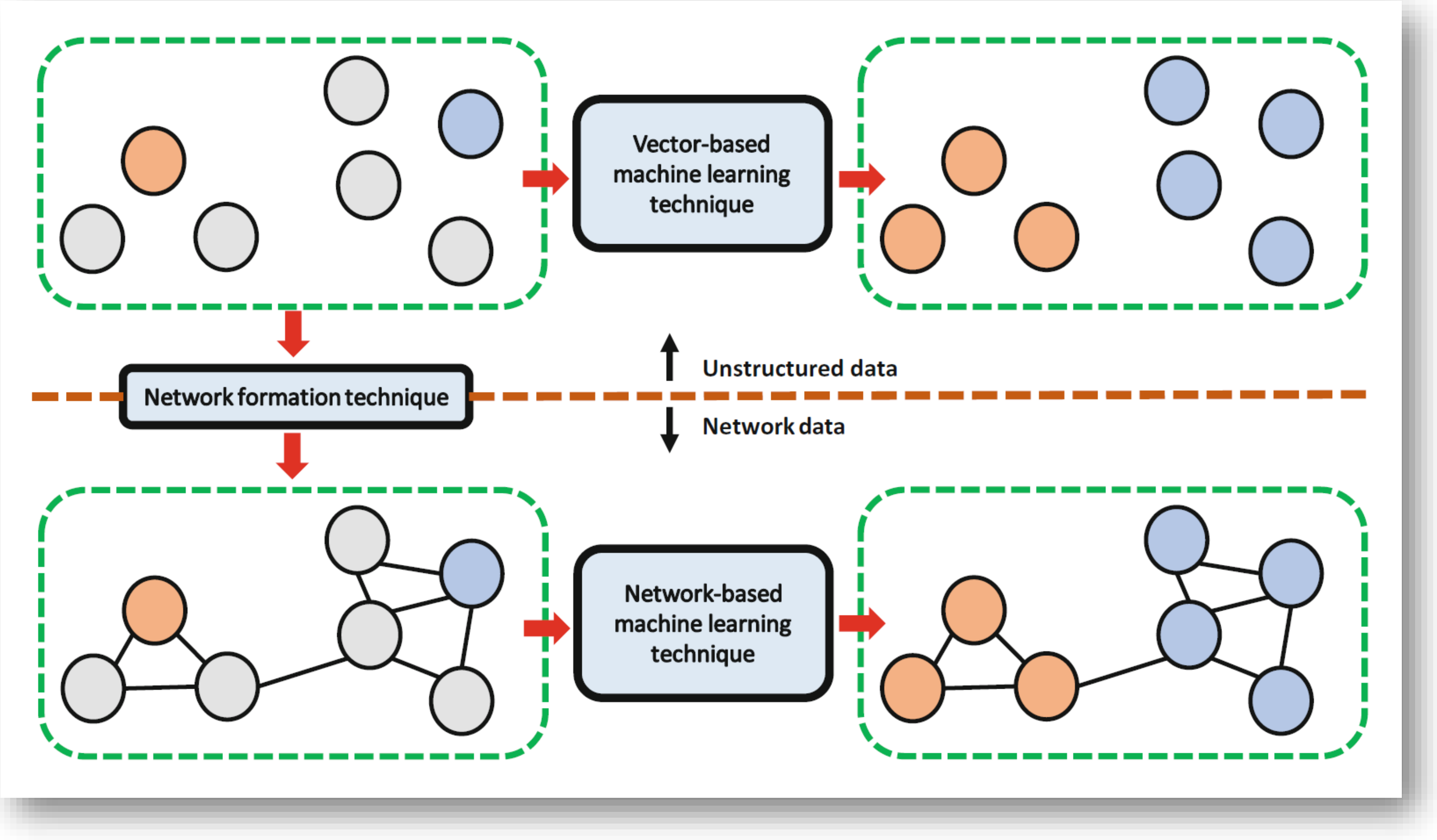
- *Network Construction* methods operate on a network where a node (vertex) represents a data instance and a pair of nodes are connected by a weighted edge.
    - Machine learning (vector data)
    - Network analysis (network data)

Silva, T. C., Zhao, L. (2016). *Machine learning in complex networks* (2016). Springer

Two approaches

- *Network Construction* methods operate on a network where a node represents a data instance and a pair of nodes are connected by a weighted edge.
  - **Task-independent Network Construction:** These methods do not use labeled data for network construction, and hence they are task-independent or unsupervised in nature.
  - **Task-dependent Network Construction:** Algorithms that fall under this category make use of both the labeled and unlabeled data for network construction. Labeled data can be used as a prior for adapting the network to the task at hand.

Vector Data x Networks


Vector-based machine learning technique
Network formation technique
Unstructured data
Network data
Network-based machine learning technique

Why networks?

- Visualization
  - Thanks to layout algorithms, there is a way how to „project“ data with a high dimensionality to 2D
- Links (edges) between nodes representing similarity
- Groups of similar nodes (clusters, communities)
- Ranking / rating nodes

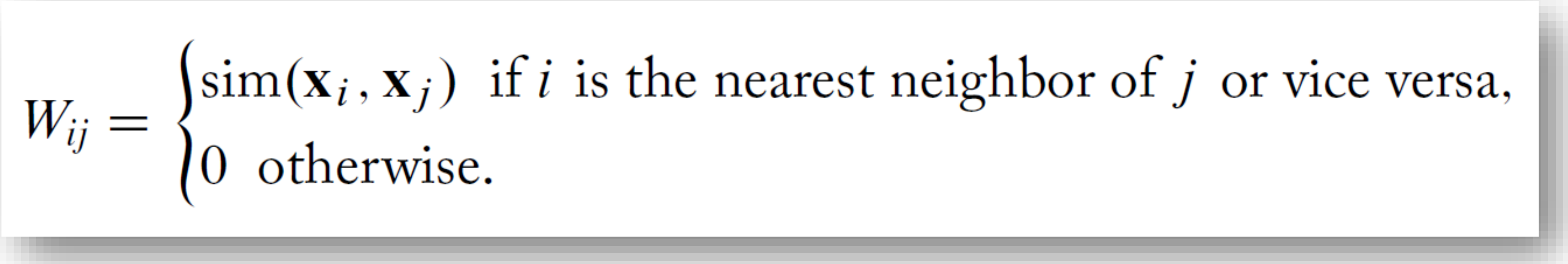
Two Factors

- **A proper similarity function *s*:** the similarity function *s* enables us to quantify how different or similar two data items are with respect to their attributes. Applying the similarity function to all of the pairs of vertices, we are able to construct
  - the similarity matrix
  - the dissimilarity (distance) matrix
- **A network formation technique:** we decide whether or not to add a link between v_i and v_j by using some rules applied on the similarity matrix or on the dissimilarity matrix.

Similarity (distance) functions

- Gaussian kernel, cosine similarity,…
- Correlation coefficient
- Co-occurrence
- Euclidean distance
- Each to each similarity (distance): symmetric matrix representing a complete graph.

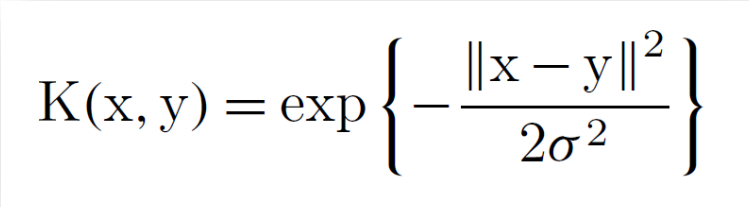
Similarity Matrix (*k* nearest neighbors)



$$W_{ij} = \begin{cases} \text{sim}(\mathbf{x}_i, \mathbf{x}_j) & \text{if } i \text{ is the nearest neighbor of } j \text{ or vice versa,} \\ 0 & \text{otherwise.} \end{cases}$$

- *W* is a symmetric matrix and an edge is added between nodes *i* and *j* if either *i* is a nearest neighbor of *j* or vice versa.

- The Gaussian kernel (a similarity function), also called the Gaussian radial basis function (RBF) kernel, is defined as



$$\mathrm{K}(\mathrm{x}, \mathrm{y}) = \exp\left\{-\frac{\|\mathrm{x}-\mathrm{y}\|^2}{2\sigma^2}\right\}$$

Graph x network

- *Graph*: if there is not a link between nodes, then, there is not a link (math).

- *Network*: If there is not a link between nodes, then we are not sure about the link (real world).

- Graph is a representation of a network…

Real-world networks

- Small world property: average lenght of the shortest path between nodes is *log N* or less

- Scale-free property: The power law distribution of node degree (a number of neighbors)

- Community structure (clusters)

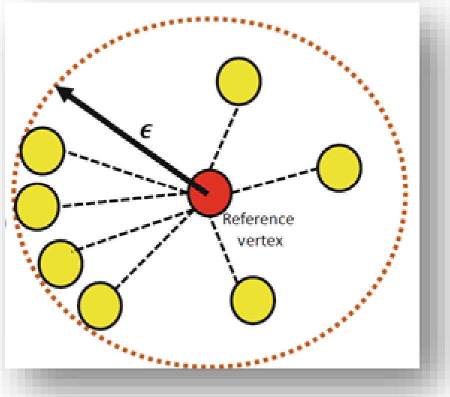
Methods

Can we trust the networks constructed from vector data?

ε-radius based method

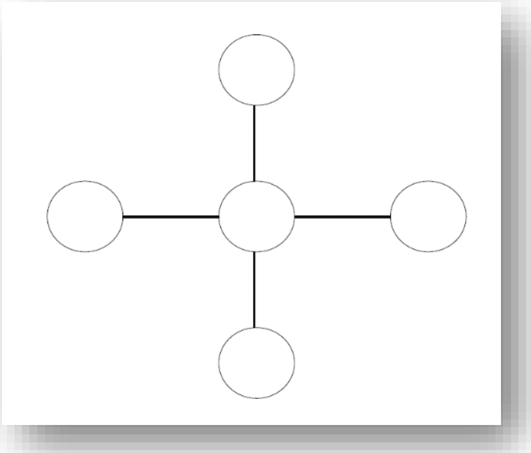
- An undirected edge between two nodes is added if the distance between them is smaller than ε, where $\varepsilon > 0$ is a predefined constant.

- Given a point x, a ball (specific to the chosen distance metric) of radius ε is drawn around it and undirected edges are added between x and all points inside this ball.

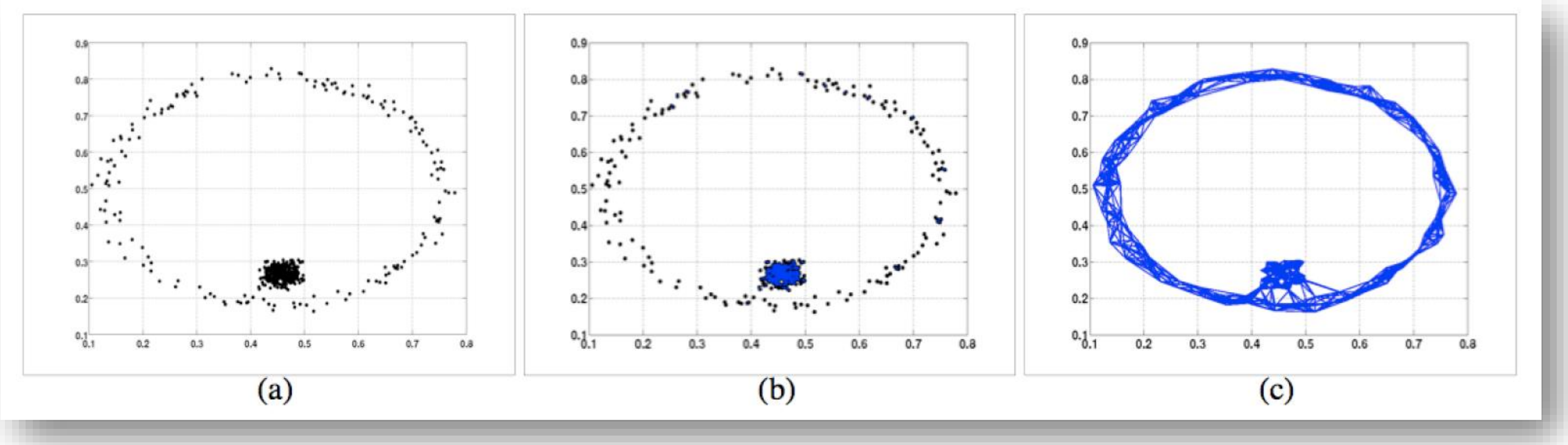


k-Nearest Neighbor (k-NN) Method

- Each node adds *k* links to its nearest neighbors.
- Some nodes in the network can have a very large number of neighbors or degree leading to irregular networks.
- The 1-NN network of a set of five points arranged in specific configuration:



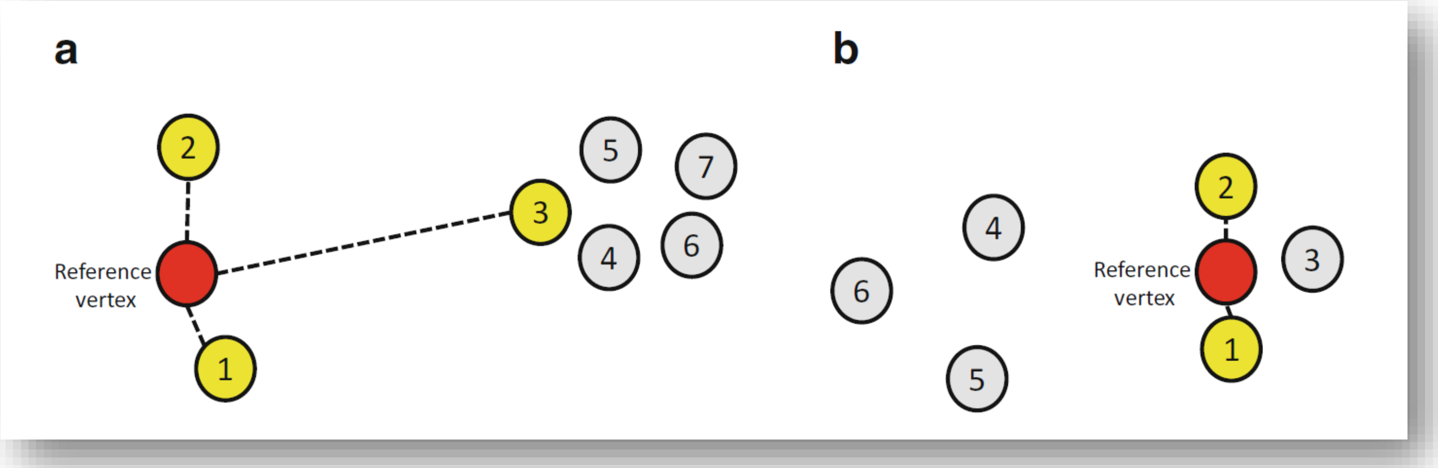
Comparison

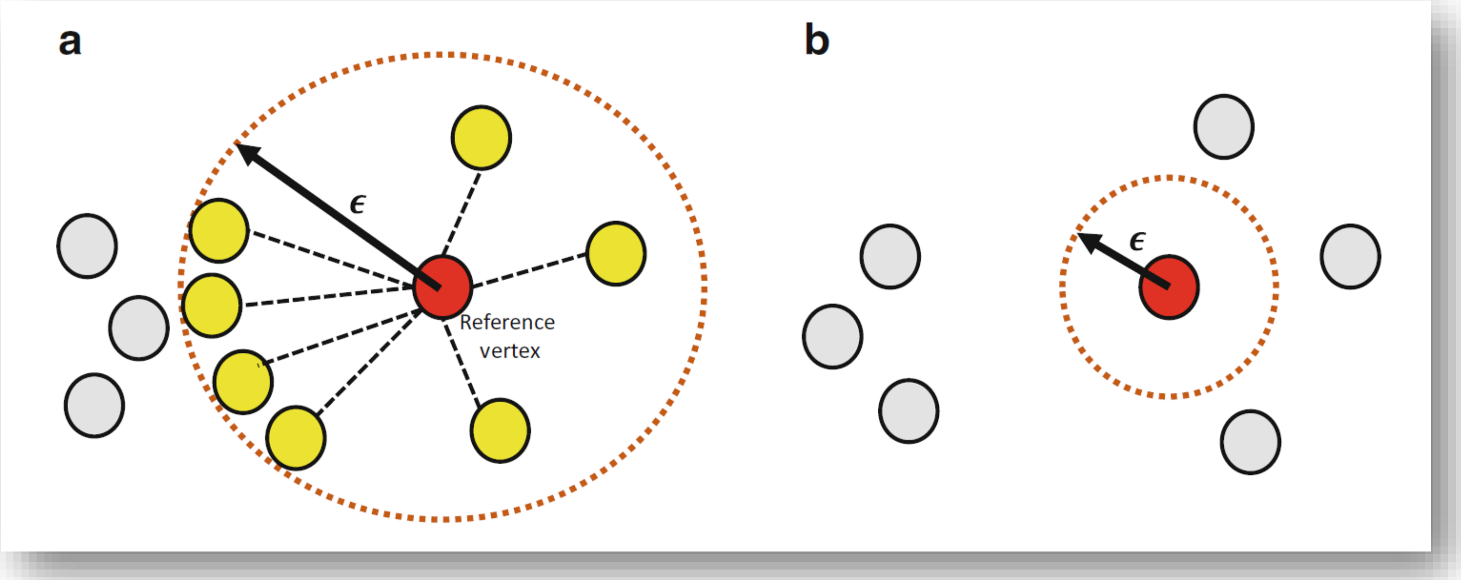


(a) (b) (c)

- *k*-NN and *ε*-radius networks constructed from a synthetic dataset:
  - (a) thee synthetic dataset,
  - (b) ε-radius network,
  - (c) k-NN network (k = 10).
- The *ε*-radius is quite sensitive to the choice of *ε* and it may return networks with disconnected components as in (b).

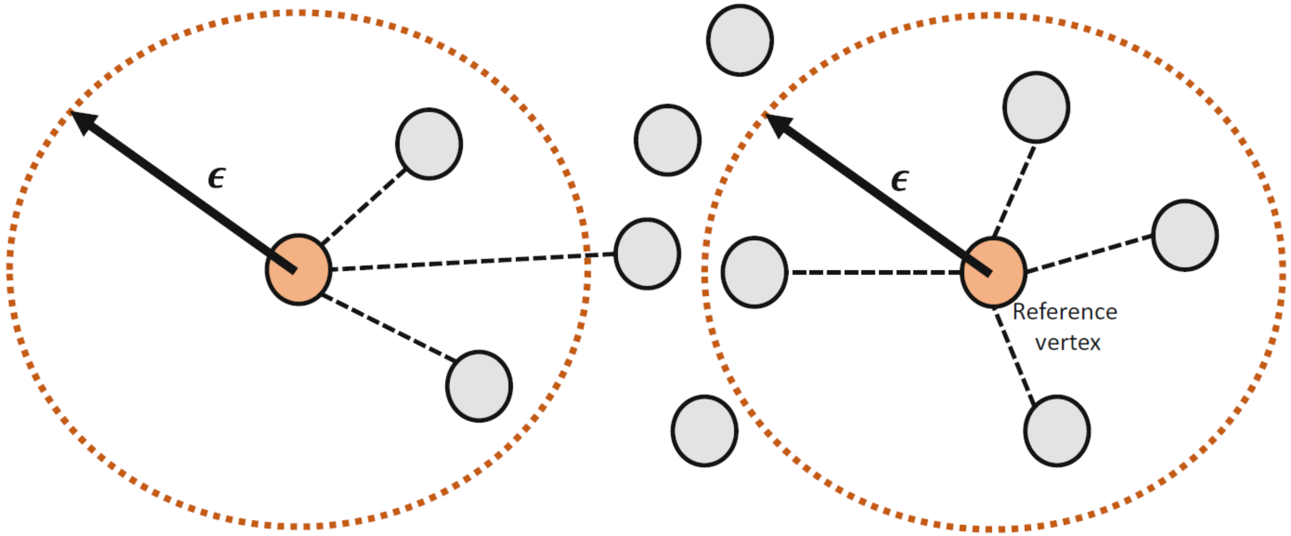
Limitations





Combination

$$\mathcal{N}(v_i) = \begin{cases} \epsilon\text{-radius}(v_i), & \text{if } |\epsilon\text{-radius}(v_i)| > k \\ k\text{-NN}(v_i), & \text{otherwise} \end{cases}$$



Declared as in a sparse region:
Uses ***k***-NN

Declared as in a dense region:
Uses **ϵ**-radius

b-matching method

- Opposed to the *k*-NN network, the b-matching network ensures that each node in the network has the same number of edges.

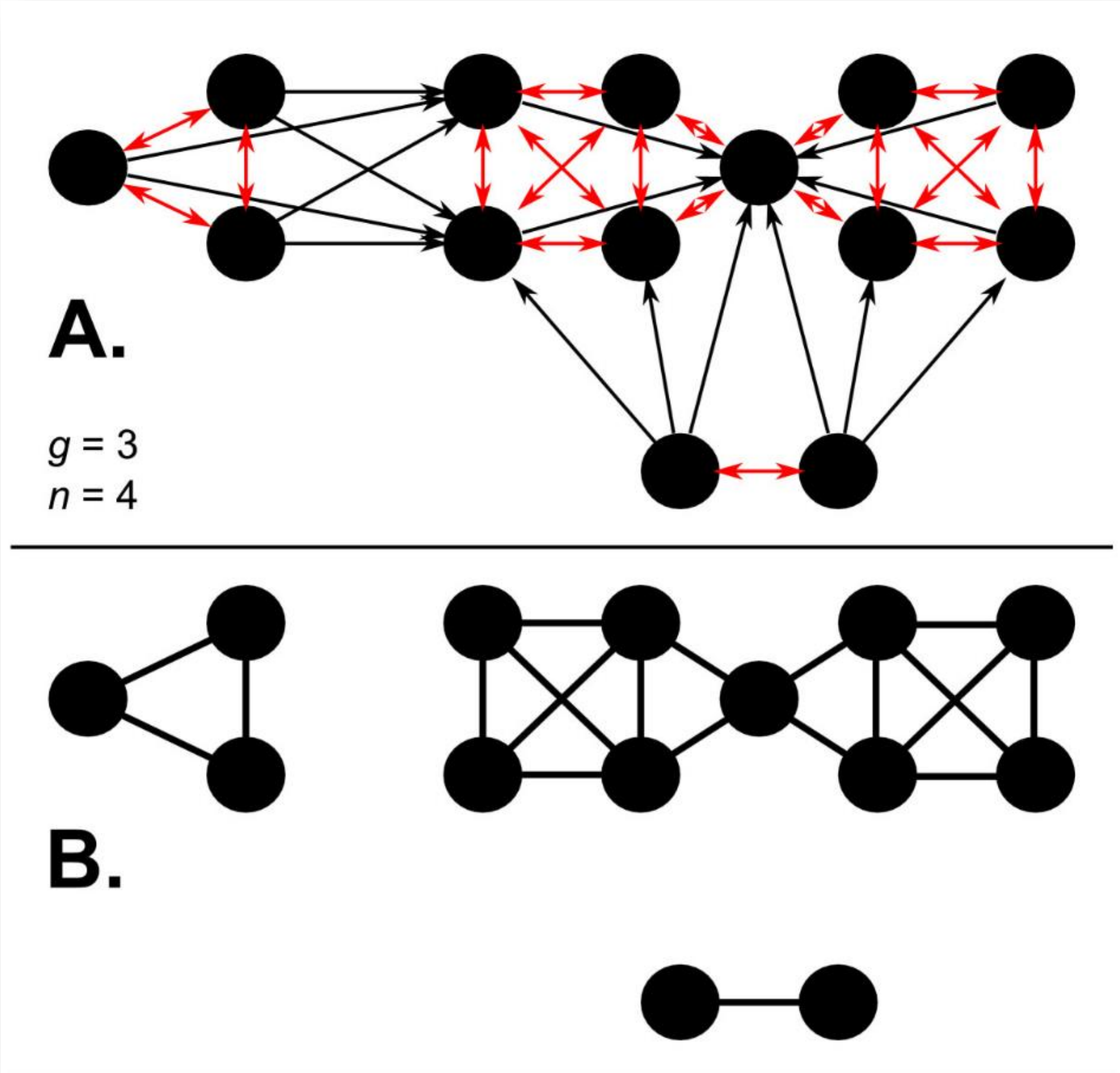
- Therefore, it produces a balanced or regular network.

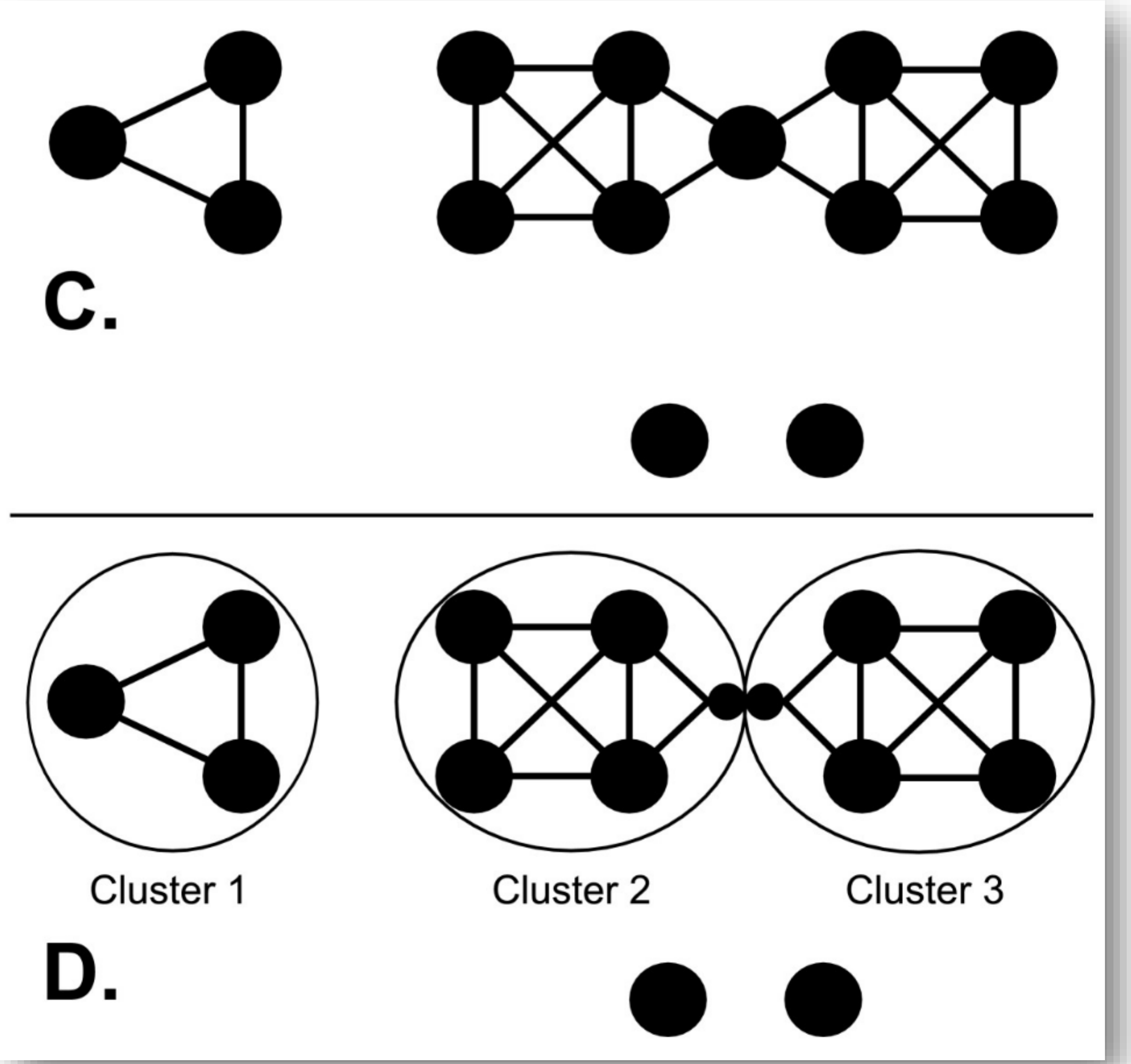
- However, the constructed network does not have real-world properties

Nearest Neighbor Networks (NNN)

- Network construction combined with clustering
- Algorithm
  - For each object, the n nearest neighbors are calculated based on the similarity measure.
  - If objects are considered to be nodes in a network, this results in a directed network in which each node is of out degree n.
  - An undirected network is then constructed by connecting any two object o_i and o_j such that $o_i \in N(o_j)$ and $o_j \in N(o_i)$, i.e. the two objects are mutual nearest neighbors.
  - All cliques (complete subgraphs) of size g within this network are identified, and overlapping cliques are merged to produce preliminary networks representing potential clusters.

Huttenhower, C., Flamholz, A. I., Landis, J. N., Sahi, S., Myers, C. L., Olszewski, K. L., ... Coller, H. A. (2007). Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *Bmc Bioinformatics*, *8*(1), 250.

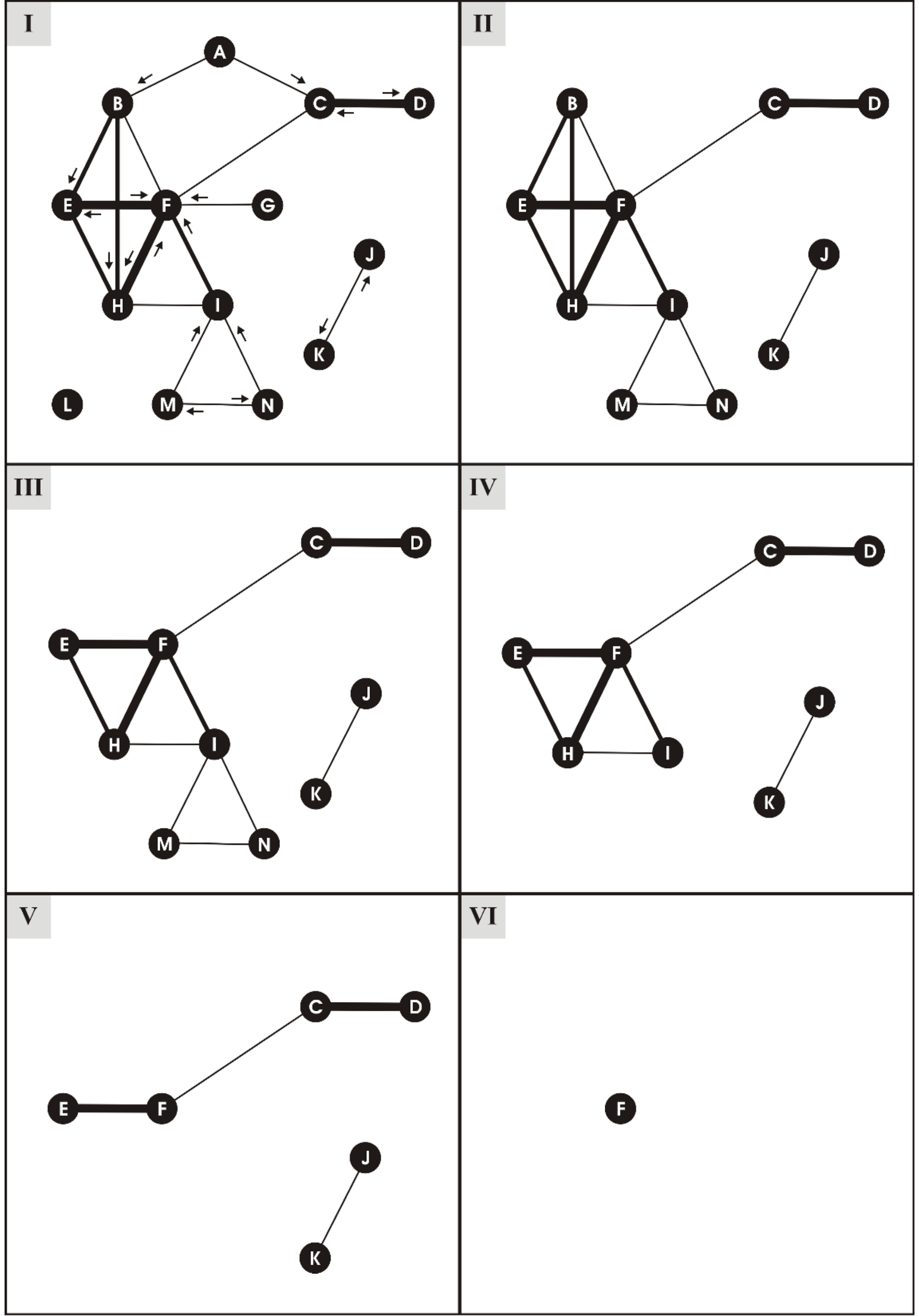

A.
g = 3
n = 4
B.


C.
Cluster 1
Cluster 2
Cluster 3
D.

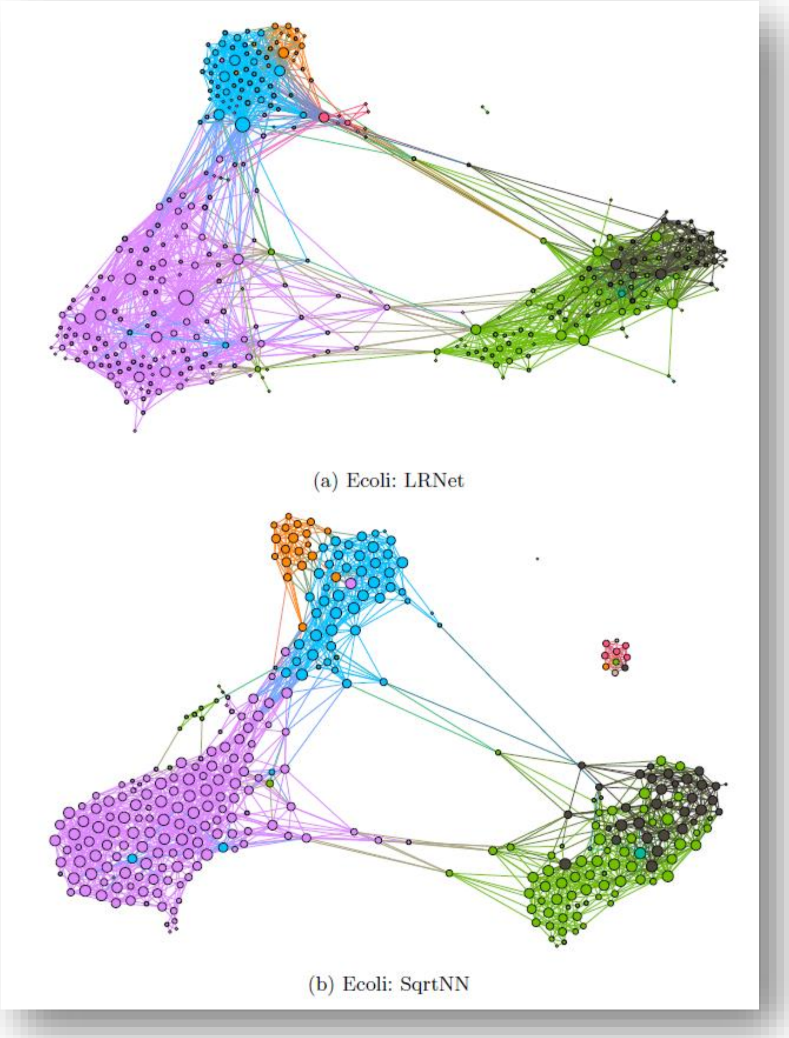
LRNet Algorithm

- The LRNet algorithm for the construction of the weighted graph utilizing local representativeness is composed of four steps:
    - Create a similarity matrix *S* of dataset *D*.
    - Calculate the representativeness of all objects O_i.
    - Create the set *V* of nodes of graph *G* so that node v_i of graph *G* represents object O_i of dataset *D*.
    - Create the set of edges *E* of graph *G* so that *E* contains an edge e_{ij} between nodes v_i and v_j $(i \neq j)$ if O_j is the representative neighbor of O_i. Nodes add a different number of links to the network.
- Representativeness is a kernel function *f(N, NN)*. For node *v*, *N* is a number of its neighbors (degree), *NN* is a number of nodes for which v is the nearest neighbor.

Ochodkova, E., Zehnalova, S., Kudelka, M. (2017). Graph Construction Based on Local Representativeness. In *International Computing and Combinatorics Conference* (pp. 654-665). Springer

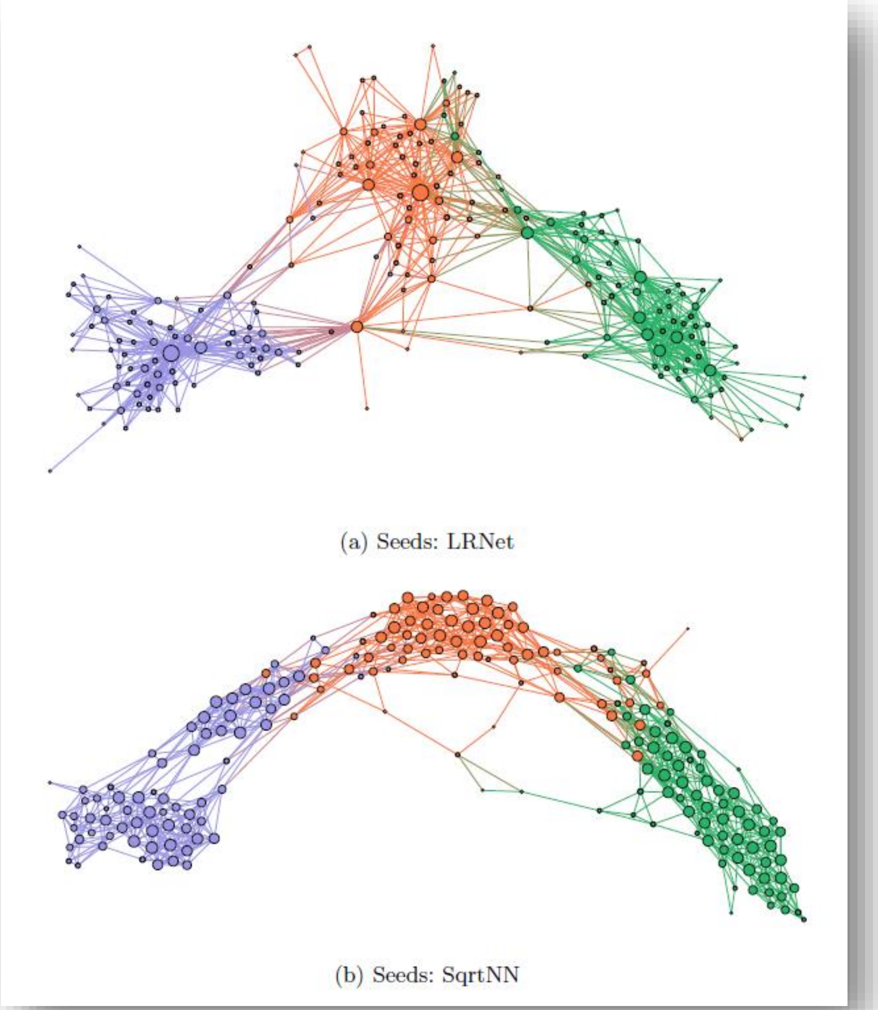

I
A
B
C
D
E
F
G
H
I
J
K
L
M
N
II
B
C
D
E
F
H
I
J
K
M
N
III
C
D
E
F
H
I
J
K
M
N
IV
C
D
E
F
H
I
J
K
V
C
D
E
F
J
K
VI
F

Biological Data



(a) Ecoli: LRNet

(b) Ecoli: SqrtNN



(a) Seeds: LRNet

(b) Seeds: SqrtNN

1-NN Co-occurrence Networks

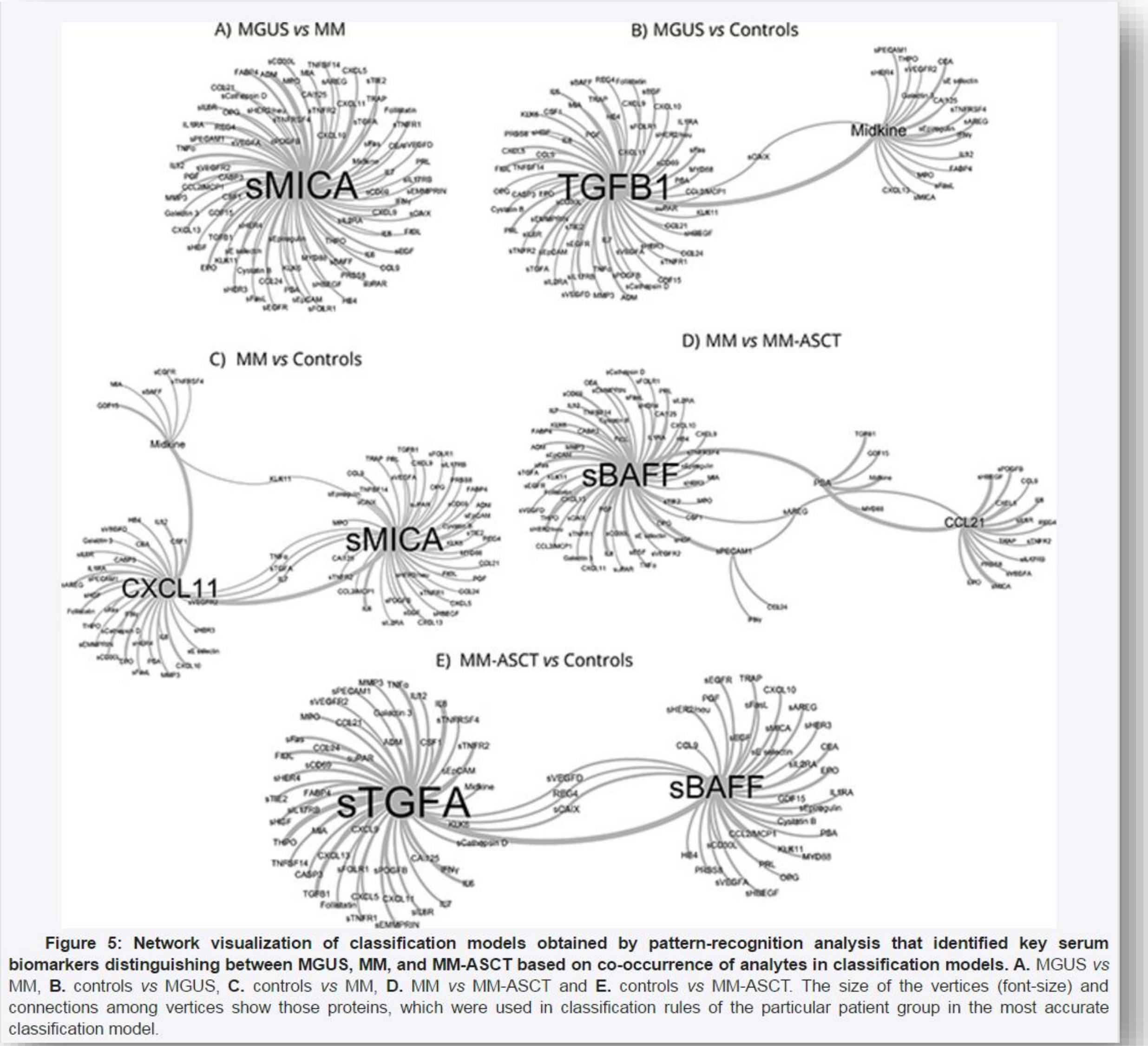
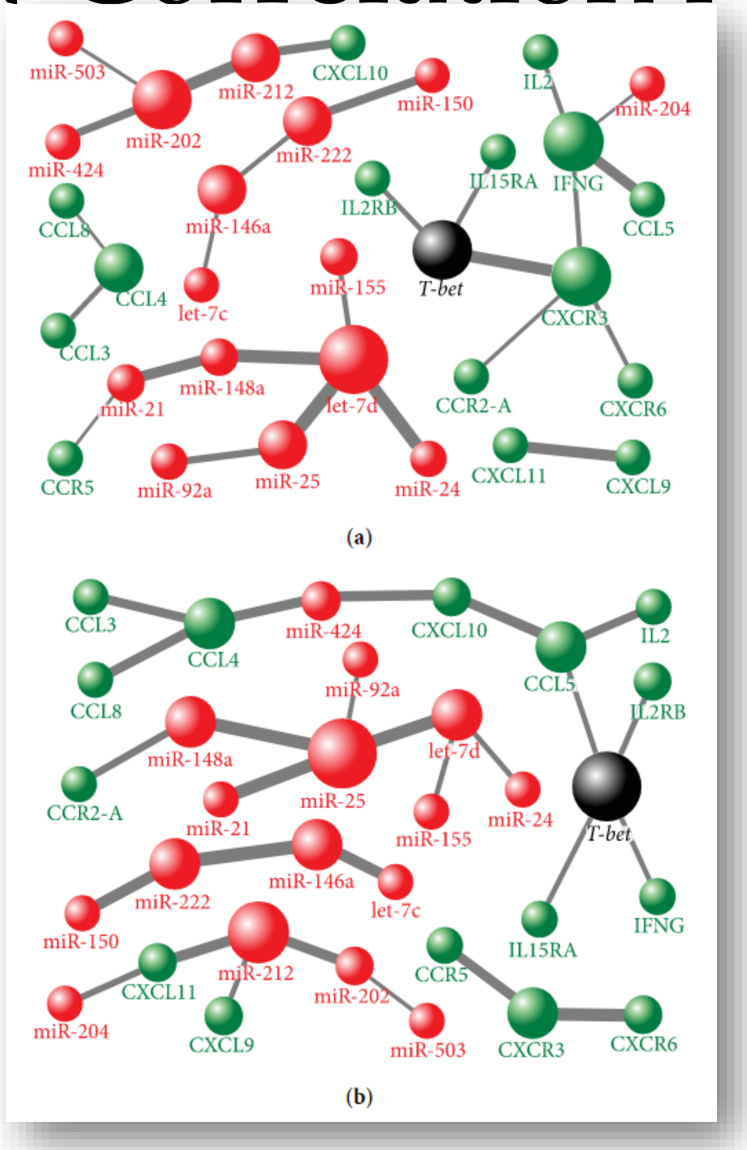
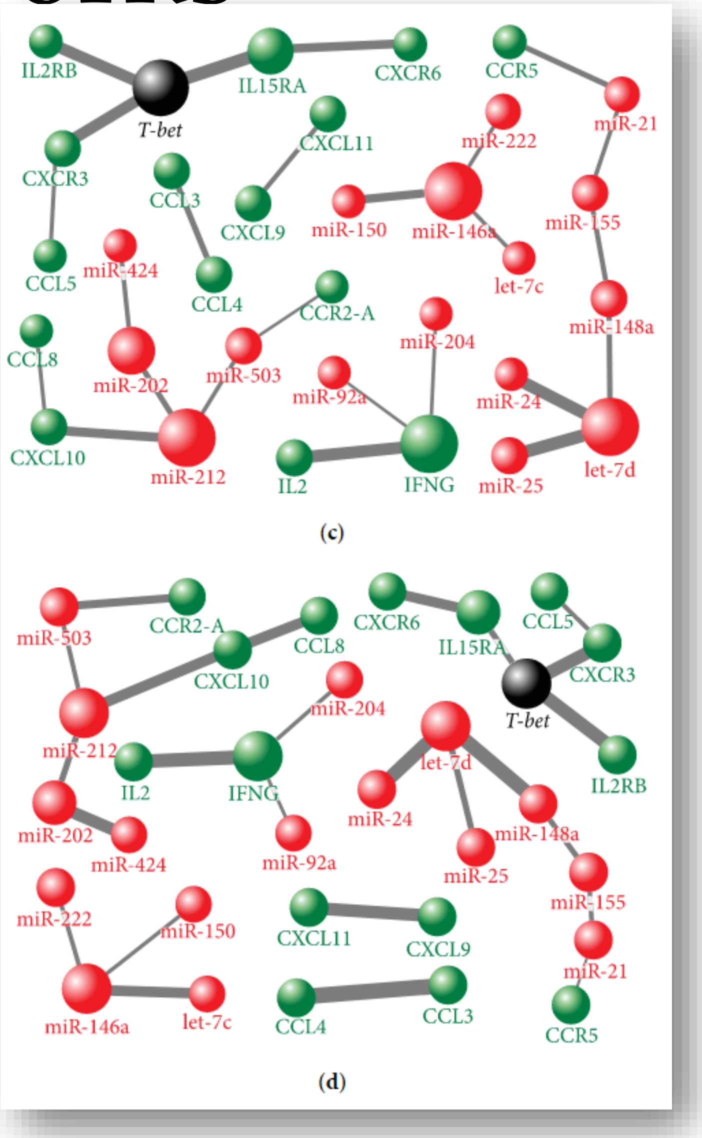


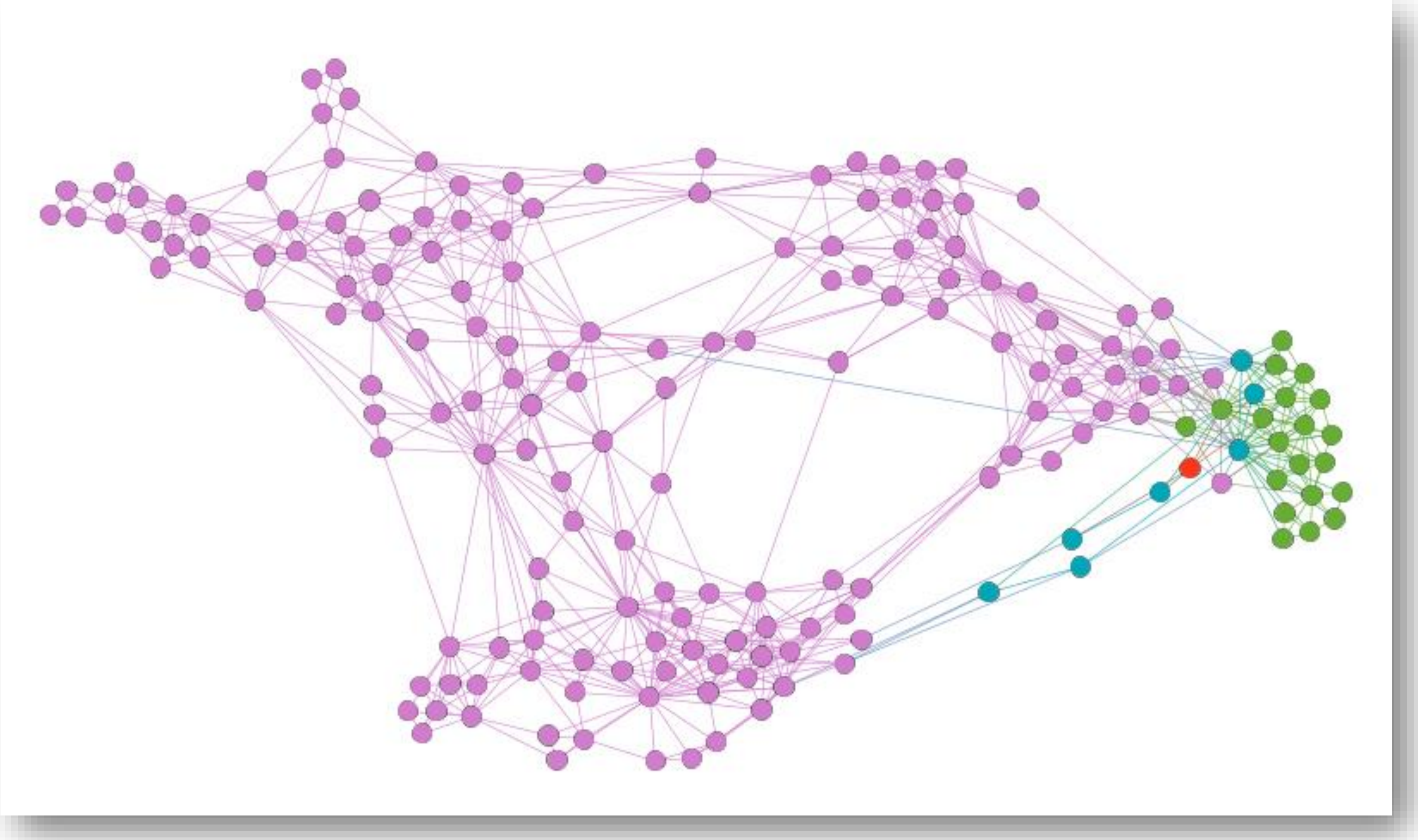
Figure 5: Network visualization of classification models obtained by pattern-recognition analysis that identified key serum biomarkers distinguishing between MGUS, MM, and MM-ASCT based on co-occurrence of analytes in classification models. A. MGUS *vs* MM, **B.** controls *vs* MGUS, **C.** controls *vs* MM, **D.** MM *vs* MM-ASCT and **E.** controls *vs* MM-ASCT. The size of the vertices (font-size) and connections among vertices show those proteins, which were used in classification rules of the particular patient group in the most accurate classification model.

1-NN Correlation Networks





LRNet – 3 Classes of Patients

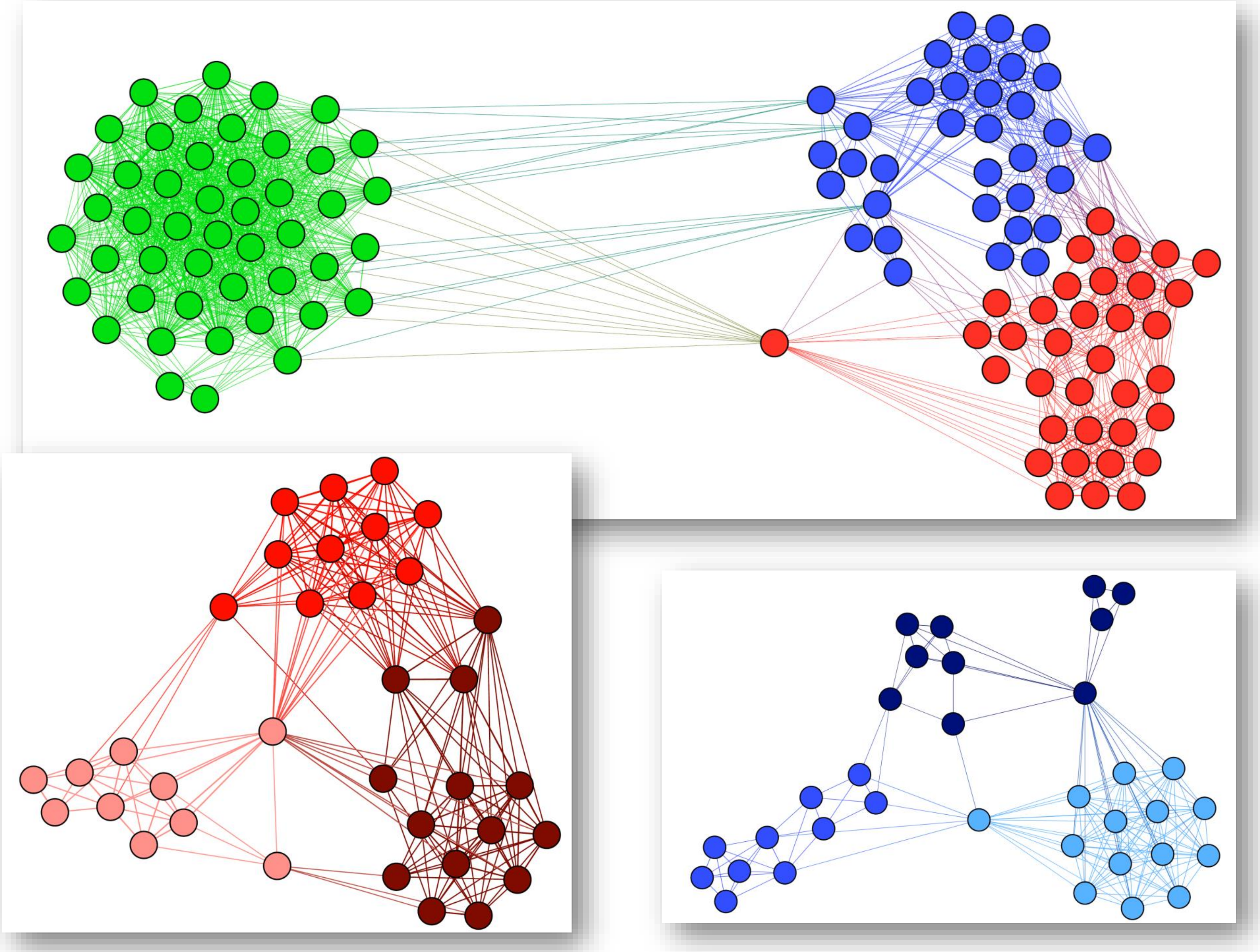


Quality of clustering

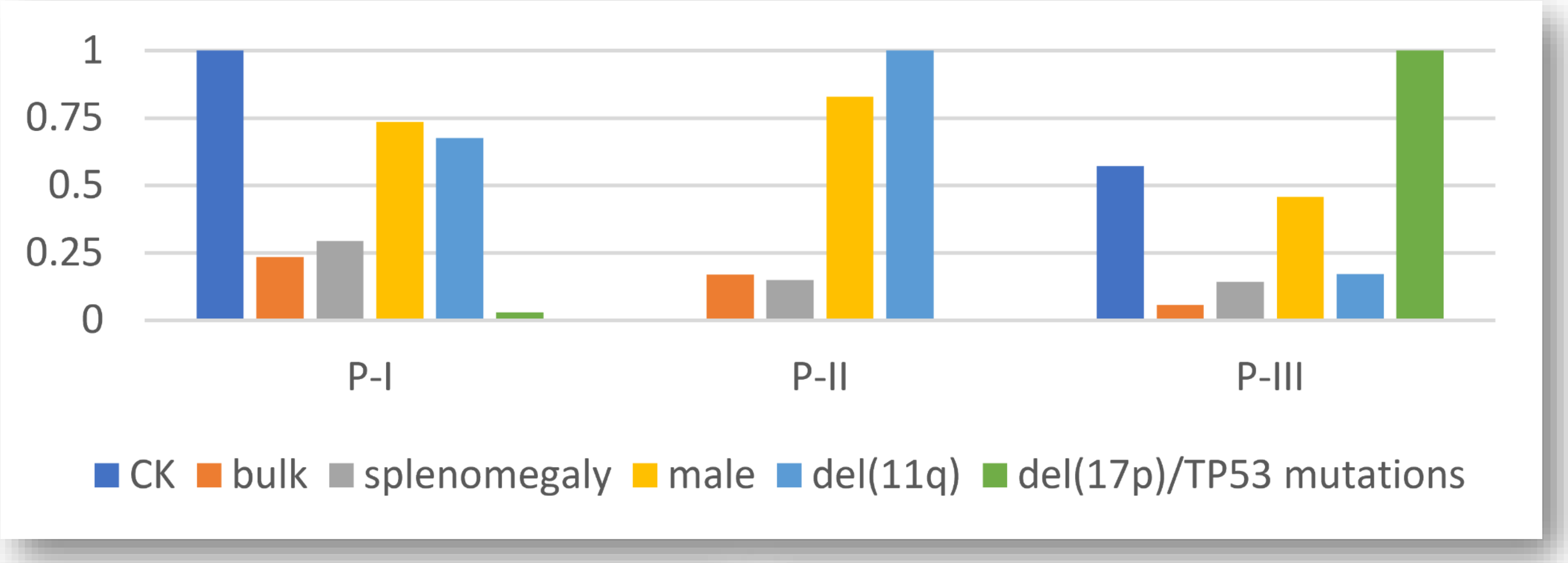
- Modularity (Louvain algorithm)

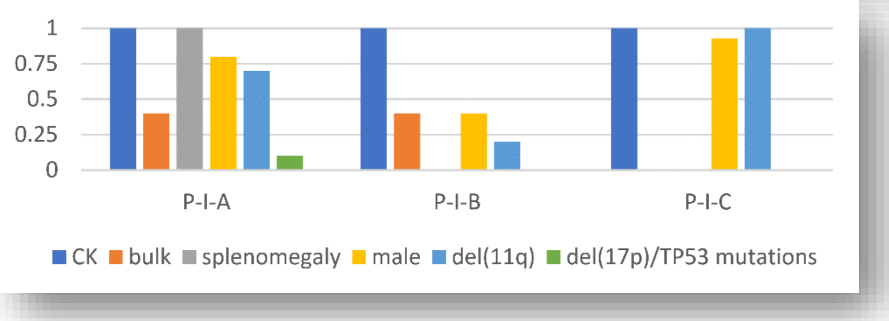
- Silhouette

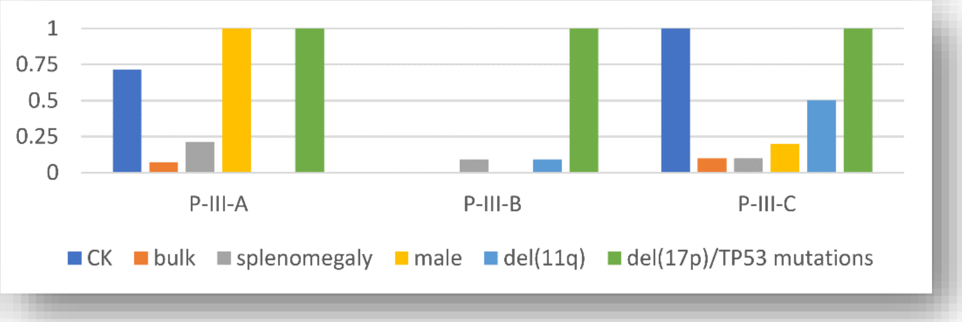
- Average values of markers (confidence intervals)



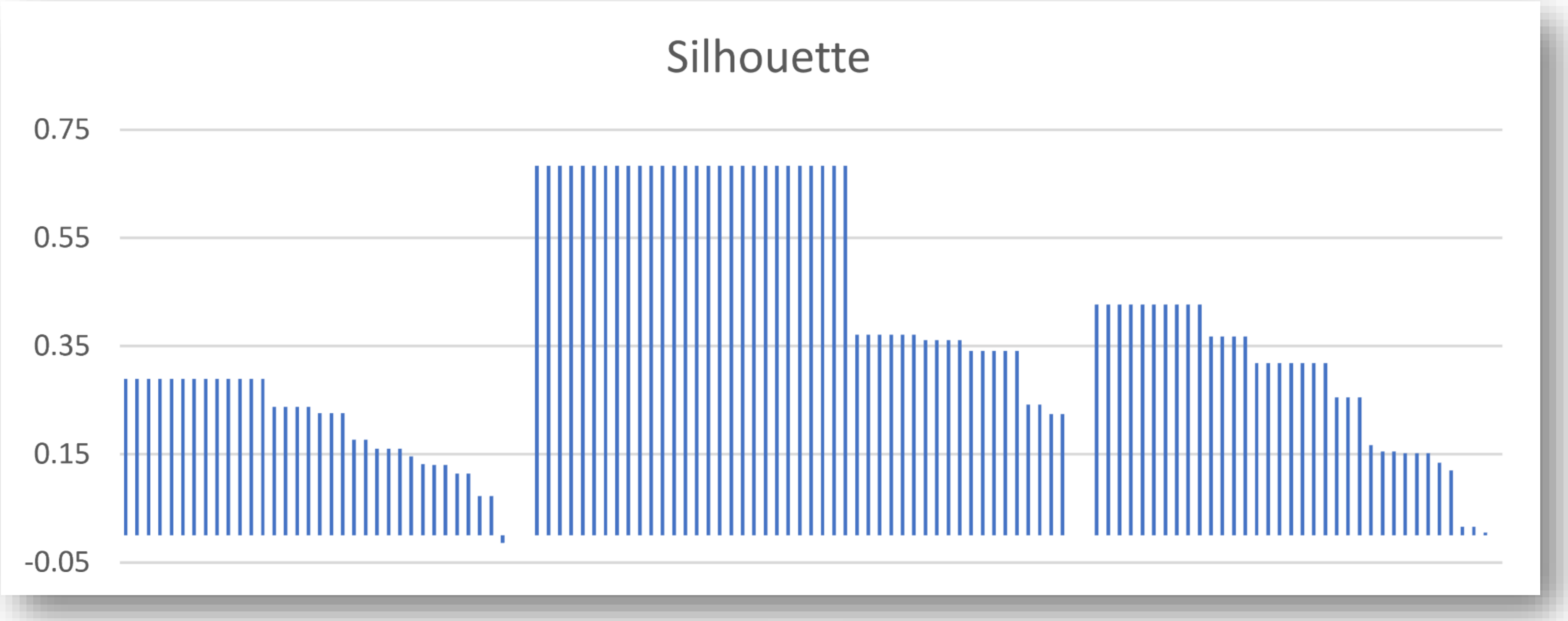
Profiles

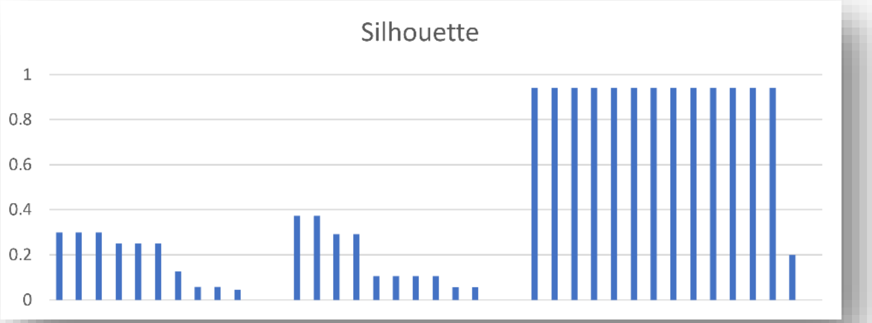

1
0.75
0.5
0.25
0
P-I
P-II
P-III
CK
bulk
splenomegaly
male
del(11q)
del(17p)/TP53 mutations

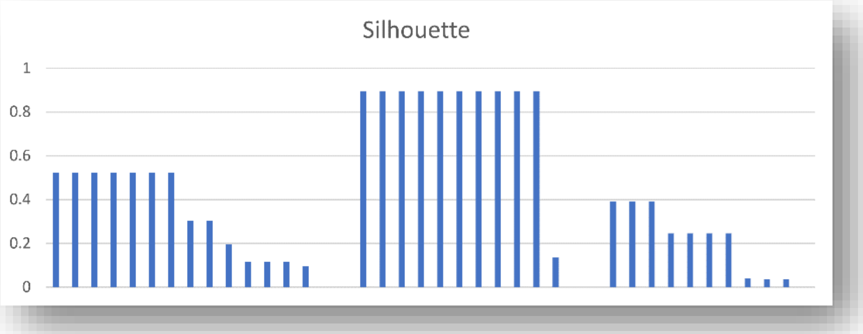

1
0.75
0.5
0.25
0
P-I-A
P-I-B
P-I-C
CK
bulk
splenomegaly
male
del(11q)
del(17p)/TP53 mutations


1
0.75
0.5
0.25
0
P-III-A
P-III-B
P-III-C
CK
bulk
splenomegaly
male
del(11q)
del(17p)/TP53 mutations

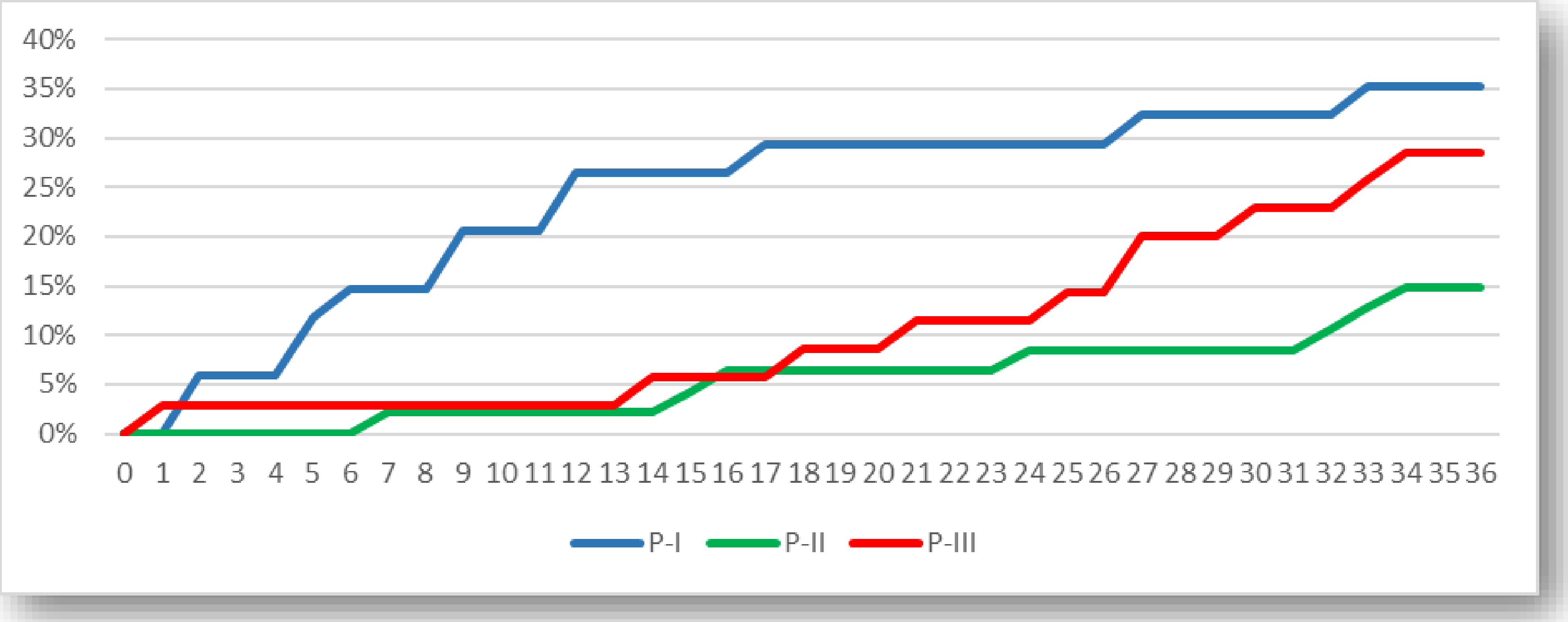
Silhouette

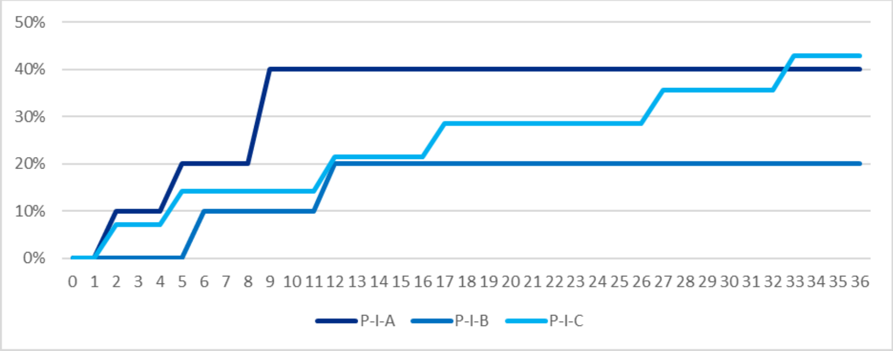

Silhouette
0.75
0.55
0.35
0.15
-0.05

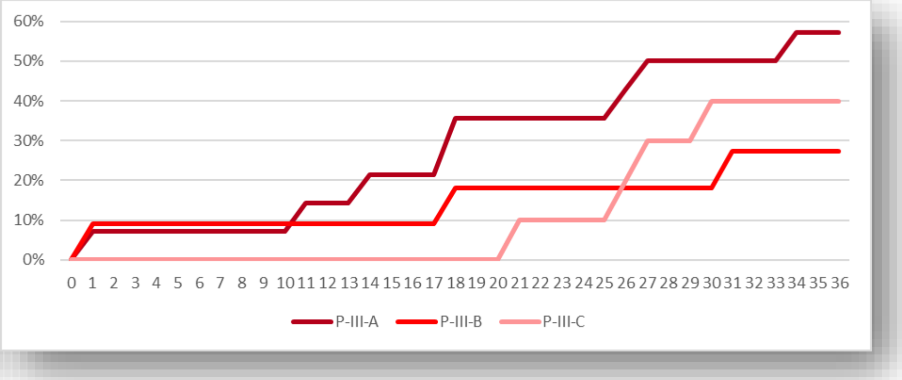

Silhouette
1
0.8
0.6
0.4
0.2
0


Silhouette
1
0.8
0.6
0.4
0.2
0

Empirical probability of death







Summary

- There are different methods of network construction with different results (regular x real-world properties).

- Combination of network construction and network clustering is suitable tool for visual data mining. However, there are limitations (time complexity).

- Asymptotic time complexity is at least $O(n^2)$, $O(n^3)$ in case of high dimension.

Assignment

- Implement eps-radius, k-NN, and a combined method of network construction from vector data.
- Apply all the implemented methods to the iris dataset and visualize the results; use Gephi system or visualization libraries for R or Python.

Sources

- Silva, T. C., Zhao, L. (2016). *Machine learning in complex networks* (Vol. 2016). Springer.

- Huttenhower, C., Flamholz, A. I., Landis, J. N., Sahi, S., Myers, C. L., Olszewski, K. L., ... Coller, H. A. (2007). Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *Bmc Bioinformatics*, *8*(1), 250. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-250

- Subramanya, A., Talukdar, P. P. (2014). Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *8*(4), 9-11.

- Ochodkova, E., Zehnalova, S., Kudelka, M. (2017). Graph Construction Based on Local Representativeness. In *International Computing and Combinatorics Conference* (pp. 654-665). Springer.