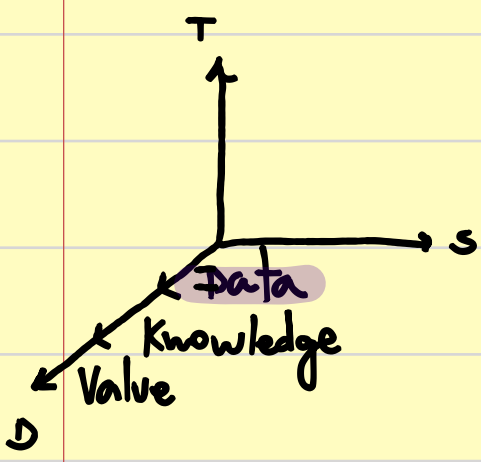


Datasets can be classified in 2 groups:

- EUCLIDEAN DATASETS
- NON-EUCLIDEAN DATASETS

Examples of Datasets:

(1) • N one-dimensional time-series. EUCLIDEAN DATASET.

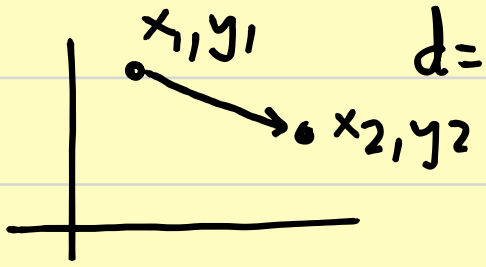
	KPI_1	KPI_2	KPI_3	...	KPI_n
0	...				
1	...				
2	...				
.					
.					
.					
1000	...				

1D × N

|||

VECTOR × N

For 2D (N=2)



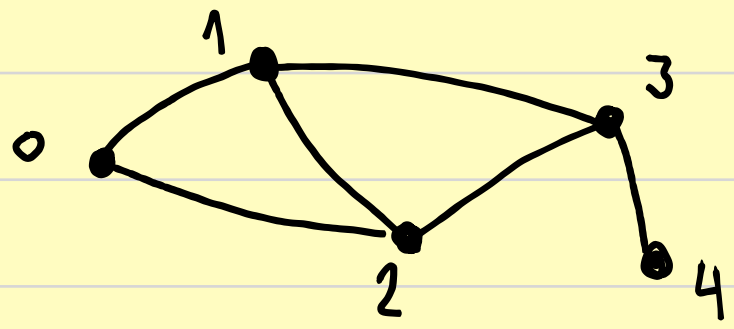
$$d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

EUCLIDEAN DISTANCE

- The hypothesis is that the underlying structure of the dataset allows for the calculation of a distance.
- This implies there is a way to measure "distance" between points in the space.

(2)

- NETWORK. NON-EUCLIDEAN DATASET.

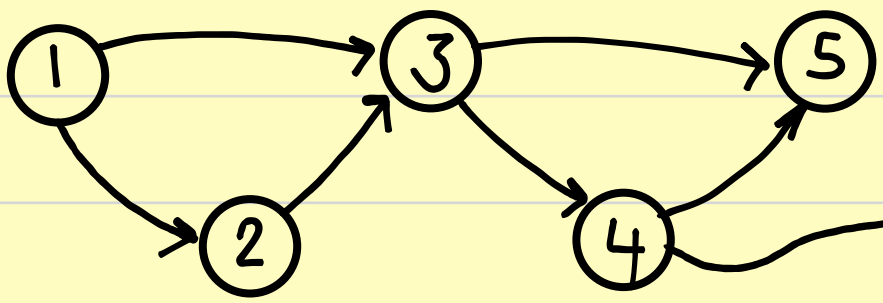


Nodes + Edges + Attributes of both Node & Edges

A network is defined (mathematically) by a set (group) of nodes and a set of edges. This group of nets is called „Graph" (G):

Netzwerk ≡ Graph ≡ G(NUE) N nodes, E edges

Beispiel:



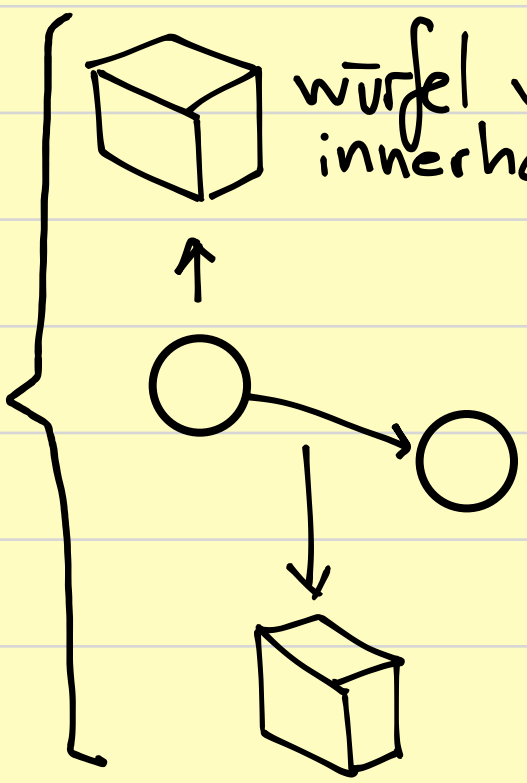
Graph

N: {1,2,3,4,5,6} E: {(1→2); (2→3); (3→4); (1→3); (3→5); (4→5); (4→6)}

(3) G = {NUE}

Beispiel:

Euklydische & nicht Euklydische Datensätze Kombiniert



würfel von euklydischen Daten innerhalb von jeder Knote.

3 KENNZAHLEN UM NETZWERKE MESSBAR ZU MACHEN

AVERAGE PATH LENGTH

Average distance (steps) between nodes in the network.

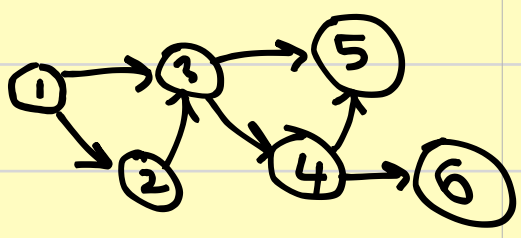
$$APL = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}$$

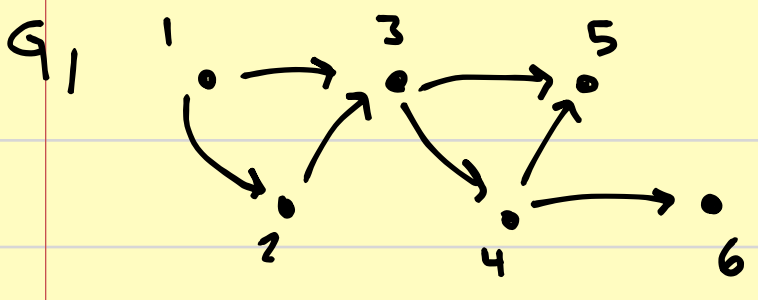
$N(N-1)$: Maximum # of relationships in the network.

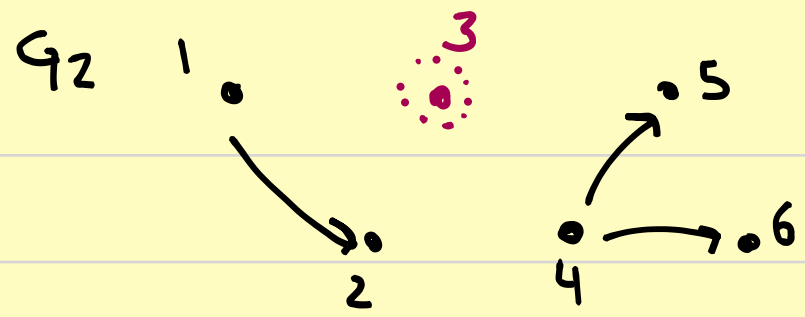
$\sum\sum d_{ij}$: sum of all paths between nodes.

$$APL = \frac{1}{6 \cdot 5} \Big[\underbrace{\big[d_{12} \; d_{13} \; d_{14} \; d_{15} \; d_{16} \big]}_{1} + \underbrace{\big[d_{21} \; d_{23} \; d_{24} \; d_{25} \; d_{26} \big]}_{2} + \underbrace{\big[d_{31} \; d_{32} \; d_{34} \; d_{35} \; d_{36} \big]}_{3} + \underbrace{\big[d_{41} \; d_{42} \; d_{43} \; d_{45} \; d_{46} \big]}_{4} + \underbrace{\big[d_{51} \; d_{52} \; d_{53} \; d_{54} \; d_{56} \big]}_{5} + \underbrace{\big[d_{61} \; d_{62} \; d_{63} \; d_{64} \; d_{65} \big]}_{6} \Big]$$

$$= \frac{1}{30} \Big[[1+1+2+2+3] + [1+1+2+2+3] + [1+1+1+1+2] + [2+2+1+1+1] + [2+2+1+1+2] + [3+3+2+1+2] \Big] = \dots$$

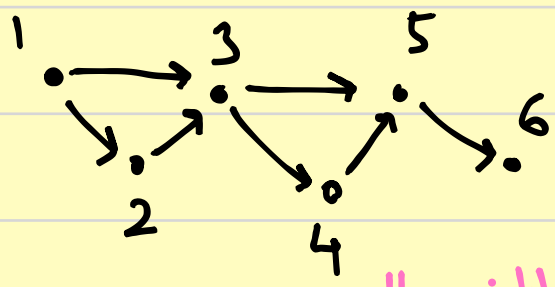






When we compare two Networks, the one with the SHORTEST APL would be usually faster / more effective.

CLUSTERING COEFFICIENT



CC describes how good groups are created in the network!

$$CC = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{2 \cdot L_i}{k_i(k_i - 1)}$$

$L_i \equiv$ Number of relationships between the neigbours of node „i"

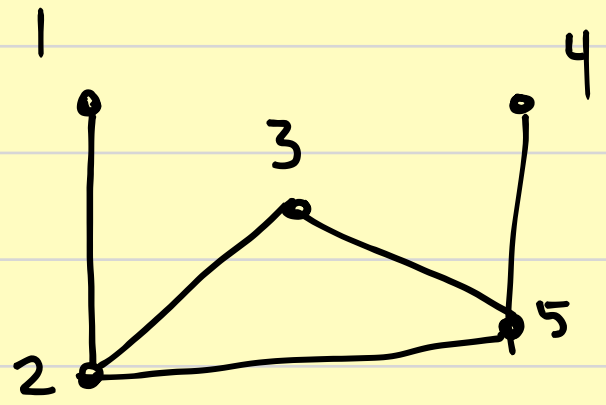
$k_i \equiv$ Number of neighbours of node „i"

L_1: the neighbours of node #1 have ONE connection

k_1: node#1 has TWO connections

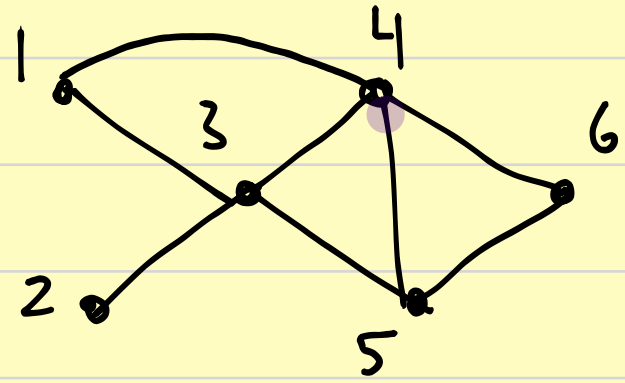
$$CC = \frac{1}{6} \cdot \left[\left[\frac{2 \cdot 1}{2 \cdot (2-1)} \right]_1 + \left[\frac{2 \cdot 1}{2(2-1)} \right]_2 + \left[\frac{2 \cdot 2}{4(4-1)} \right]_3 + \left[\frac{2 \cdot 1}{2(2-1)} \right]_4 + \left[\frac{2 \cdot 1}{3(3-1)} \right]_5 + \left[\frac{2 \cdot 0}{\cdot} \right]_6 \right]$$

Beispiel



$L_1=0$; $L_2=1$; $L_3=1$; $L_4=0$; $L_5=1$

Beispiel



$L_1=1$; $L_2=0$; $L_3=2$; $L_4=3$; $L_5=2$; $L_6=1$

the higher the clustering coefficient, the better the communication between the elements of the network.