Úvod do analýzy sociálních sítí

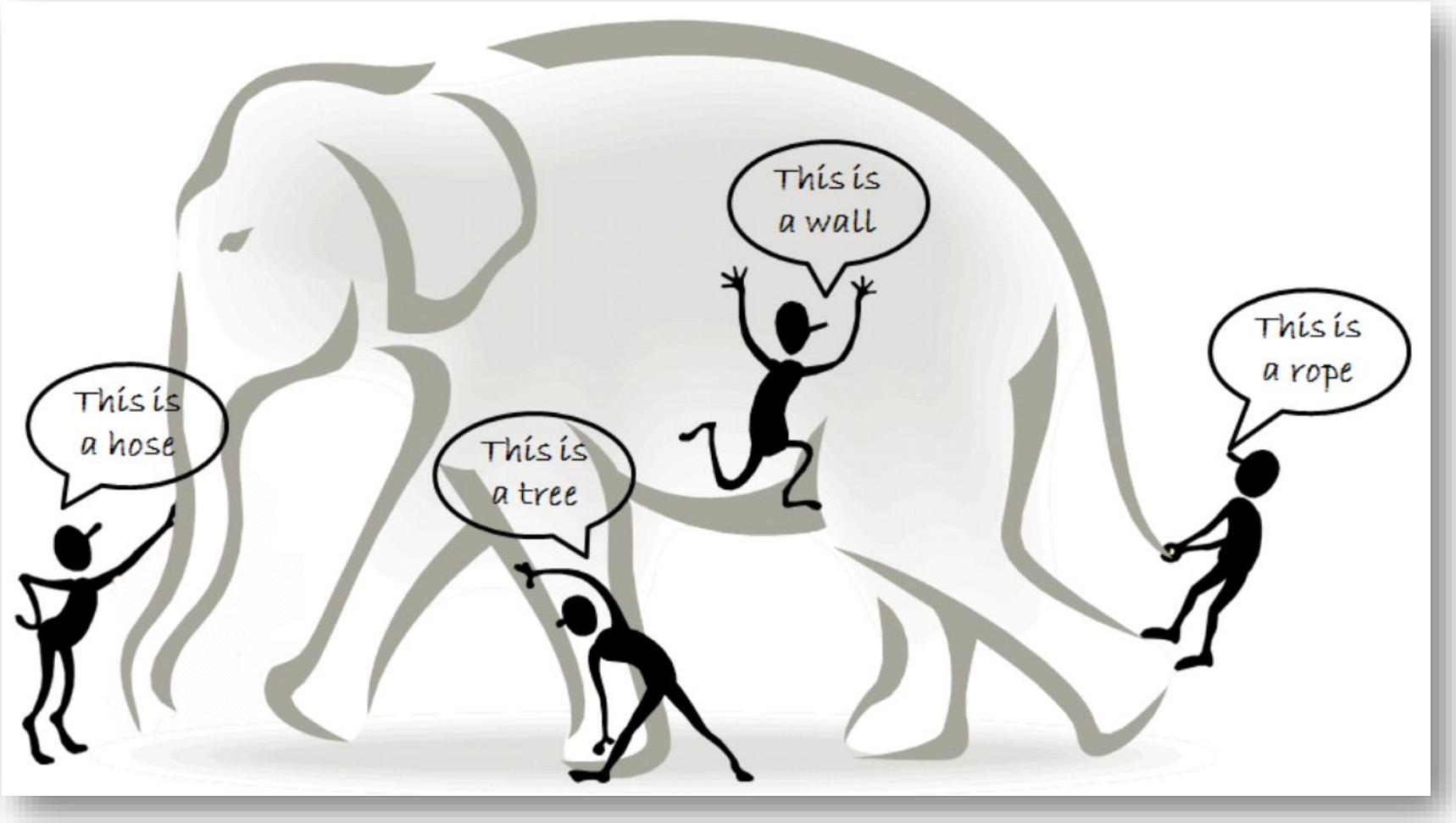
Konstrukce sítě z vektorových dat

2023-2024

Data – informace – znalosti

- Data, informace, znalosti:
  - **Data** nejsou dobře lidsky čitelná a přímočaře interpretovatelná.
  - **Informace** jsou čitelné, ale vyžadují správnou interpretaci.
  - **Znalosti** jsou výsledkem správné interpretace informací.
- K získání znalostí potřebujeme co nejúplnější pohled…

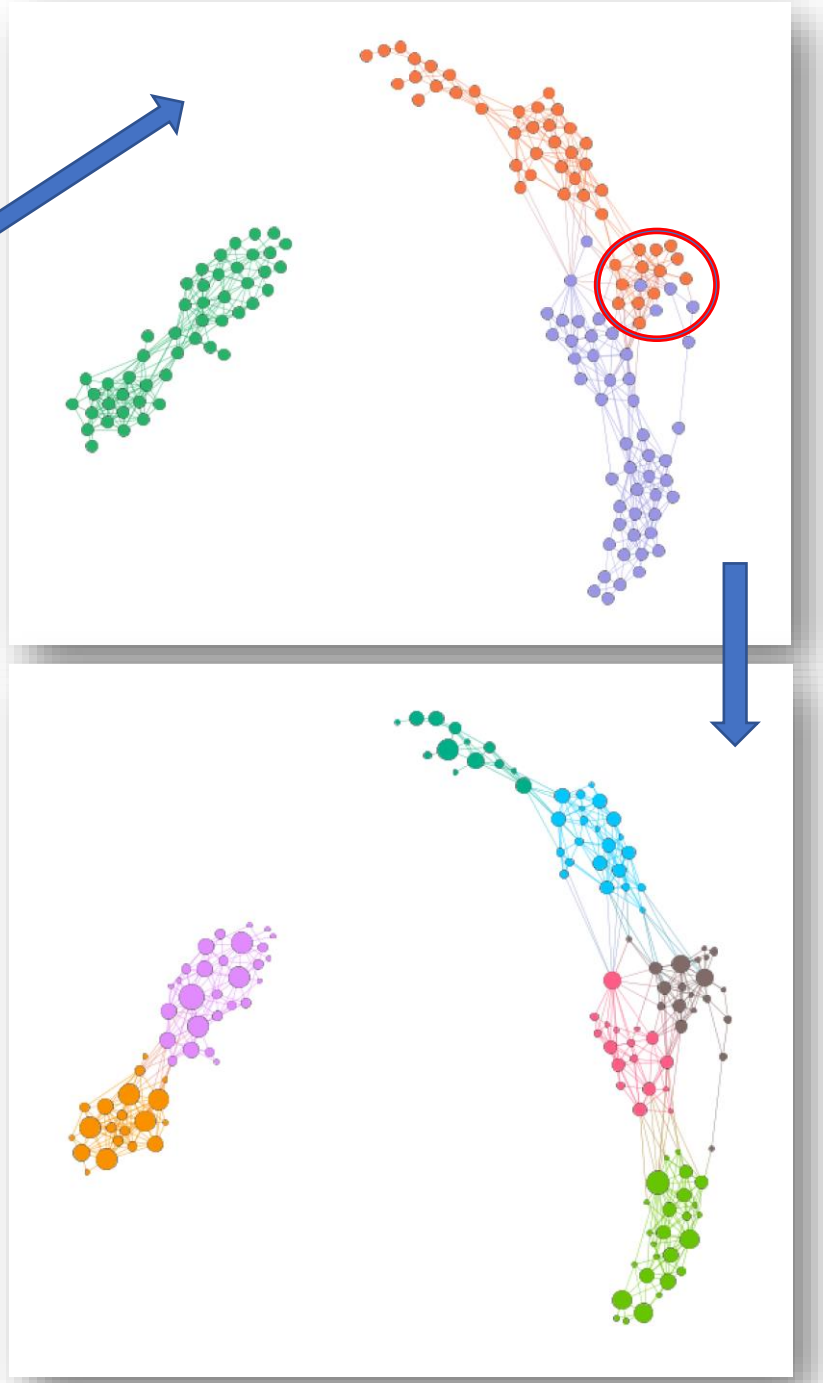
Copak to asi je?



Wu, X., Zhu, X., Wu, G. Q., Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering.

Od dat ke komplexním sítím

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
…					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
…					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
…					



Proč komplexní sítě?

- Často se podstatné věci ukážou až při pohledu na detaily.
- Komplexní sítě reprezentují komplexní systém. Tedy systém který nelze celý jednoduše popsat, protože v různých částech může mít různé vlastnosti (struktury, chování apod.).
- **Pro analýzu komplexních sítí je jednou z klíčových metod vizualizace.**
- Reálná data zaznamenávají realitu... A tradiční analytické přístupy mají problém s vizualizací.

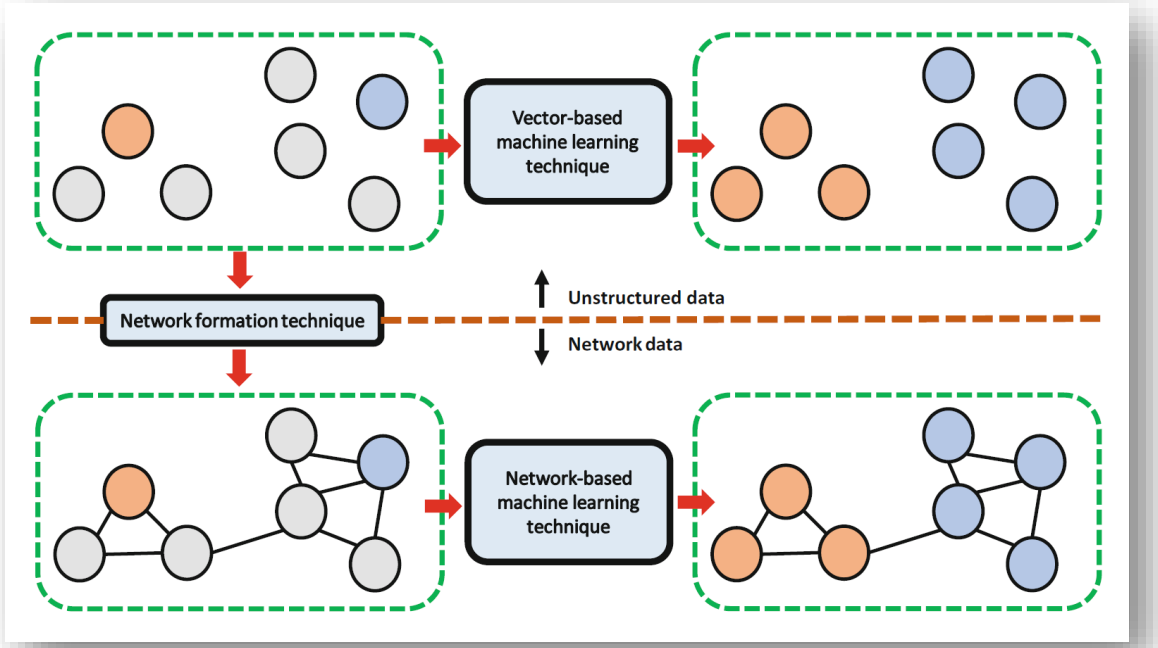
Vlastnosti komplexních sítí

- Vzorce i anomálie mohou dostat vizuální podobu.
- Lokálně mají skupiny vrcholů různou hustotu propojení reprezentovanou hranami mezi vrcholy.
- Sítě mají obvykle komunitní strukturu (oddělené nebo překrývající se shluky vrcholů).
- V sítích existují vrcholy s různou mírou důležitosti (centralita, reprezentativnost).

Jak na to?

- Nejdříve se musí z vektorových dat zkonstruovat síť, ve které vrcholy reprezentují instance a hrany podobnost mezi instancemi.

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
…					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
…					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
…					



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

Krok 1: Matice vzdálenosti/podobnosti

- Jak na to?
  - Eukleidovská vzdálenost
  - Gaussova funkce (převod Eukleidovské vzdálenosti na podobnost)
  - Kosínová míra (proporční podobnost)
  - Korelace
  - Společný výskyt (co-occurrence, Jaccard)
- Předpokládá se symetrická podobnost, je proto nutné spočítat podobnost každého s každým (kvadratická složitost). Výsledkem je symetrická matice reprezentující tzv. úplný graf s $n * (n-1) / 2$ hranami.

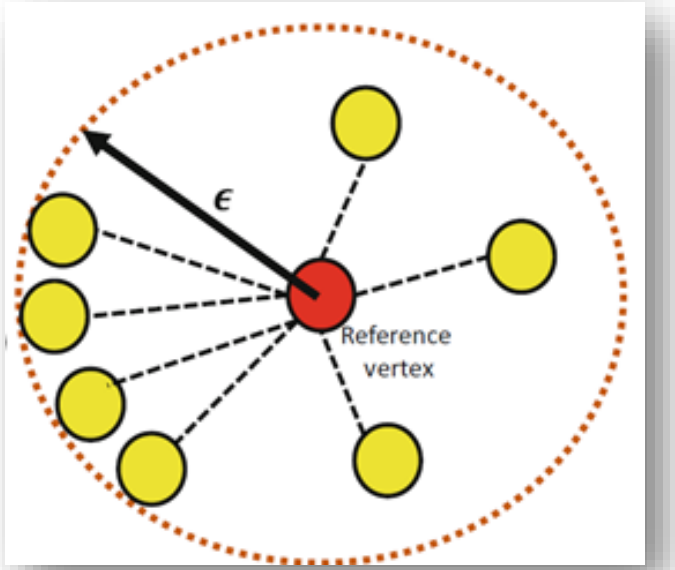
Krok 2: Výběr hran

- Do analýzy se berou v úvahu všechny instance. Proto počet vrcholů sítě odpovídá počtu záznamů v datové sadě.

- Musíme zvolit algoritmus či pravidlo, na jehož základě s využitím matice podobnosti vybereme hrany do sítě. Výsledkem by měl být řídký graf (počet hran je malým násobkem počtu vrcholů).

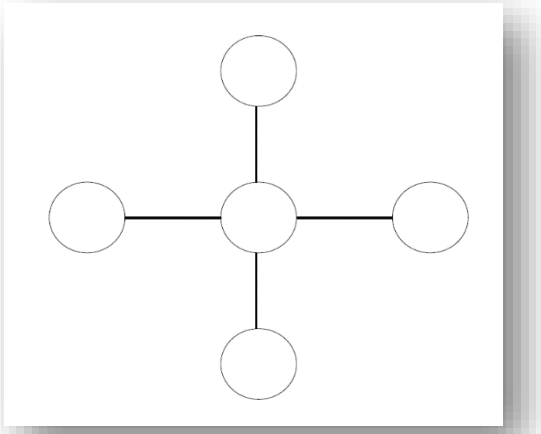
Krok 3_algA: ε-okolí

- Hrana je mezi vrcholy přidána tehdy, pokud je podobnost větší (vzdálenost menší) než předdefinovaná hodnota $\varepsilon > 0$.



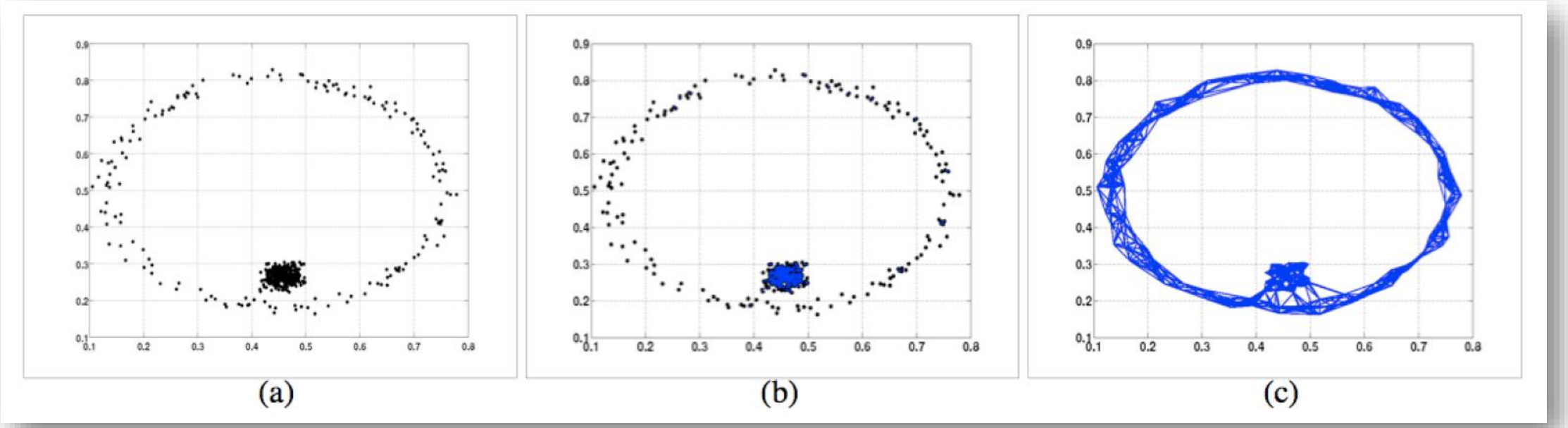
Krok 3_algB: *k* nejbližších sousedů (k-NN)

- Každý vrchol je spojen hranou s *k* vrcholy, se kterými má největší podobnost.
- Některé vrcholy mohou mít hodně sousedů, jiné málo.
- Pro *k = 1* to může dopadnout takto:

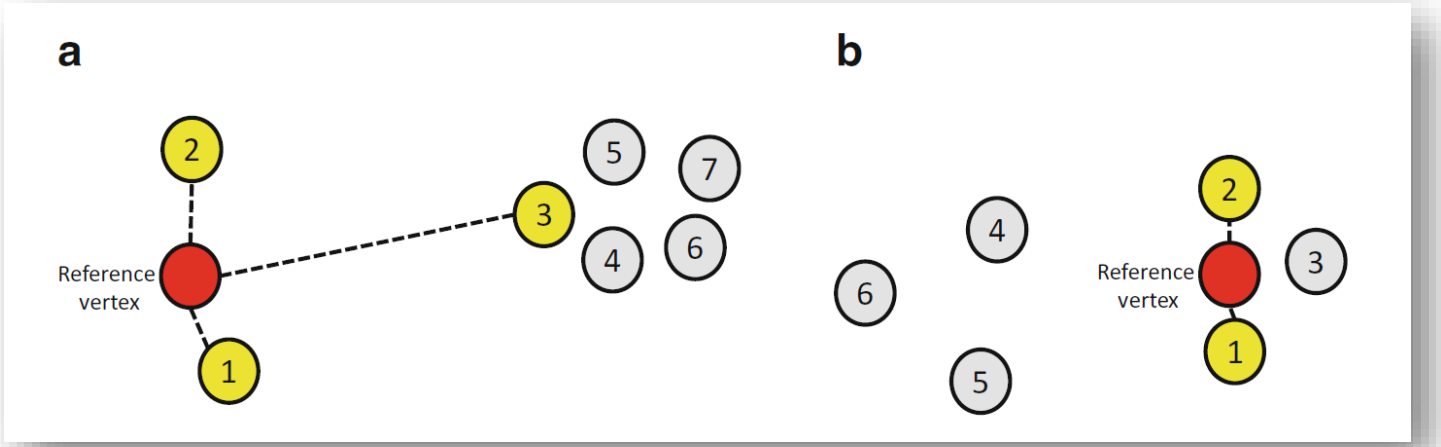


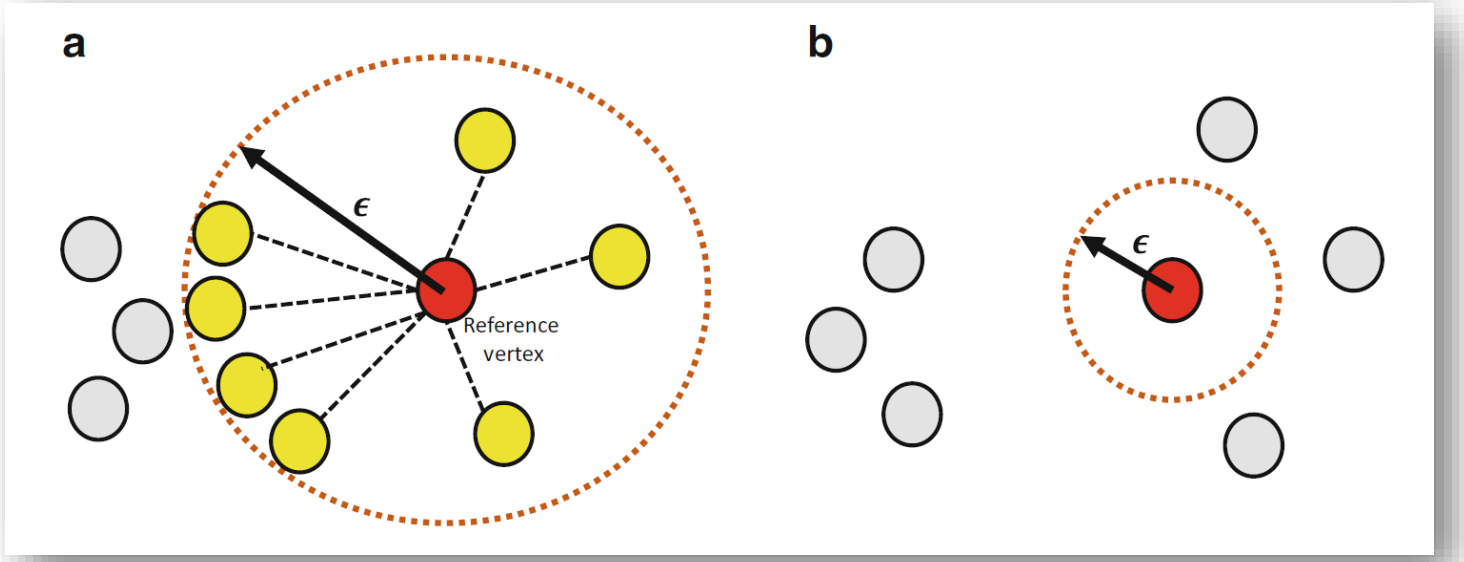
Srovnání

- *k*-NN and ε-radius sítě:
  - (a) umělá datová sada,
  - (b) ε-okolí síť,
  - (c) *k*-NN síť pro *k = 10*.



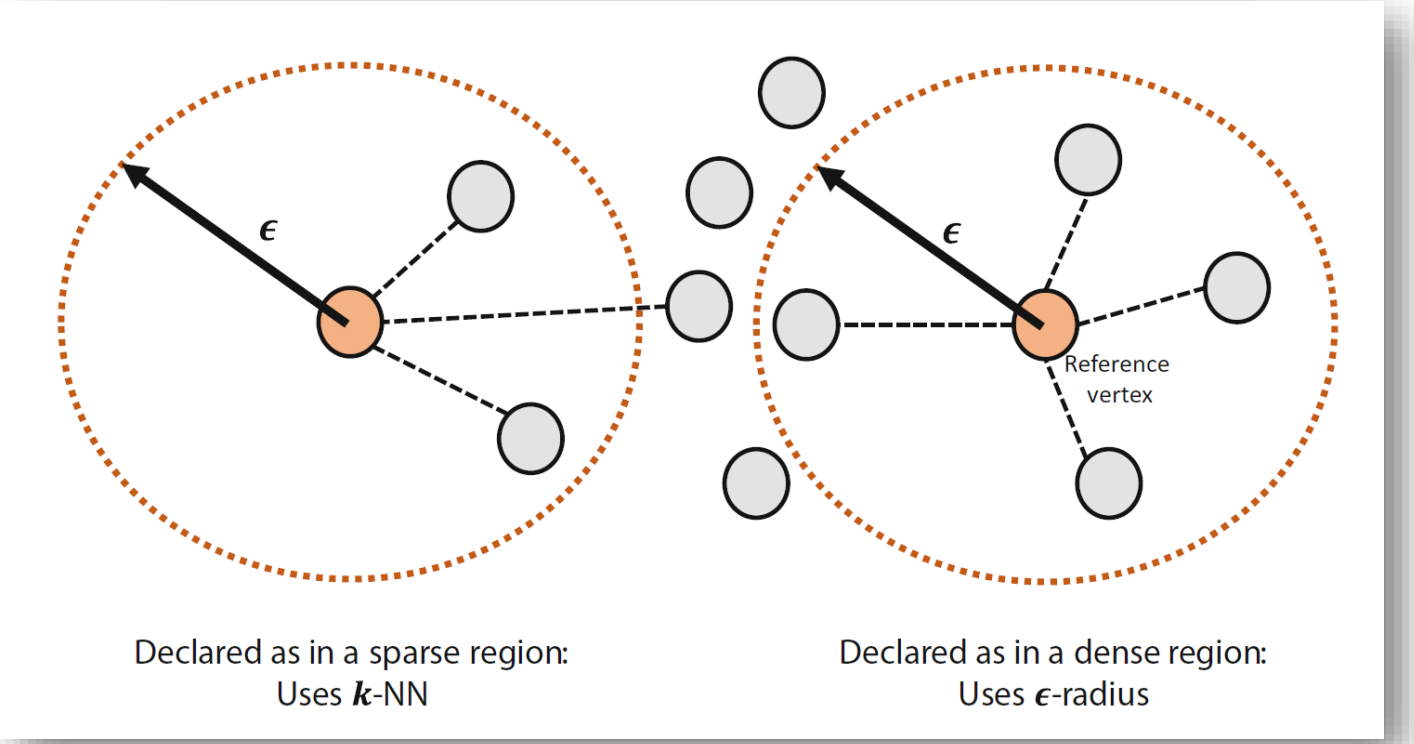
Problémy


a
b
2
3
5
7
4
6
1
Reference
vertex
4
6
5
2
3
1
Reference
vertex


a
b
ϵ
Reference
vertex
ϵ

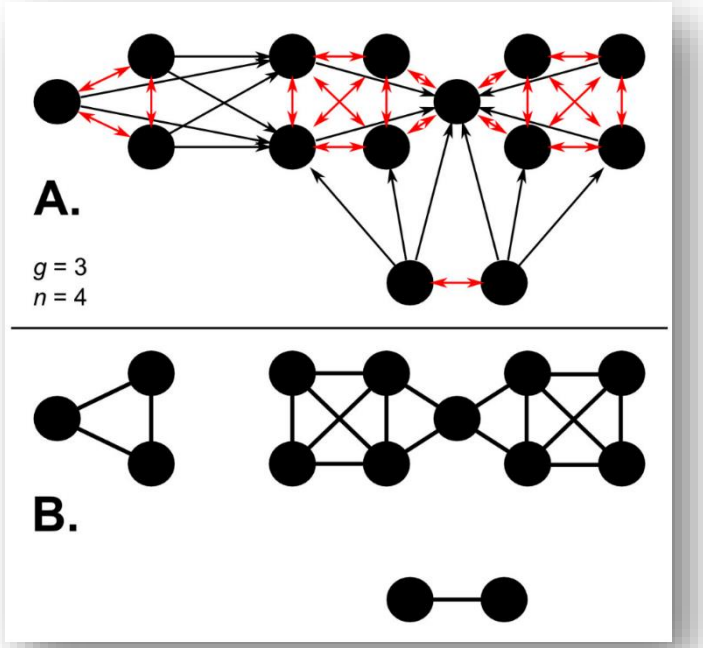
Kombinace algA a algB

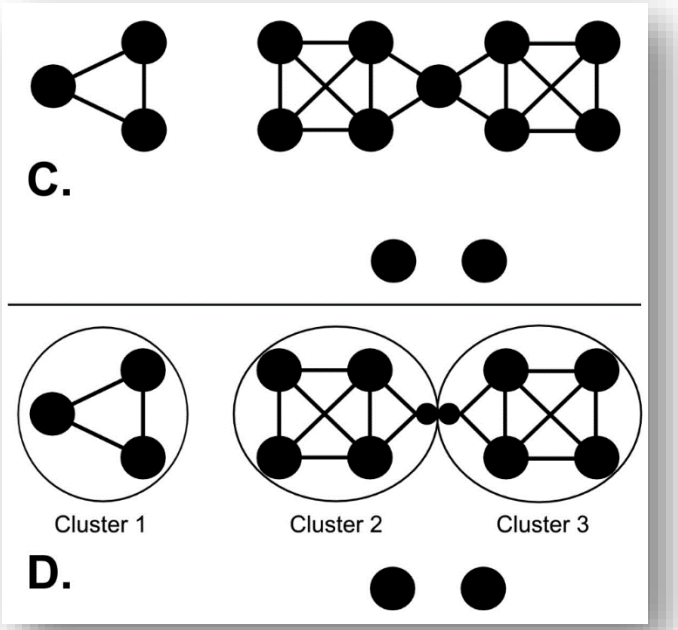
$$\mathcal{N}(v_i) = \begin{cases} \epsilon\text{-radius}(v_i), & \text{if } |\epsilon\text{-radius}(v_i)| > k \\ k\text{-NN}(v_i), & \text{otherwise} \end{cases}$$



Krok 3_algC: Nearest Neighbor Networks

- Zvolí se počet nejbližších sousedů *n*.
- Spojí se jen ty vrcholy, které jsou „oboustranně“ nejbližšími sousedy.





Huttenhower, C., Flamholz, A. I., Landis, J. N.,... (2007). Nearest Neighbor Networks: clustering expression data based on gene neighborhoods. *Bmc Bioinformatics*, *8*(1), 250.

Krok 3_algD: LRNet

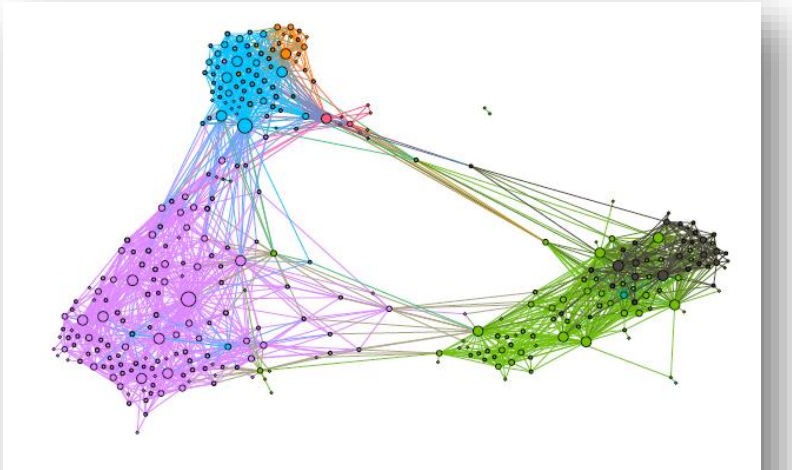
- Algoritmus využívá lokální reprezentativnost vrcholů k volbě počtu nejbližších sousedů.
- Lokální reprezentativnost je pro algoritmus kernel funkcí *f(N, NN)*, která definuje pro různé vrcholy různé *k*. Počet nejbližších sousedů vrcholu závisí na celkovém počtu jeho sousedů a počtu těch sousedů, pro které je tento vrchol nejbližším sousedem.
- Dále algoritmus funguje podobně jako *k-NN*.

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

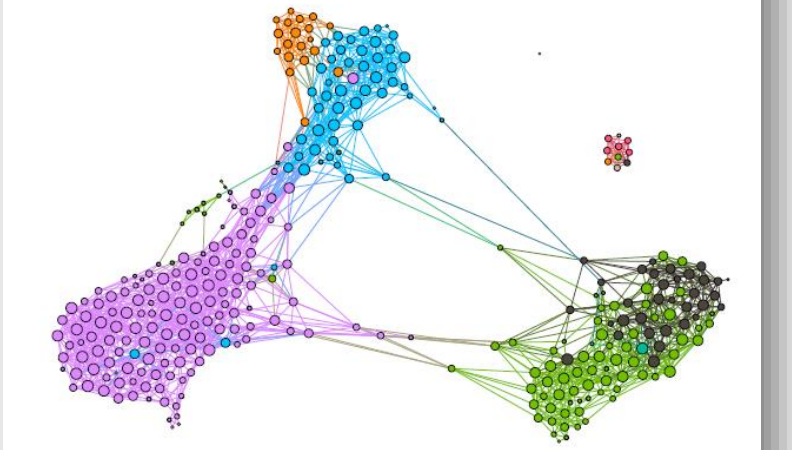
Referenční biologická data

Najdete rozdíl?

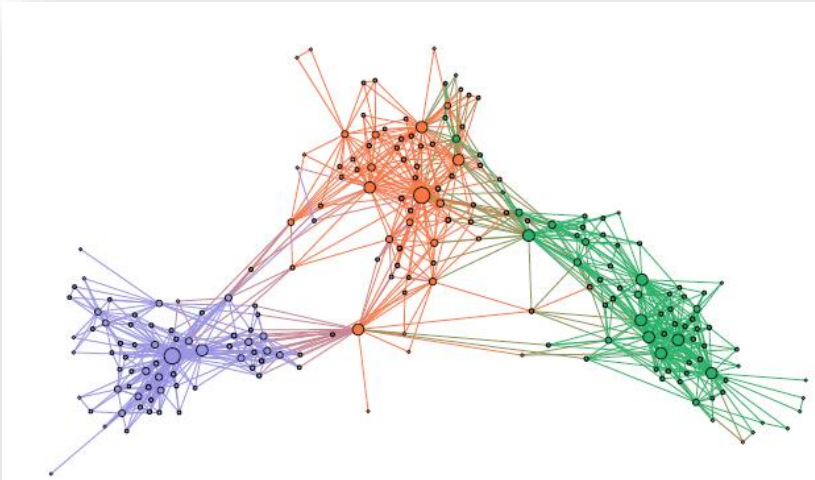
Velikost vrcholu odpovídá počtu sousedů…



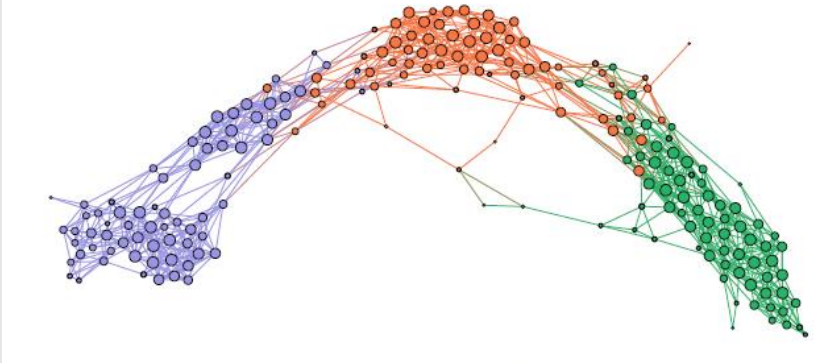
(a) Ecoli: LRNet



(b) Ecoli: SqrtNN

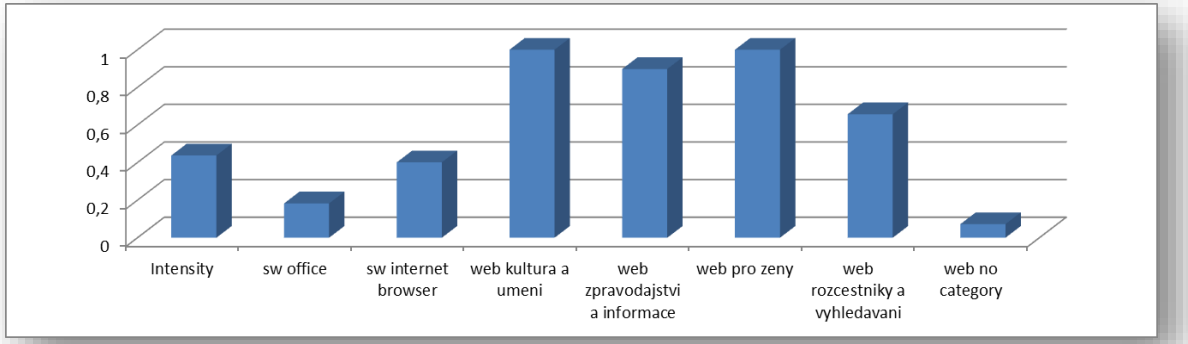


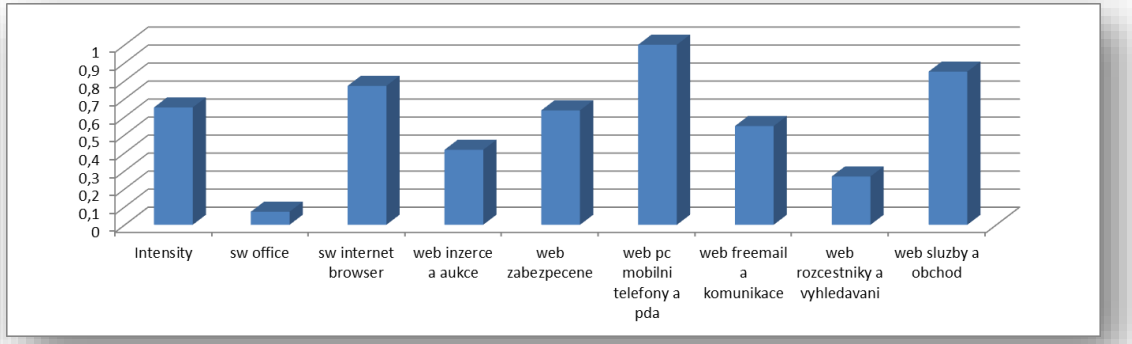
(a) Seeds: LRNet

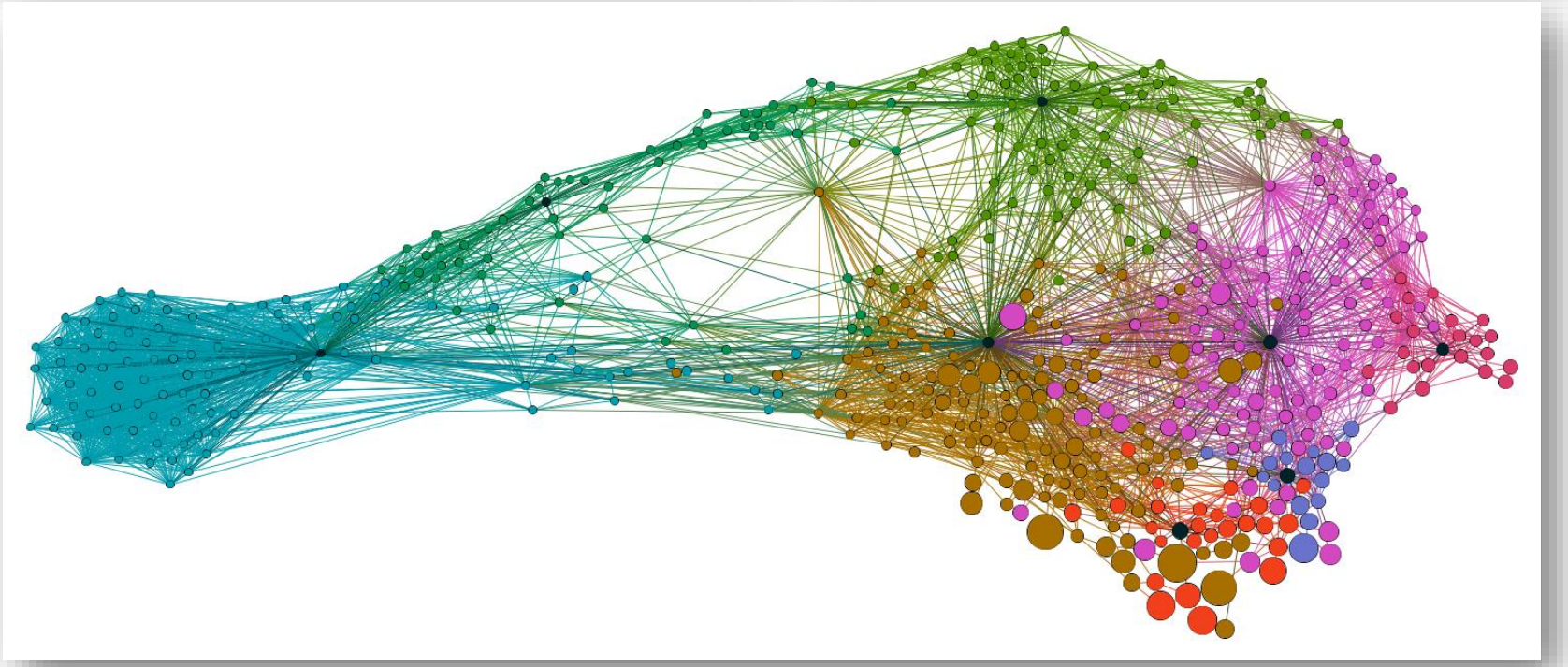


(b) Seeds: SqrtNN

Hodinové chování

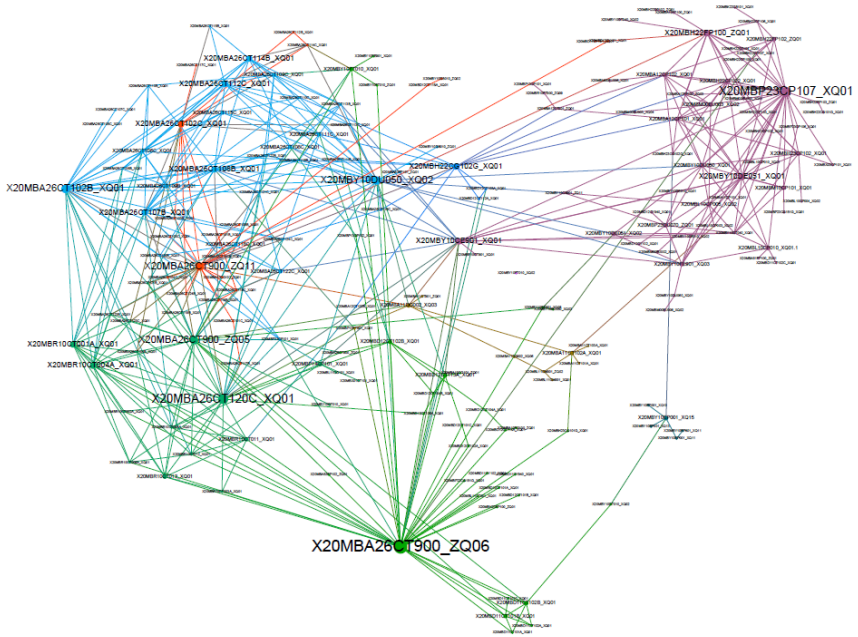


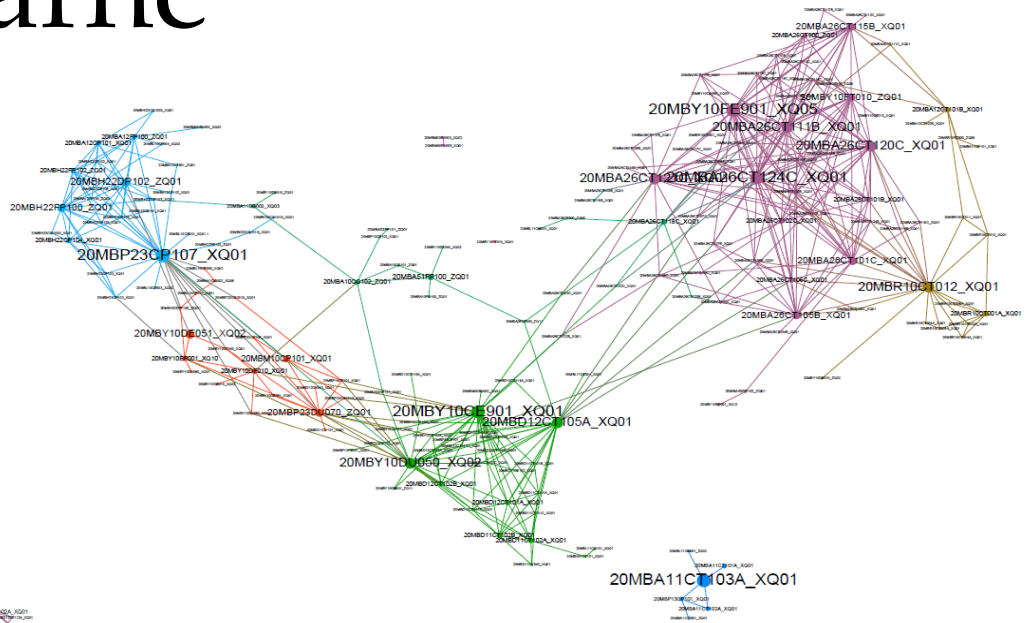


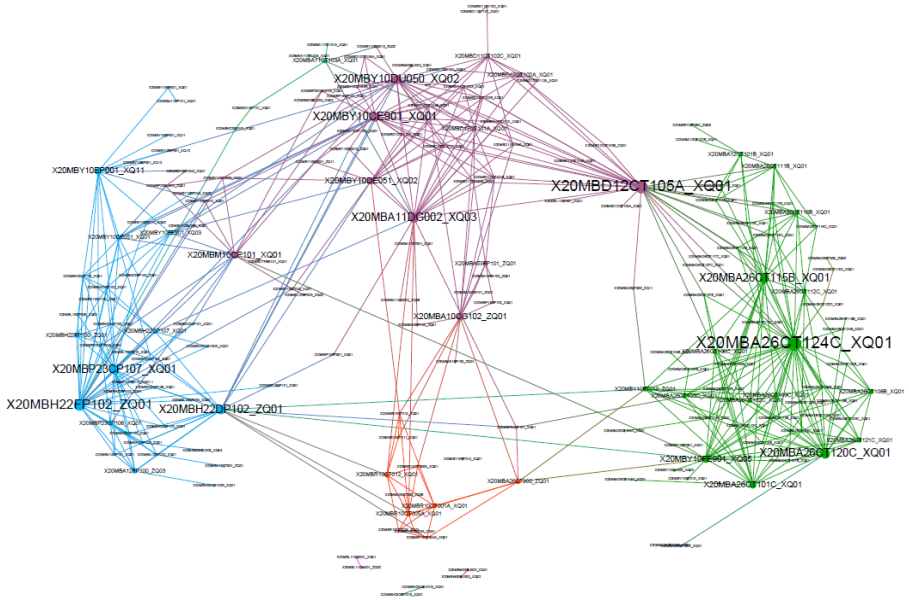


Data ze senzorů v elektrárně

- Sítě: Y-h M-m D-s

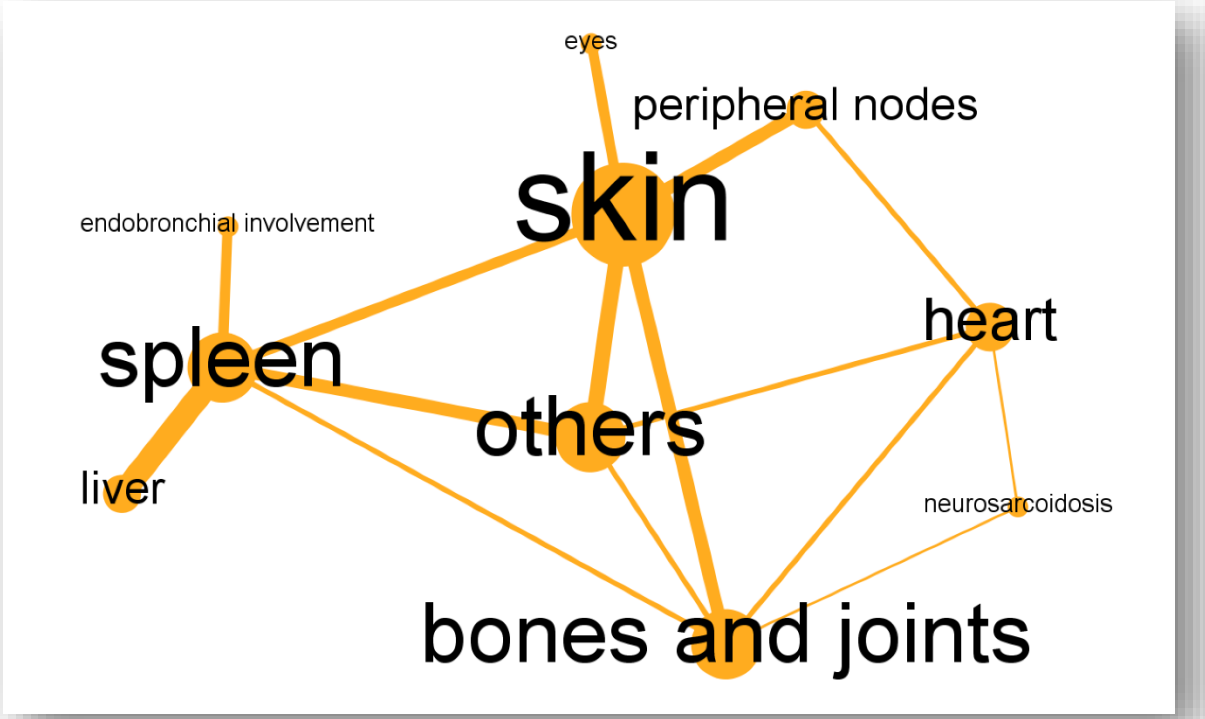




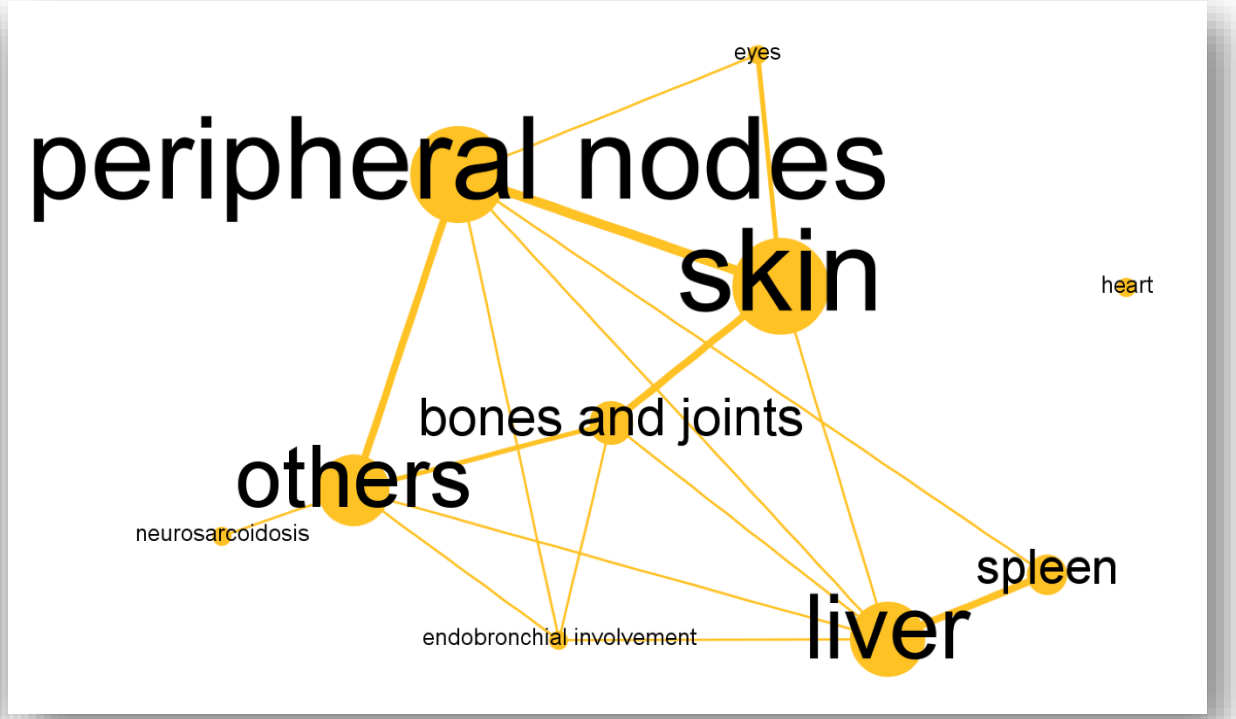


Společný výskyt postižených orgánů

Ženy



Muži



Iris

- https://archive.ics.uci.edu/ml/datasets/iris
- 150 instancí, 4 atributy

Seeds

- https://archive.ics.uci.edu/ml/datasets/seeds

- 210 instancí, 7 atributů

Ecoli

- https://archive.ics.uci.edu/ml/datasets/ecoli

- 336 instancí, 7 atributů

Shrnutí

- Máme různé metody, které se liší ve výsledcích.
- Kombinace převodu a shlukování je výborný nástroj pro pochopení, jak data vypadají.
- Časová složitost je pro řídké sítě $O(n^2)$, $O(n^3)$ pro husté sítě nebo velmi vysoký počet atributů.

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.

- LRNet algoritmus: https://homel.vsb.cz/~kud007/lrnet_files/