Présentation du projet de méthodologie de recherche
Option : AMSD

Non-linear Attributed Graph Clustering by Symmetric NMF with PU Learning

Réalisé par:

- Chaimae Hilal
- Kenza Ouazzani Chahdi
- Sana Oubenyahya
- Mouad Et-tali

2021/2022

OUTLINE

I. Datasets

II. Methods

❏ NMF

- **Disadvantage of traditional NMF in Attributed Graphs**

❏ Symmetric NMF

❏ PU learning

❏ NAGC

III. Findings

The project is based on the scientific paper **"Non-linear Attributed Graph Clustering by Symmetric NMF with PU Learning"** by Maekawa, Takeuchi and Onizuka.

The objective of the paper is to design an effective and efficient clustering method that accurately captures the hidden relationship between topology and attributes in real world graphs.

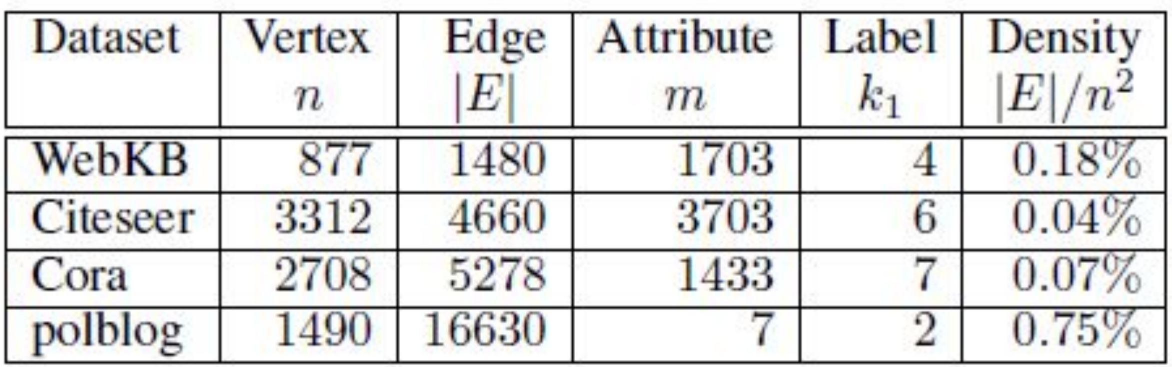
Datasets

❏ **WebKB :** the web graph of four universities, the label for a vertex indicates the owner university of the page. The attributes of a vertex represent the words appeared in the page.

❏ **Citeseer and Cora** are citation networks. The label for a vertex corresponds to a research field of the paper. The attributes of a vertex consist of the words appeared in the paper.

❏ **Polblog4** is a network of hyperlinks between blogs on US politics, the label of a vertex indicates whether the blog is liberal or conservative. The attributes of a vertex represent the sources of the blogs.

The following table summarizes the characteristics of each dataset.



Dataset	Vertex	Edge	Attribute	Label	Density
	n	$\|E\|$	m	k_1	$\|E\|/n^2$
WebKB	877	1480	1703	4	0.18%
Citeseer	3312	4660	3703	6	0.04%
Cora	2708	5278	1433	7	0.07%
polblog	1490	16630	7	2	0.75%

Methods

NMF : Non-negative Matrix Factorization

(NMF) provides a lower rank approximation of a nonnegative matrix, and has been successfully used as a clustering method

NMF : Non-negative Matrix Factorization

(NMF) provides a lower rank approximation of a nonnegative matrix, and has been successfully used as a clustering method

NMF's Loss function is defined as :

NMF : Non-negative Matrix Factorization

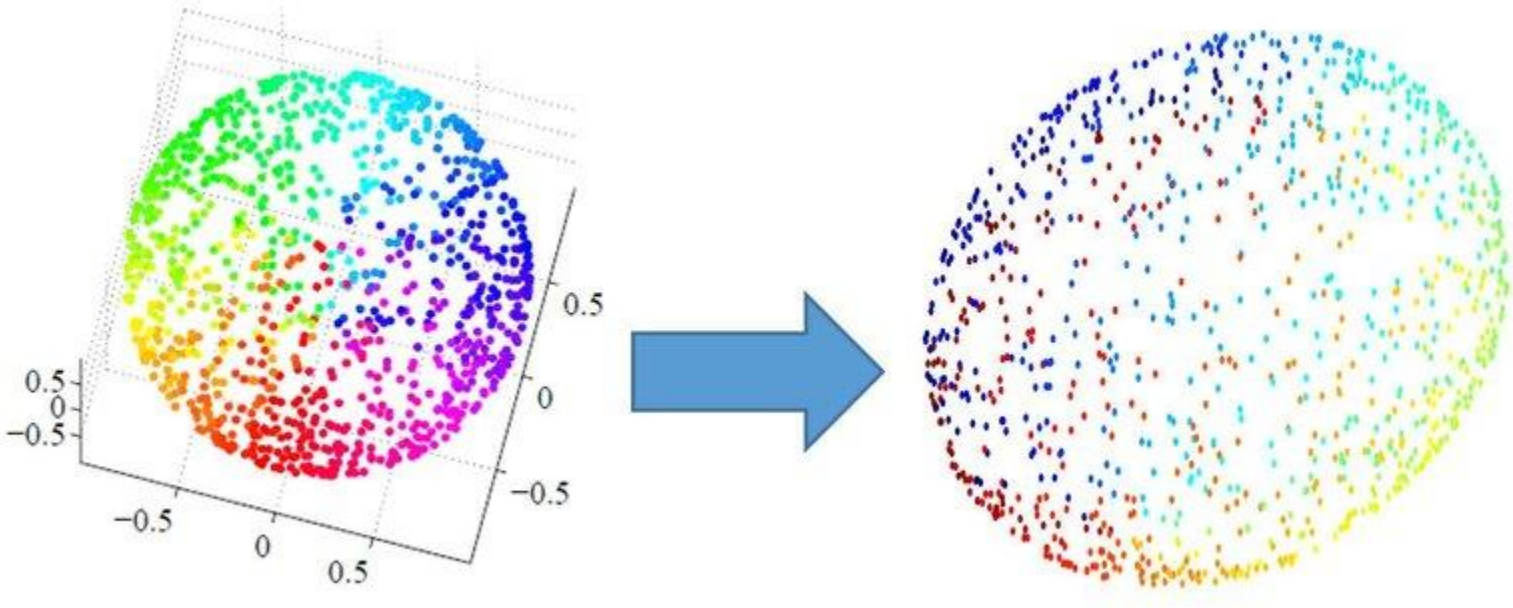
(NMF) provides a lower rank approximation of a nonnegative matrix, and has been successfully used as a clustering method.

NMF’s Loss function is defined as :

$$\min_{\boldsymbol{U},\boldsymbol{V}\geq 0} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\top}\|_{\mathcal{F}}^{2}$$

Disadvantage of traditional NMF in Attributed Graphs

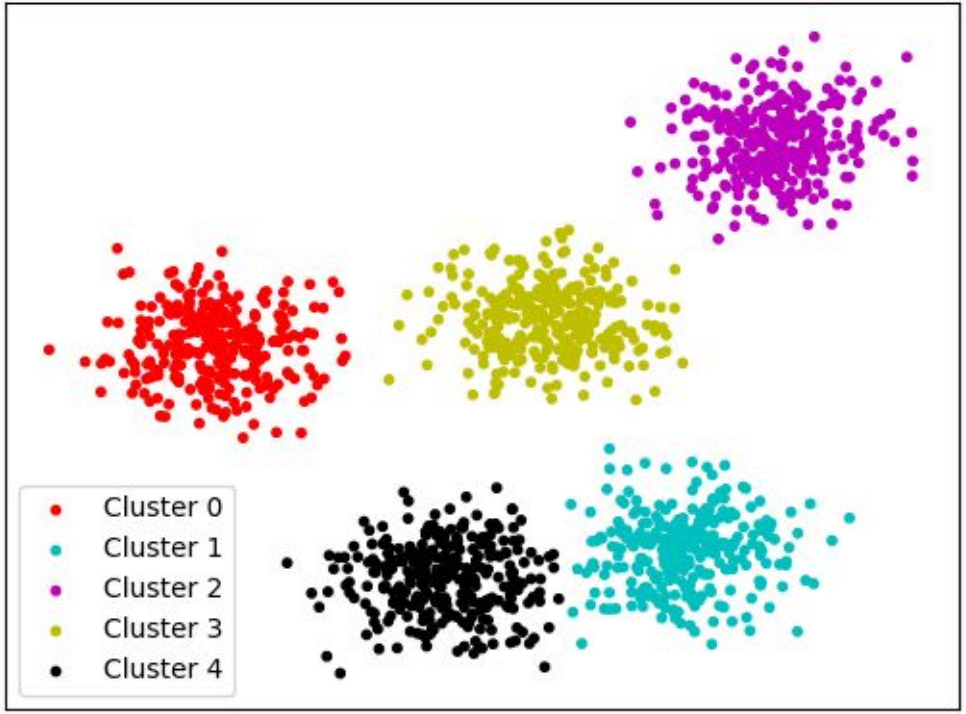
Although NMF has been shown to be effective to perform clustering, the goals of clustering and dimension reduction are different. (and only coincide under a certain condition)



3D

2D

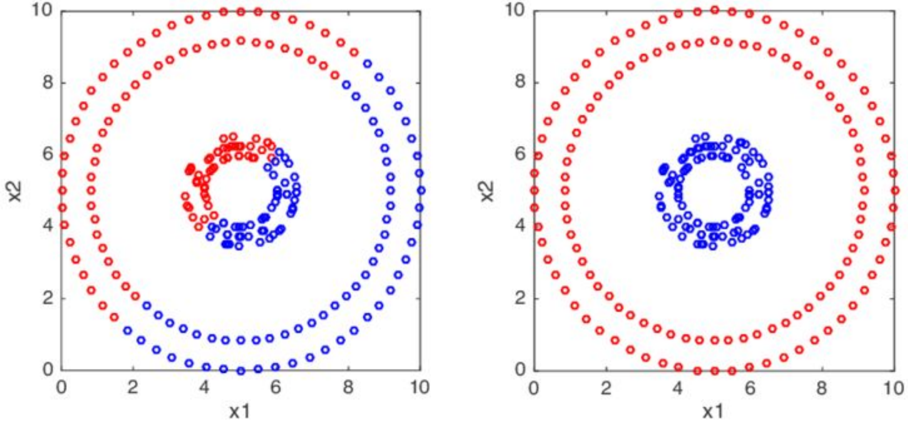
While a dimension reduction method uses a few basis vectors that well approximate the data matrix,



The goal of clustering is to find a partitioning of the data points where similarity is high within each cluster and low across clusters

When these two goals coincide, i.e. a basis vector is a suitable representation of one cluster, NMF is able to achieve good clustering results!...

However, this assumption is violated when the data have nonlinear cluster structures.ption is violated when the data have nonlinear cluster structures.



NMF (left) vs a Non Linear Clustering Method (right)

Symmetric NMF

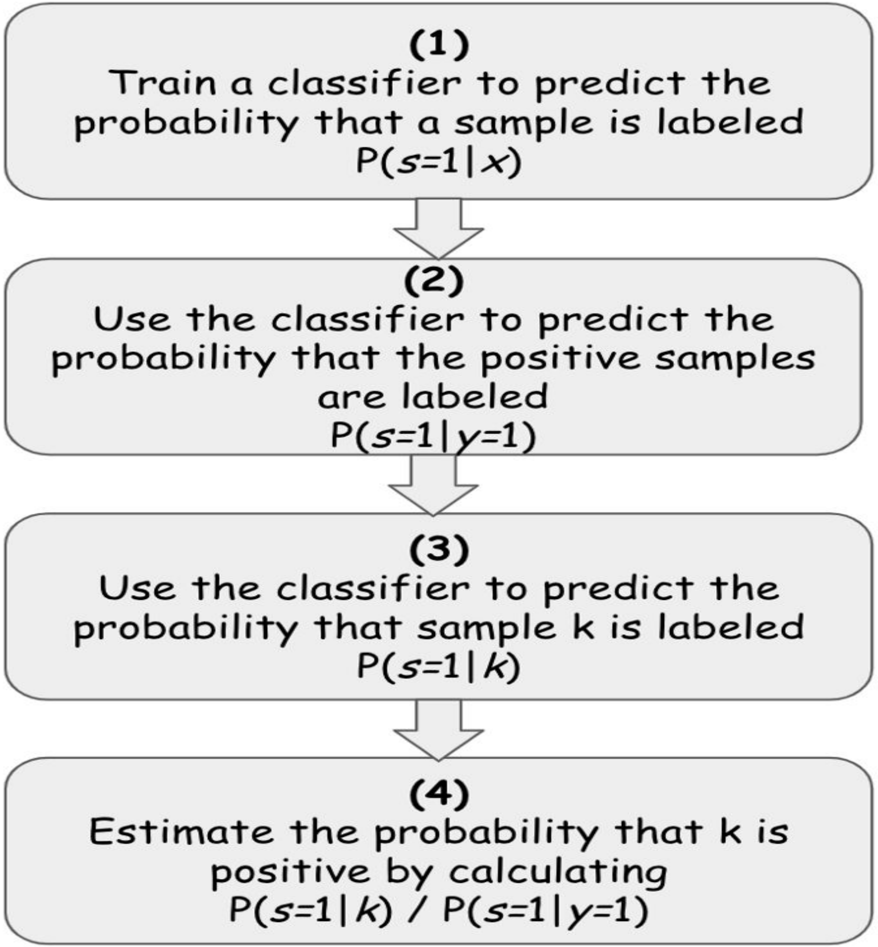
SymNMF is based on a similarity measure between data points, and factorizes a symmetric matrix containing pairwise similarity values (not necessarily nonnegative) which is why it estimates a cluster assignment matrix U by minimizing a non-convex loss function that uses S as input :

$$\min_{\boldsymbol{U} \geq 0} \|\boldsymbol{S} - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathcal{F}}^2$$

The main goal of graph clustering is to find a partition of vertices in a graph where the similarity between vertices is high within the same cluster and low across different clusters.

Positive Unlabeled learning

Given a set of examples of an particular class **P** (called the positive class) and a set of unlabeled examples **U**, which contains both class **P** and non-class **P** (called the negative class) instances, the goal is to build a binary classifier to classify the test set **T** into two classes, positive and negative, where **T** can be **U**.



(1)
Train a classifier to predict the probability that a sample is labeled
P(s=1|x)

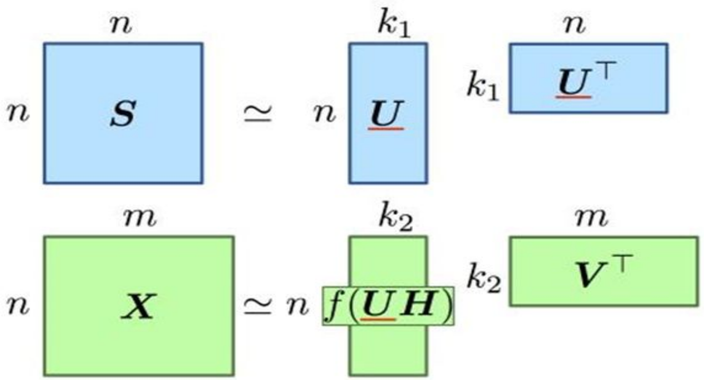
(2)
Use the classifier to predict the probability that the positive samples are labeled
P(s=1|y=1)

(3)
Use the classifier to predict the probability that sample k is labeled
P(s=1|k)

(4)
Estimate the probability that k is positive by calculating
P(s=1|k) / P(s=1|y=1)

Non Linear Attributed Graph

This method jointly decomposes the adjacency matrix S and the attribute matrix X into factor matrices with learning a non-linear projection function.

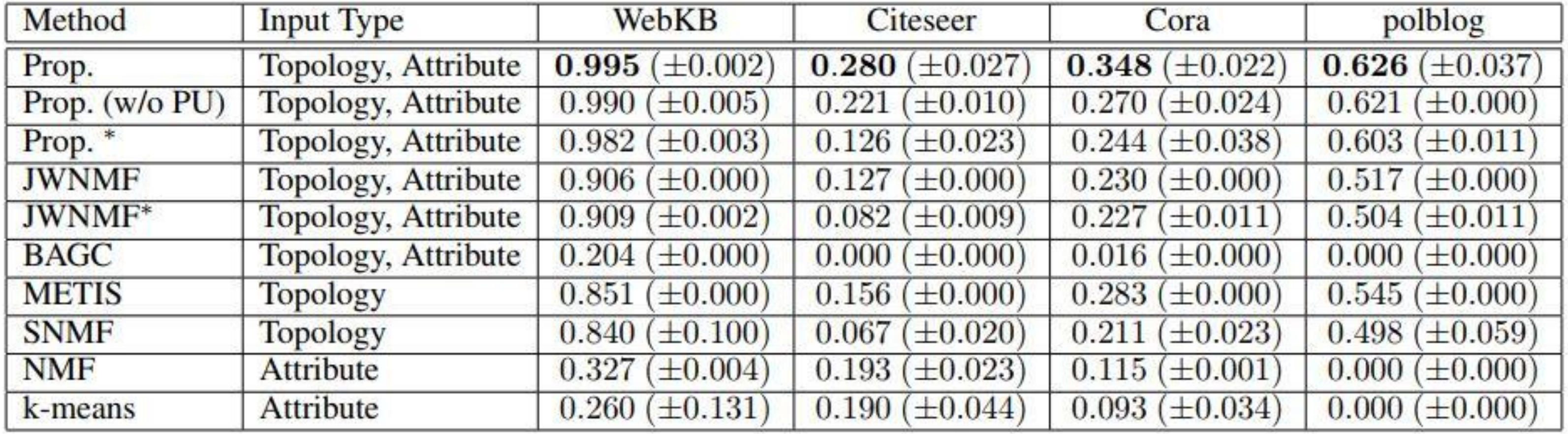


Non Linear Attributed Graph

It’s Loss function is defined as follows:

$$\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{H}\geq 0} \mathcal{L}_{\rho}(\boldsymbol{S}-\boldsymbol{U}\boldsymbol{U}^{\top})+\frac{\lambda}{2}\|\boldsymbol{X}-f(\boldsymbol{U}\boldsymbol{H})\boldsymbol{V}^{\top}\|_{\mathcal{F}}^{2}$$

Results



Method	Input Type	WebKB	Citeseer	Cora	polblog
Prop.	Topology, Attribute	**0.995** (±0.002)	**0.280** (±0.027)	**0.348** (±0.022)	**0.626** (±0.037)
Prop. (w/o PU)	Topology, Attribute	0.990 (±0.005)	0.221 (±0.010)	0.270 (±0.024)	0.621 (±0.000)
Prop. *	Topology, Attribute	0.982 (±0.003)	0.126 (±0.023)	0.244 (±0.038)	0.603 (±0.011)
JWNMF	Topology, Attribute	0.906 (±0.000)	0.127 (±0.000)	0.230 (±0.000)	0.517 (±0.000)
JWNMF*	Topology, Attribute	0.909 (±0.002)	0.082 (±0.009)	0.227 (±0.011)	0.504 (±0.011)
BAGC	Topology, Attribute	0.204 (±0.000)	0.000 (±0.000)	0.016 (±0.000)	0.000 (±0.000)
METIS	Topology	0.851 (±0.000)	0.156 (±0.000)	0.283 (±0.000)	0.545 (±0.000)
SNMF	Topology	0.840 (±0.100)	0.067 (±0.020)	0.211 (±0.023)	0.498 (±0.059)
NMF	Attribute	0.327 (±0.004)	0.193 (±0.023)	0.115 (±0.001)	0.000 (±0.000)
k-means	Attribute	0.260 (±0.131)	0.190 (±0.044)	0.093 (±0.034)	0.000 (±0.000)

Table: The average and standard deviation (in parenthesis) of ARI.

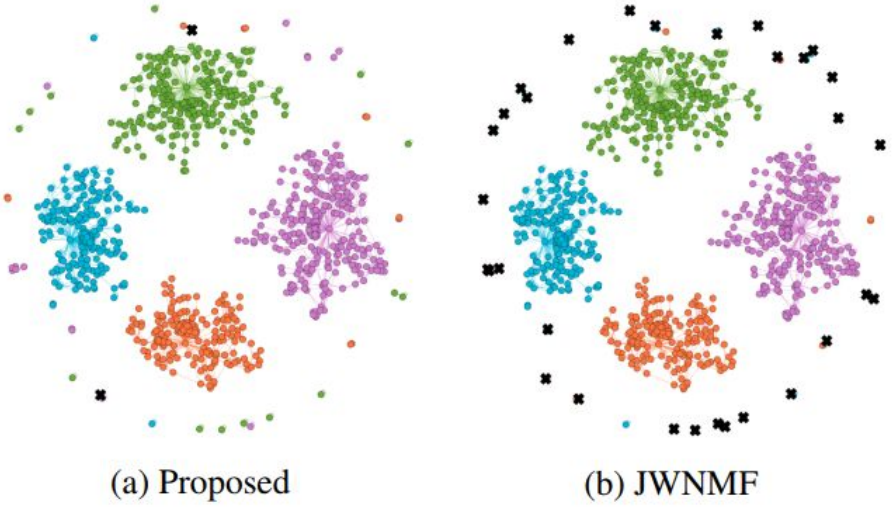
- Our method (Prop) initialized by K-means results outperforms all other methods for all datasets. This result confirms the power of nonlinear projection and PU learning to the quality of clustering.

- The results of our method without PU Learning (Prop. (w/o PU)) show the advantage of the nonlinear approach. In all the datasets, it is more efficient than the competing methods except for the Cora dataset on which the METIS method took first place.

- Moreover, the result of initialization by k-means always improves performance (Prop) compared to that initialized by random values (Prop*).

- The JWNMF method took second place for the WebKB dataset but resulted in poor performance on other datasets, while METIS, which is a graph clustering method, took second/third place for the WebKB, Cora and Polblog datasets.

- However, for the Citeseer, NMF and k-means dataset, attribute-based clustering methods took second and third place respectively. These results demonstrate that the topology or the attributes of the graphs remarkably affect the quality of the clustering, but it is more efficient to combine the two.


(a) Proposed

(b) JWNMF

Visualization of the results of WebKB with four clusters.

The colors of the vertices correspond to the clusters computed by our proposed method and JWNMF. The vertices with the cross mark ("x") indicate the wrong cluster assignment based on the ground truth. Our method assigns only two vertices to the wrong clusters out of 877 vertices.

References

- D. Kuang, C. Ding and H.Park. Symmetric Nonnegative Matrix Factorization for Graph Clustering, 1:3, 2012.

- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization, 1:2, 1999.

- https://medium.com/octavian-ai/how-to-get-started-with-machine-learning-on-graphs-7f0795c83763

- https://iksinc.online/2016/03/21/what-is-nmf-and-what-can-you-do-with-it /