# Réduction de dimension et classification non supervisée d'un graphe attribué : une approche simultanée

Lazhar Labiod and Mohamed Nadif

LIPADE, University of Paris, 45 rue des Saints-Pères,
Paris, France.

04/09/2019

# Plan

# Attributed Network

## Attributed Network

- AN are used to model a large variety of real-world networks, such as academic networks and health care systems where both node links and attributes/features are available for analysis.
- AN embedding and clustering has received a significant amount of attention as an important problem with many applications including social networks, academic citation networks and protein-protein interaction networks.
- Unlike plain networks in which only node links and dependencies are observed, each node in an AN is often associated with a valuable set of features.

# Attributed Network(AN)

## Definition

An attributed network $G = (V, E, X)$ consists of $V$ the set of nodes, $E \subseteq V \times V$ the set of links, and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}2, \ldots, \mathbf{x}_n]$ where $n = |V|$ and $\mathbf{x}_i \in \mathbb{R}^d$ is the feature/attribute vector of the node $v_i$.

## Attributed Network

Formally, the graph can be represented by two types of information ;

- the content information $\mathbf{X} \in \mathbb{R}^{n \times d}$.
- the structure information $\mathbf{A} \in \mathbb{R}^{n \times n}$, where $\mathbf{A}$ is an adjacency matrix of $G$ and $a_{ij} = 1$ if $e_{ij} \in E$ otherwise 0 ; we consider that each node is a neighbor of itself, then we set $a_{ii} = 1$ for all nodes.

# AN embedding and clustering

## Clustering

The aim of cluster analysis is the discovery of a finite number of homogeneous classes from data.

## Embedding

AN embedding aims to seek a continuous low-dimensional matrix representation for nodes in a network, such that original network topological structure and node attribute proximity can be preserved in the this new low-dimensional embedding.

## Embedding and clustering

- PCA + k-means.
- Spectral embedding + k-means.

# Spectral clustering : a tandem approach

## Spectral Clustering (SC) [NJW, 2001]

- **Step 1**. finding a spectral embedding by using an eigenvector/eigenvalue decomposition of a Laplacian matrix.
- **Step 2**. K-means clustering is performed on the first few eigenvectors.

## Remark

In Spectral Clustering, spectral embedding (Step 1) and clustering (Step 2) tasks are performed separately.

# Graph Embedding and clustering : a tandem approach

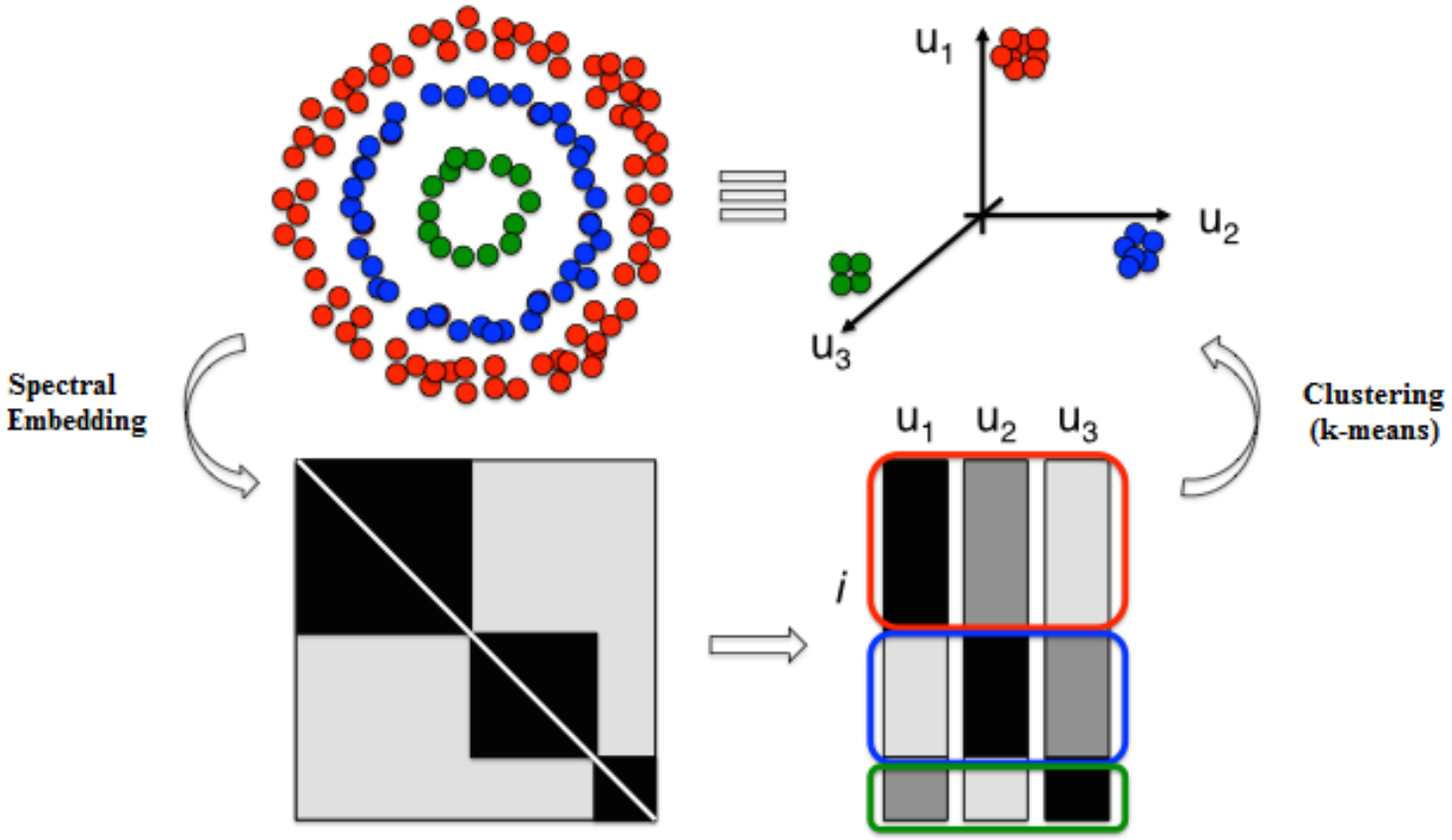



FIGURE – Spectral Clustering.

# AN Embedding and clustering : a tandem approach

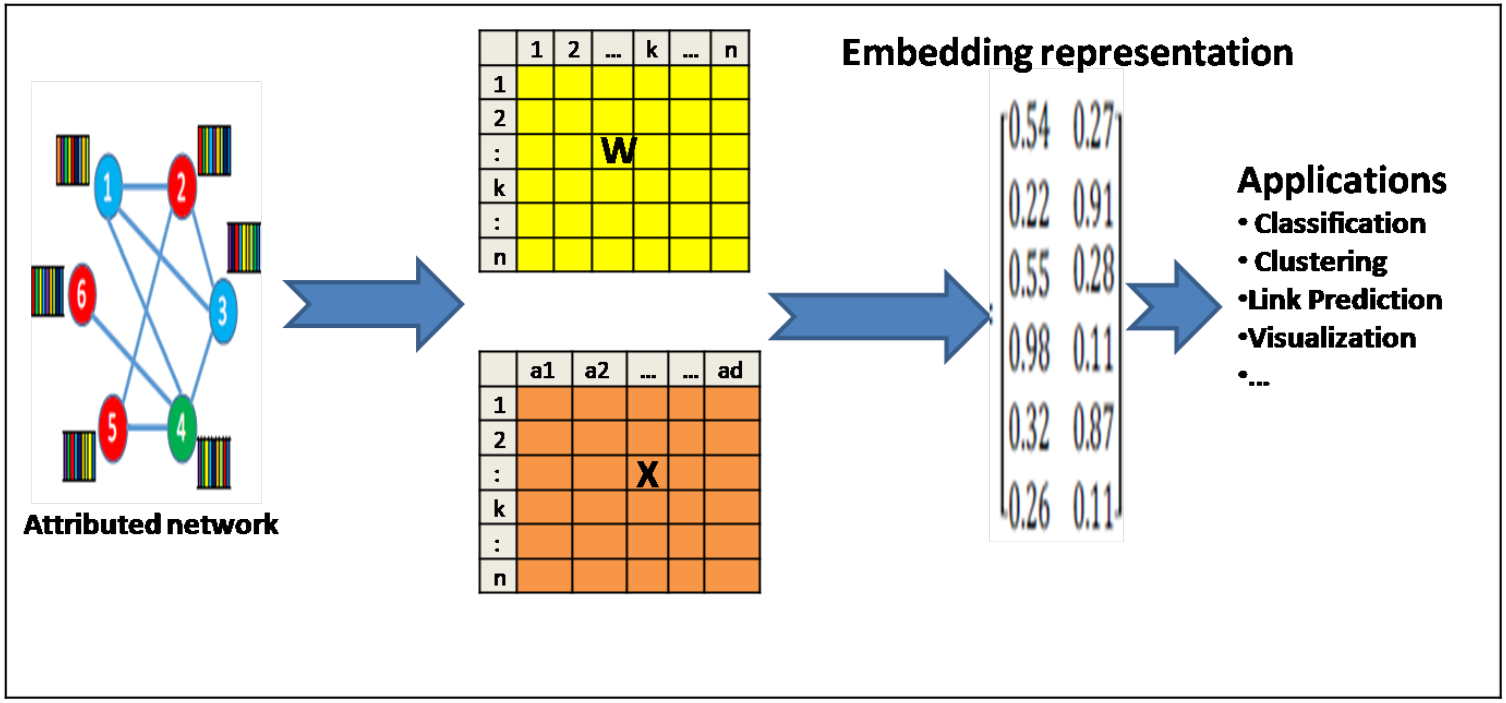



FIGURE – AN Embedding.

# limitations of tandem approach

Although existing AN clustering has been applied in practice widely, they may easily achieve poor performance due to the following drawbacks :

- The disadvantage of this approach is that consists in optimizing two different objectives.
- High risk of severe deviation of approximate continuous embedding solution from the good discrete clustering.
- Information loss among separate independent stages, i.e., continuous embedding generation, embedding discretization.

# Simultaneous attributed network embedding and clustering

## SANEC

We propose a novel simultaneous AN Embedding and clustering scheme which jointly ;

- learns embedding from both network and attributes information.
- learns continuous embedding and discrete clustering labels.
- Specifically, we explicitly enforce a discrete transformation on the intermediate continuous labels (embedding).
- This leads to a tractable optimization problem with a discrete solution.

# SANEC framework

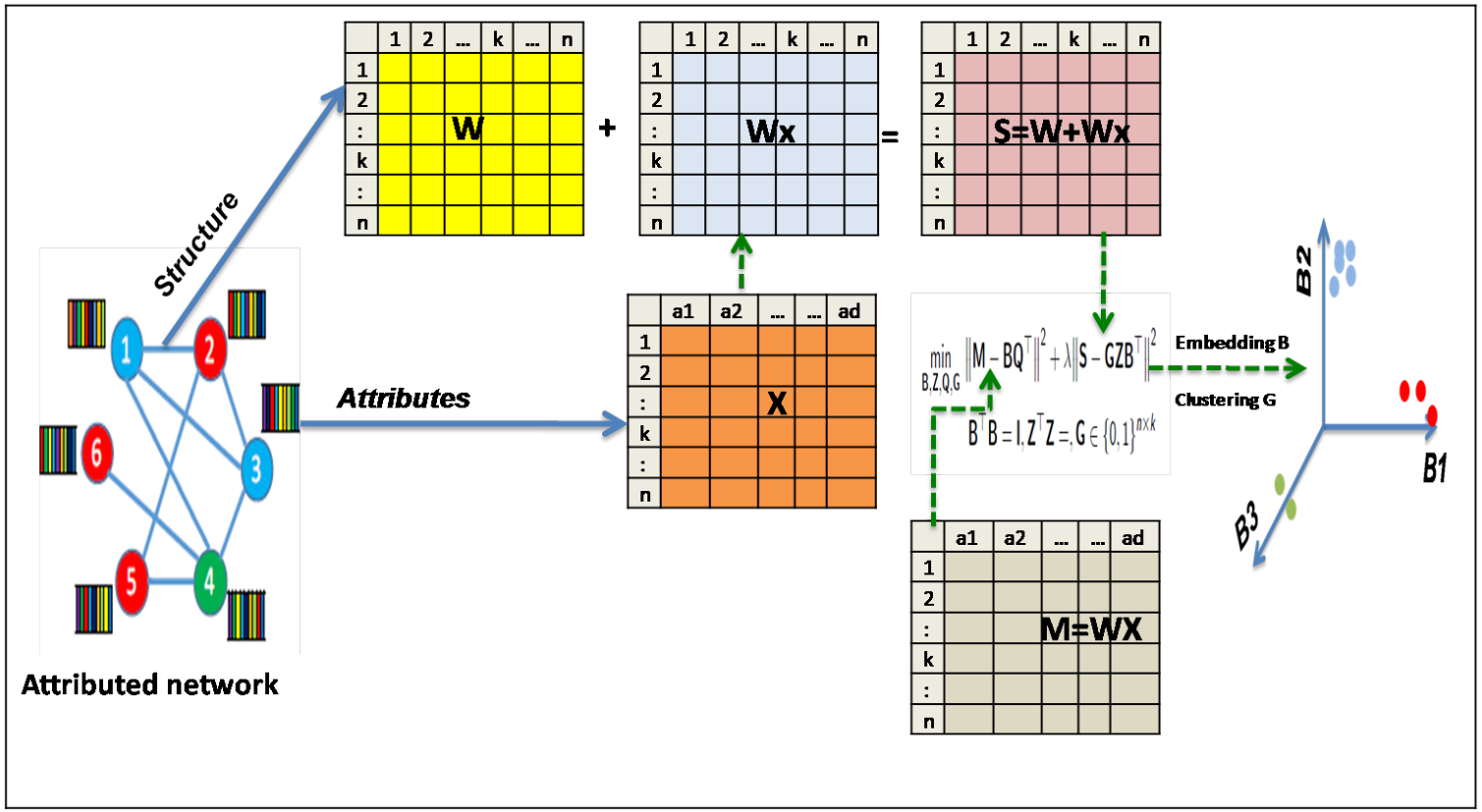



FIGURE – SANEC Framework

# SANEC framework

## W, M and S

- Nodes proximity : $\mathbf{W} = \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}$ is the degree matrix of $\mathbf{A}$ defined by $d_{ii} = \sum_{i'=1}^{n} a_{i'i}$.
- New data representation ; $\mathbf{M} = (m_{ij})$ can be considered as a multiplicative integration of both $\mathbf{W}$ and $\mathbf{X}$ by replacing each node by the centroid of their neighborhood (barycenter) : i.e,$\mathbf{m}_{ij} = \sum_{k=1}^{n} \mathbf{w}_{ik}\mathbf{x}_{kj}, \forall i, j$ or $\mathbf{M} = \mathbf{WX}$.
- Additional information about nodes similarity from $\mathbf{X}$ : we choose to take $\mathbf{S}$ defined by $\mathbf{S} = \mathbf{W} + \mathbf{W_X}$

# SANEC Objective Function

$$\min_{\mathbf{B},\mathbf{Z},\mathbf{Q},\mathbf{G}} \left\|\mathbf{M} - \mathbf{B}\mathbf{Q}^\top\right\|^2 + \lambda\left\|\mathbf{S} - \mathbf{G}\mathbf{Z}\mathbf{B}^\top\right\|^2 \quad (1)$$

$$\mathbf{B}^\top\mathbf{B} = \mathbf{I}, \mathbf{Z}^\top\mathbf{Z} = \mathbf{I}, \mathbf{G} \in \{0,1\}^{n\times k}$$

- $\mathbf{G} = (g_{ij})$ of size $(n \times k)$ is a cluster membership matrix,
- $\mathbf{B} = (B_{ij})$ of size $(n \times k)$ is the embedding matrix ;
- $\mathbf{Z} = (q_{ij})$ of size $(k \times k)$ is an orthonormal rotation matrix which most closely maps $B$ to $G$.

**Spectral Embedding criterion**

$$\min_{\mathbf{B},\mathbf{Q}} \left\|\mathbf{M} - \mathbf{B}\mathbf{Q}^\top\right\|^2.$$

**Spectral rotation (in low embedding subspace)**

$$\left\|\mathbf{S} - \mathbf{G}\mathbf{Z}\mathbf{B}^\top\right\|^2 = \left\|\mathbf{S} - \mathbf{B}\mathbf{B}^\top\mathbf{S}\right\|^2 + \|\mathbf{S}\mathbf{B} - \mathbf{G}\mathbf{Z}\|^2$$

# SANEC : Optimization

To solve (1) we rely on the following theorem.

**Theorem**

*Let* $\mathbf{X}_{n\times k}$ *and* $\mathbf{B}_{n\times k}$ *be two matrices. Consider the constrained minimization problem*

$$\mathbf{Z}_* = \arg\min_{\mathbf{Z}} \left\|\mathbf{X} - \mathbf{B}\mathbf{Z}^T\right\|^2 \quad s.t \quad \mathbf{Z}^T\mathbf{Z} = I.$$

*Suppose the SVD of* $\mathbf{X}^T\mathbf{B}$ *is* $\mathbf{U}\mathbf{D}\mathbf{V}^T$*, then* $\mathbf{Z}_* = \mathbf{U}\mathbf{V}^T$*.*

We can observe that this problem turns out to be similar to the well known orthogonal Procrustes problem (Schonemann, 1966)

# SANEC : Optimization

**Compute Z**. Fixing **G** and **B** the problem which arises in (1) is equivalent to

$$\max_{\mathbf{Z}} \; Tr(\mathbf{G}^\top \mathbf{S}\mathbf{B}\mathbf{Z}) \quad \text{s.t.} \quad \mathbf{Z}^\top \mathbf{Z} = \mathbf{I}. \tag{2}$$

**Compute Q.** Given **G**, **Z** and **B**, the opimization problem (1) is equivalent to

$$\min_{\mathbf{Q}} \left\| \mathbf{M} - \mathbf{B}\mathbf{Q}^\top \right\|^2. \tag{3}$$

**Compute B**. Given **G**, **Q** and **Z**, the problem (1) is equivalent to

$$\max_{\mathbf{B}} \; Tr((\mathbf{M}^\top \mathbf{Q} + \lambda \mathbf{S}\mathbf{G}\mathbf{Z})\mathbf{B}^\top) \quad \text{s.t.} \quad \mathbf{B}^\top \mathbf{B} = \mathbf{I}. \tag{4}$$

**Computation of G.** As **G** is a cluster membership matrix, its computation is done as follow : We fix **Q**, **Z**, **B**. Let $\tilde{\mathbf{B}} = \mathbf{S}\mathbf{B}$ and calculate

$$g_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_{k'} ||\tilde{\mathbf{b}}_i - \mathbf{z}_{k'}||^2 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In summary, the steps of the SDRC algorithm can be deduced in Algorithm 1.

**Algorithme 1 :** : SANEC algorithm

**Input :** **M** and **S** from structure matrix **W** and content matrix **X** ;
**Initialize :** **B**, **Q** and **Z** with arbitrary orthonormal matrix ;
**repeat**
(a) - Compute **G** using (5)
(b) - Compute **B** using (4)
(c) - Compute **Q** using (3)
(d) - Compute **Z** using (2)
**until** convergence
**Output :** **G** : clustering matrix, **Z** : rotation matrix, **B** : nodes embedding and **Q** : attributes embedding

TABLE – Description of Data sets

| Data sets | Characteristics | | | |
|---|---|---|---|---|
| | #Nodes | # Attributes | #Edges | #classes |
| Cora | 2708 | 1433 | 5294 | 7 |
| Citeseer | 3312 | 3703 | 4732 | 6 |
| Wiki | 2405 | 4973 | 17981 | 19 |

- Cora : A citation network with 2708 nodes and 5294 links between them ; the nodes correspond to publications described by binary vectors of 1433 dimensions and classified into 7.
- Citeseer : A citation network consisting of 3312 publications labeled into 6 sub-fields. Each publication is described by a binary vector of 3703 dimensions and there are 4732 links between them.
- Wiki : A network with 2405 documents and 17981 links. These documents have 4973-dimension vectors representing them and are divided into 19 classes.

# Compared algorithms

- Graph Encoder [16] learns graph embedding for spectral graph clustering.
- DNGR[2] trains a stacked denoising autoencoder for graph embedding.
- RTM [3] learns the topic distributions of each document from both text and citation.
- RMSC[18] employs a multi-view learning approach for graph clustering.
- TADW [20] applies matrix factorization for network representation learning.
- DeepWalk[10] is a network representation approach which encodes social relations into a continuous embedding space.
- Spectral Clustering[14] is a widely used approach for learning social embedding.
- GAE [8] is an autoencoderbased unsupervised framework for attributed network data embedding.
- VGAE [8] is a variational graph autoencoder approach for graph embedding with both node links and node attributes information.
- ARGAis the most recent adversarially regularized autoencoder algorithm which uses graph autoencoder to learn the embedding.
- ARVGA [9] algorithm, which uses a variational graph autoencoder to learn the embedding.

TABLE – Clustering performances (Acc % , NMI % and ARI % ) on Cora, Citeseer and Wiki datasets

| Methods | Datasets | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Cora | | | Citeseer | | | Wiki | | |
| | Acc | NMI | ARI | Acc | NMI | ARI | Acc | NMI | ARI |
| K-means | 49.22 | 32.10 | 22.96 | 54.01 | 30.54 | 27.86 | 41.72 | 44.02 | 15.07 |
| Spectral | 36.72 | 12.67 | 03.11 | 23.89 | 05.57 | 01.00 | 22.04 | 18.17 | 01.46 |
| GraphEncoder | 32.49 | 10.93 | 00.55 | 22.52 | 03.30 | 01.00 | 20.67 | 12.07 | 0.49 |
| DeepWalk | 48.40 | 32.70 | 24.27 | 33.65 | 08.78 | 09.22 | 38.46 | 32.38 | 17.03 |
| DNGR | 41.91 | 31.84 | 14.22 | 32.59 | 18.02 | 04.29 | 37.58 | 35.85 | 17.97 |
| RTM | 43.96 | 23.01 | 16.91 | 45.09 | 23.93 | 20.26 | 43.64 | 44.95 | 13.84 |
| RMSC | 40.66 | 25.51 | 08.95 | 29.50 | 13.87 | 04.88 | 39.76 | 41.50 | 11.16 |
| TAWD | 56.03 | 44.11 | 33.20 | 45.48 | 29.14 | 22.81 | 30.96 | 27.13 | 04.54 |
| VGAE | 50.20 | 32.92 | 25.47 | 46.70 | 26.05 | 20.56 | 45.09 | 46.76 | 26.34 |
| ARGE | 64.0 | 44.9 | 35.2 | 57.3 | 35.0 | 34.1 | 47.34 | 47.02 | 28.16 |
| ARVGE | 63.8 | 45.0 | 37.74 | 54.4 | 26.1 | 24.5 | 46.45 | 47.8 | 29.65 |
| SANEC | **67.38** | **47.14** | **39.88** | **66.77** | **40.60** | **41.78** | **49.57** | **48.30** | **33.14** |

# Cora visualization

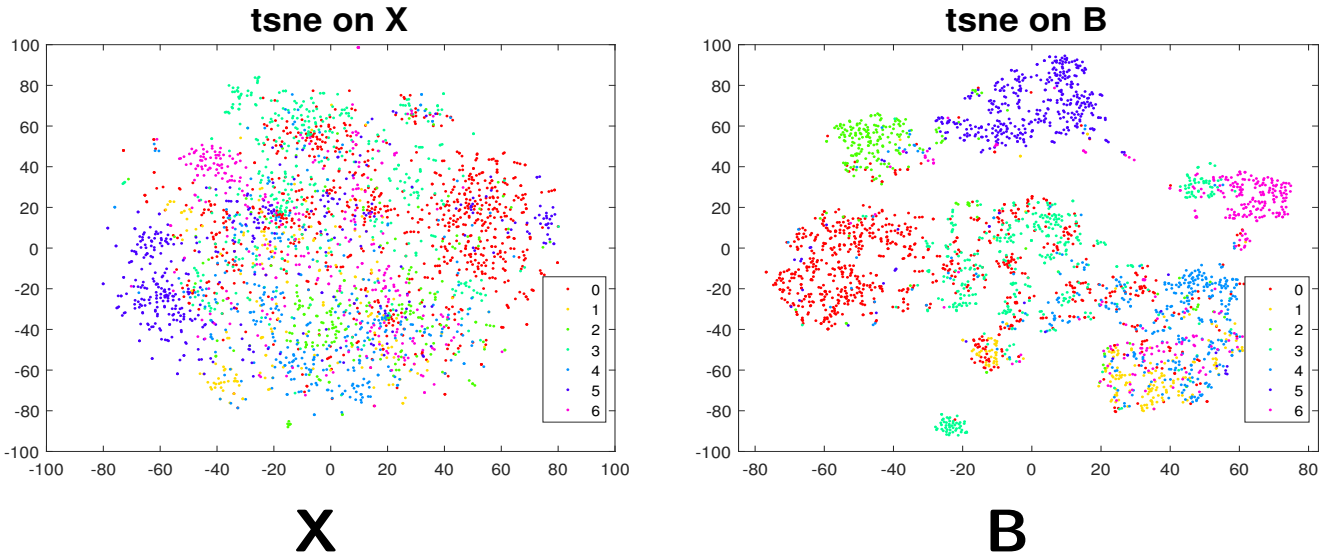



FIGURE – Cora visualization.

# Visualization of Citeseer dataset

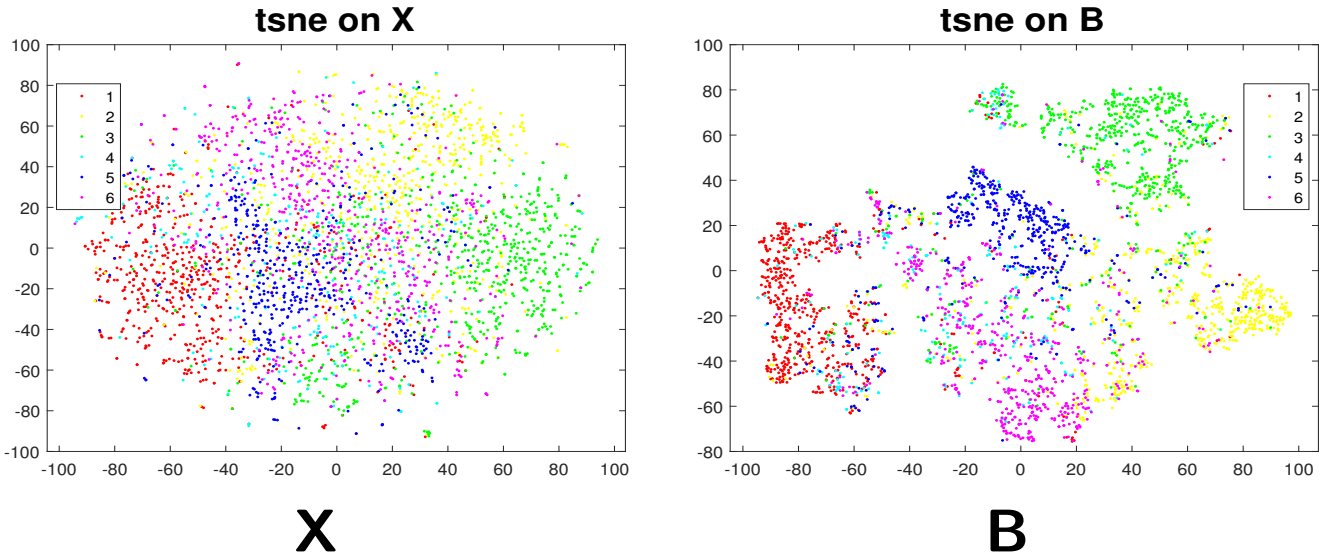
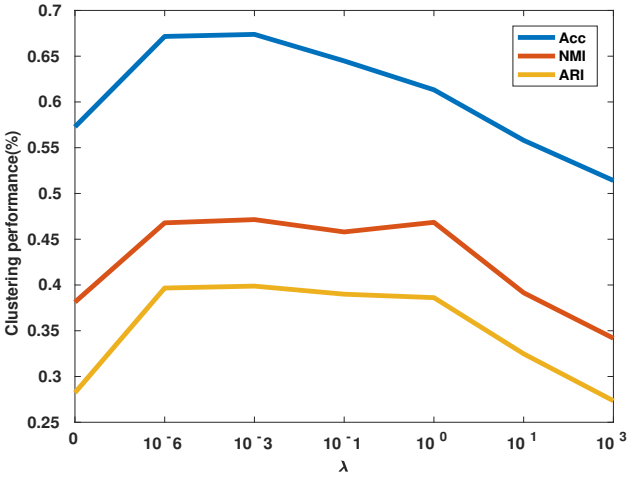
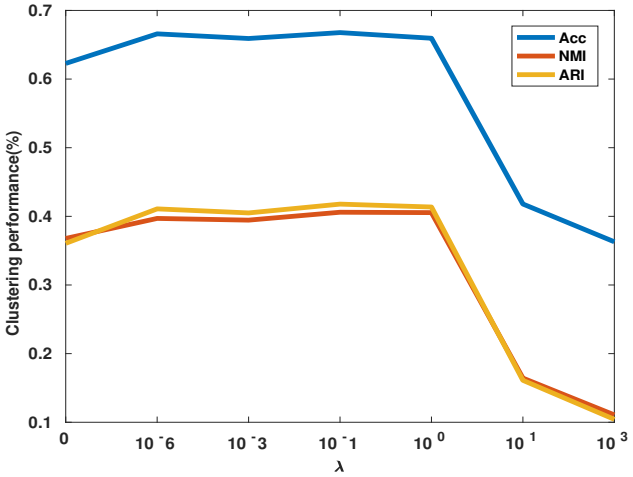



FIGURE – Visualization of Citseer dataset.

# Sensitivity analysis according to $\lambda$




Cora dataset




Citeseer dataset

FIGURE – Sensitivity analysis.

# Conclusion

## Contributions

- We proposed a novel matrix decomposition framework for simultaneous attributed network data embedding and clustering.
- We show that the objective learning of SANEC can be decomposed into three terms, the first one is the objective function of PCA applied to $M = WX$, the second is the graph embedding criterion in a low-dimensional space, and the third term is the clustering criterion.
- We also integrate a discrete rotation functionality, which allows a smooth transformation from the relaxed continuous embedding to a discrete solution and guarantees a tractable optimization problem with a discrete solution.

# Thank You