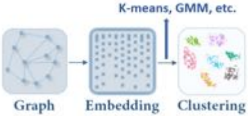
Clustering by Deep Latent Position Model with Graph Convolutional Network

Dingge Liang
(Joint work with C. Bouveyron, P. Latouche and M. Corneli)

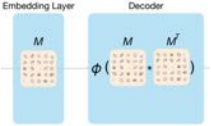
Université Côte d'Azur, Inria Maasai team

Most of the existing works for graph clustering

- are two-step embedding models $\Longrightarrow$ not **end-to-end** methods



- use an simple inner product as decoder $\Longrightarrow$ no **modelling of distances**



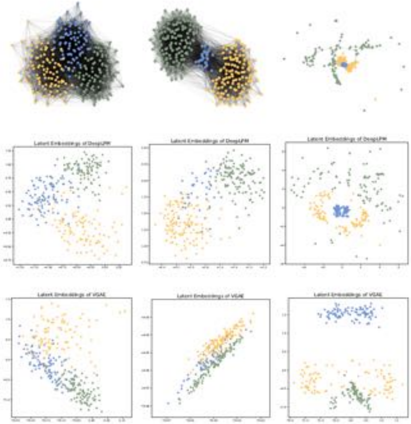


Figure: From top to bottom one sees: the original simulated graphs, the latent embeddings learned by DeepLPM and latent embeddings learned by VGAE.

Key-features of DeepLPM:

- a LPM-based decoder, modelling the distance between each pair of nodes in the latent space.
- it automatically assigns each node to its cluster, without any additional clustering step.

The generative process is as follows:

$$c_i \overset{iid}{\sim} \mathcal{M}(1, \pi), \text{ with } \pi \in [0,1]^K, \sum_{k=1}^{K} \pi_k = 1, \quad (1)$$

$$z_i | (c_{ik} = 1) \sim \mathcal{N}(\mu_k, \sigma_k^2 I_P), \text{ with } \sigma_k^2 \in \mathbb{R}^{+*}, \quad (2)$$

$$A_{ij} | z_i, z_j \sim \mathcal{B}(f_{\alpha,\beta}(z_i, z_j)), \quad (3)$$

with

$$f_{\alpha,\beta}(z_i, z_j) = \sigma(\alpha + \beta^T y_{ij} - ||z_i - z_j||^2). \quad (4)$$

We rely on a variational approach to approximate the log-likelihood with $\Theta = \{\pi, \mu_k, \sigma_k^2, \alpha, \beta\}$

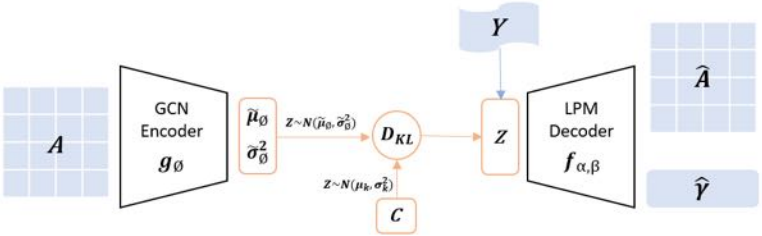
$$\log p(A|\Theta) = \mathcal{L}(q(Z, C); \Theta) + D_{KL}(q(Z, C)||p(Z, C|A, \Theta)), \quad (5)$$

and made following assumptions:

$$q(Z, C) = q(Z)q(C) = \prod_{i=1}^{N} q(z_i)q(c_i), \quad (6)$$

$$q(z_i) = \mathcal{N}(\tilde{\mu}_\phi(\overline{A})_i, \tilde{\sigma}_\phi^2(\overline{A})_i I_P), \quad (7)$$

$$q(C) = \prod_{i=1}^{N} \mathcal{M}(c_i; 1, \gamma_i). \quad (8)$$



Scenario A is simulated based on LPCM model.

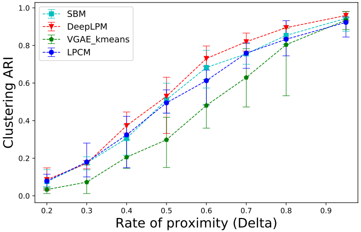


Figure: Clustering ARI in scenario A.

Scenario B is simulated according to SBM model.

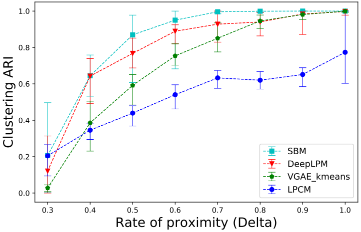


Figure: Clustering ARI in scenario B.

Scenario C is simulated based on circular-structured data.

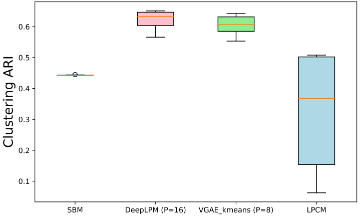


Figure: Clustering ARI in scenario C.

The real-world data comes from Medieval history of Europe.

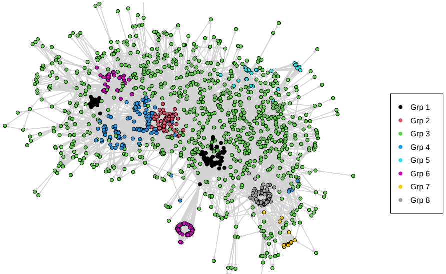


Figure: Cluster partition on the ecclesiastical network.