The Deep Latent Position Topic Model for Clustering and Representation of Networks with Textual Edges

Rémi Boutin [1] Pierre Latouche [2, 1] Charles Bouveyron [3] Dingge Liang [3]

[1]Université Paris Cité, CNRS, Laboratoire MAP5, UMR 8245, Paris, France
[2]Université Clermont Auvergne, CNRS, Laboratoire LMBP, UMR 6620, Aubière, France
[3]Université Côte d'Azur, INRIA, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France

Motivations

In this work, we are interested in data represented by a directed network $\mathcal{N} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ denotes the set of vertices $\{1, \dots, N\}$ and $\mathcal{E}$ corresponds to the set of edges. Let us denote $\mathbf{A} \in \mathcal{M}_{\mathrm{N}\times\mathrm{N}}(\{0,1\})$, the binary adjacency matrix, such that the coefficient A_{ij} is equal to 1 if $(i,j) \in \mathcal{E}$ and 0 otherwise. Moreover, $A_{ij} = 1$ means that node i sent a document to node j, denoted $W_{ij} = (W_{ijv})_{v=1}^{\mathrm{V}} \in \mathbb{N}^{\mathrm{V}}$, where V is the vocabulary size and W_{ijv}, the number of occurrence of word v. We aim at simultaneously **clustering nodes**, **finding meaningful topics** and **representing those results without post processing**.

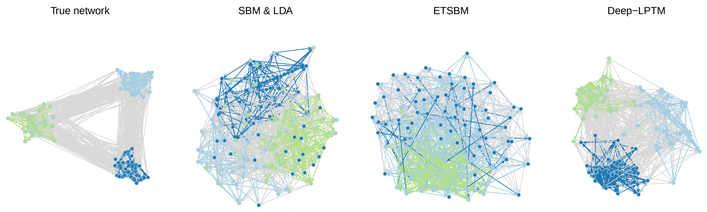

Figure 1. Representation of the node clustering (denoted by the node colours) as well as the topic modelling (the majority topic is illustrated by the edge colours) provided by SBM & LDA, ETSBM and Deep-LPTM respectively.

Model

Q	number of node clusters
K	number of topics in the documents
V	size of the vocabulary
P	dimension of the node embeddings
L	dimension of the word and topic embeddings
$\rho \in \mathcal{M}_{\mathrm{L}\times\mathrm{V}}(\mathbb{R})$	word embeddings
$\alpha_k \in \mathbb{R}^{\mathrm{L}}$	embedding of the k-th topic
$C = (C_i)_i$	the set of one hot encoded node cluster memberships
$Z = (Z_i)_i$	P-dimensional node embeddings
$Y = (Y_{ij})_{ij}$	K-dimensional edge embeddings
$\theta = (\theta_{ij})_{ij}$	$\theta_{ij} = \mathrm{softmax}(Y_{ij})$ the topic proportions in W_{ij}
$\mathbf{W} = (\mathbf{W_{ij}})_{ij}$	the set of the bag of words representations of the documents
$\mathbf{A}$	the binary adjacency matrix

Table 1. The **observed** and latent variables of Deep-LPTM as well the model quantities.

Generative assumptions

C_i	$\sim \mathcal{M}_{\mathrm{Q}}(1;\pi)$	π the topic proportions
$Z_i \mid C_{iq} = 1$	$\sim \mathcal{N}_{\mathrm{P}}(\mu_q, \sigma_q I_{\mathrm{P}})$	μ_q, σ_q the node embedding parameters
$\mathbf{A_{ij}} \mid Z_i, Z_j$	$\sim \mathcal{B}(\eta_{ij})$	κ, the threshold parameter
$Y_{ij} \mid C_{iq}C_{jr} = 1$	$\sim \mathcal{N}_{\mathrm{K}}(m_{qr}, \mathrm{diag}(s_{qr}))$	m_{qr}, s_{qr} document position parameters
$\mathbf{W_{ij}} \mid \theta_{ij}$	$\sim \mathcal{M}_{\mathrm{V}}(N_{ij}; \beta^\top \theta_{ij})$	$\beta = (\beta_1 \cdots \beta_K)^\top$ the topics distributions

where $\eta_{ij} = \mathrm{logit}(\kappa - \|Z_i - Z_j\|)$, $\beta_k = \mathrm{softmax}(\rho^\top \alpha_k) \in \Delta_{\mathrm{V}}$ the distribution of the vocabulary according to the k-th topic.

Factorisation of the complete-likelihood

Set of parameters $\Theta := \{\pi, \mu, \sigma, \kappa, m, s, \alpha, \rho\}$, and model parameters $m = (m_{qr})_{qr}$, $s = (s_{qr})_{qr}$, $\mu = (\mu_q)_q$ and $\sigma = (\sigma_q)_q$. The complete likelihood factorizes as:

$$p(\mathbf{A}, \mathbf{W}, C, Z, Y \mid \Theta) = p(\mathbf{A} \mid Z, \Theta) p(Z \mid C, \Theta) p(\mathbf{W} \mid Y, \Theta) p(Y \mid C, \Theta) p(C \mid \Theta).$$

Link with the embedded topic model [2] and the latent position and cluster model (LPCM) [3].

Graphical Model

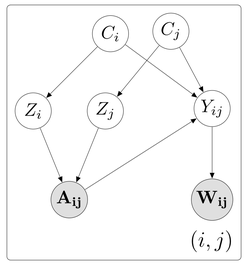

Figure 2. Graphical representation of the model without the parameters for the sake of clarity.

Inference

Likelihood

The marginal log-likelihood of the network and texts is given by:

$$\mathcal{L}(\Theta; \mathbf{A}, \mathbf{W}) = \log p(\mathbf{A}, \mathbf{W} \mid \Theta) = \log \left(\sum_C \int_Z \int_Y p(\mathbf{A}, \mathbf{W}, C, Z, Y \mid \Theta) dZ dY \right).$$

This quantity is not tractable. We choose to rely on a variational inference approach for approximation purposes. For any distribution $R(C, Z, Y)$, the following decomposition holds:

$$\mathcal{L}(\Theta; \mathbf{A}, \mathbf{W}) = \mathscr{L}(R(\cdot); \Theta) + \mathrm{KL}(R(\cdot) || p(C, Z, Y \mid \mathbf{A}, \mathbf{W})),$$

where

$$\mathscr{L}(R(\cdot); \Theta) = \mathbb{E}_R \left[\log \frac{p(\mathbf{A}, \mathbf{W}, C, Z, Y \mid \Theta)}{R(C, Z, Y)} \right]. \quad (1)$$

Variational distributions

For the quantity (1) to be tractable, we make the following assumptions about the variational distribution $R(\cdot)$:

$$\begin{aligned}
R(C, Z, Y) &= R(C) R(Z) R(Y), \\
R_\tau(C) &= \prod_{i=1}^{N} \mathcal{M}_Q(C_i; 1, \tau_i), \\
R_{\phi_Z}(Z \mid \mathbf{A}) &= \prod_{i=1}^{N} \mathcal{N}(Z_i; \mu_{\phi_Z}(\overline{\mathbf{A}})_i, \sigma^2_{\phi_Z}(\overline{\mathbf{A}})_i I_p), \\
R_{\phi_Y}(Y \mid \mathbf{A}, \mathbf{W}) &= \prod_{i \neq j} \mathcal{N}\big(Y_{ij}; \mu_{\phi_Y}(W_{ij}), \mathrm{diag}\left(\sigma^2_{\phi_Y}(W_{ij})\right)\big)^{\mathbf{A_{ij}}},
\end{aligned}$$

where $\tau = (\tau_i)_{i=1}^{\mathrm{N}}$ with $\forall i \in \{1, \dots, \mathrm{N}\}$, $\tau_i \in \Delta_{\mathrm{Q}}$. Also, $\mu_{\phi_Z}, \sigma^2_{\phi_Z}$ ($\mu_{\phi_Y}, \sigma^2_{\phi_Y}$ respectively) are the neural networks encoding the posterior mean and variances of the node embeddings (document embeddings respectively) which is based on [4].

IC2L: a new model selection criterion

We propose a new model selection criterion, IC2L, extending the integrated complete likelihood (ICL) [1] to our model. For a model $\mathcal{M}$ with Q clusters, K topics and a P-dimensional node latent space, assuming that the prior fully factorises, the IC2L criterion is given by:

$$\begin{aligned}
IC2L(\mathcal{M}, Q, K, P, \hat{Z}, \hat{Y}, \hat{C}) = &\max_\theta \log p(A, W, \hat{Z}, \hat{Y}, \hat{C} \mid \theta, \mathcal{M}, Q, K, P) \\
&- \Omega(\mathcal{M}, Q, K, P), \quad (2)
\end{aligned}$$

with $\hat{Z}, \hat{Y}$ and, $\hat{C}$ the maximum-a-posteriori estimates, and Ω a penalty coming from the BIC-like approximation.

Evaluation of the methodology

Scenarios and synthetic datasets

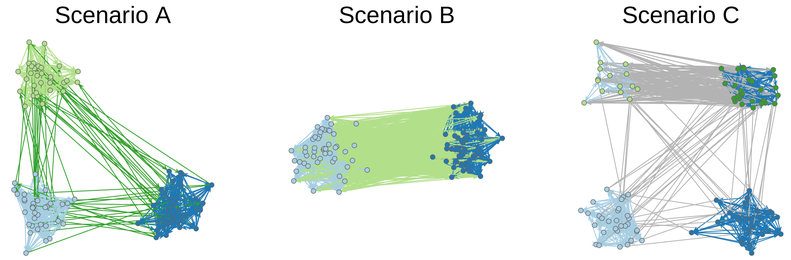

Figure 3. Networks sampled from each scenario. The node colours denote the node cluster memberships and the edge colours denote the majority topic in the corresponding documents.

IC2L evaluation

We evaluate the model selection criterion on Scenario C. The triplet $(\mathrm{K}, \mathrm{P}, \mathrm{Q})$ with the highest IC2L value is selected over 10 networks. The selected dimension for the node embedding space is always $P = 2$. Therefore, Table 2 only presents those models:

	$K=2$	$\mathbf{K=3}$	$K=4$	$K=5$	$K=6$
$Q=2$	0	0	0	0	0
$Q=3$	0	0	0	0	0
$\mathbf{Q=4}$	0	10	0	0	0
$Q=5$	0	0	0	0	0
$Q=6$	0	0	0	0	0

Table 2. Number of times a triplet $(\mathrm{K}, \mathrm{P}, \mathrm{Q})$ is associated with the highest IC2L over 10 graphs simulated according to Scenario C ($Q^\star = 4$ and $K^\star = 3$). All the models with the highest IC2L value correspond to $P = 2$. Therefore, only the table corresponding to this value is shown.

Benchmark study

To evaluate the methodologies, we use an inter-cluster (respectively intra-cluster) probability of connection set to 0.1 (0.01) to make it harder for the model to recover the true structure of the gaph. Moreover, denoting θ^* the proportions of a pure topic (with 1 on the true topic and 0 elsewhere), the mixture proportions of the topic are:

$$\theta_{ij} = (1 - \zeta)\theta^\star_{qr} + \zeta * \left(\frac{1}{K}, \dots, \frac{1}{K}\right)^\top,$$

Table 3. ARI of the node clustering averaged over 10 graphs in all three scenarios. Deep-LPTM, as well as ETSBM, are presented with and without pre-trained embeddings (denoted PT). Moreover, STBM and SBM are also provided as baselines.

	Scenario A	Scenario B	Scenario C
SBM	0.97 ± 0.03	0.00 ± 0.00	0.62 ± 0.1
STBM	0.63 ± 0.23	1.00 ± 0.00	0.66 ± 0.19
ETSBM	0.96 ± 0.10	0.90 ± 0.30	0.72 ± 0.25
ETSBM - PT	0.99 ± 0.01	1.00 ± 0.00	0.74 ± 0.21
Deep-LPTM	0.99 ± 0.02	1.00 ± 0.00	$\mathbf{0.89 \pm 0.15}$
Deep-LPTM - PT	$\mathbf{1.00 \pm 0.01}$	1.00 ± 0.00	0.85 ± 0.18

References

[1] C. Biernacki et al. "Assessing a mixture model for clustering with the integrated completed likelihood". *IEEE TPAMI* (2000).
[2] A. B. Dieng et al. "Topic modeling in embedding spaces". *TACL* (2020).
[3] M. S. Handcock et al. "Model-based clustering for social networks". *JRSS* (2007).
[4] D. Liang et al. "Deep latent position model for node clustering in graphs". *ESANN* (2022).