THE EMBEDDED TOPIC MODEL MEETS THE STOCHASTIC BLOCK MODEL TO CLUSTER GRAPH WITH TEXTUAL DATA.

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

Université Paris Cité[1], Université Côte d'Azur[2], INRIA MAASAI team[3]

Introduction

Communication networks (emails, social networks, IOT, ...) are now ubiquitous and their analysis has become a strategic field to secure our numerical lives. Unfortunately, **most of the existing methods focus on analyzing quantitative networks whereas communication networks come with textual data on the edges**. We consider in this paper networks for which two nodes are linked if and only if they share textual data. We introduce the embedded topics for the stochastic block model (ETSBM) in order to **simultaneously perform clustering on the nodes while modeling the topics** used between the different clusters. A variational-Bayes expectation-maximisation algorithm (VBEM) combined with a stochastic gradient descent (SGD) is used to perform inference. The methodology is evaluated on simulated data.

Model

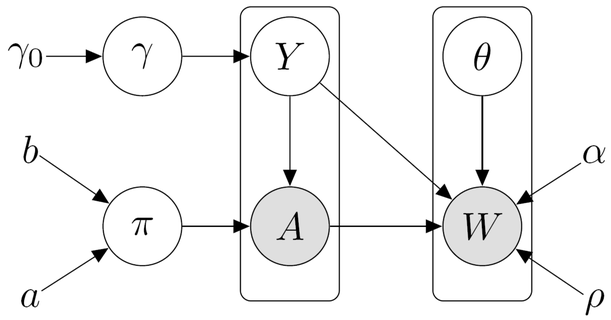

Fig. 1: Graphical representation of the model.

Modelling the interactions

This is based on the **stochastic block model (SBM)** first proposed by Nowicki and Snijders 2001. The random variable γ is used to model the **proportion of the nodes in each cluster**. We assume a Dirichlet distribution as a prior on γ:

$$\gamma \sim \mathcal{D}ir_Q(\gamma_0). \tag{1}$$

Let Y_i denotes the **cluster membership** of node i for all $i \in \{1, \dots, M\}$. They are assumed to follow a multinomial distribution and to be *i.i.d* such that:

$$Y_i \mid \gamma \overset{\text{i.i.d}}{\sim} \mathcal{M}_Q(Y_i; 1, \gamma).$$

The connection between two nodes are supposed to be *i.i.d* given their cluster memberships:

$$A_{ij} \mid Y_{iq}Y_{jr} = 1, \pi_{qr} \overset{\text{i.i.d}}{\sim} \mathcal{B}(\pi_{qr}). \tag{2}$$

Finally, we assume a beta distribution as a prior on the probability matrix $\pi \in \mathcal{M}_{Q\times Q}(\mathbb{R})$ such that:

$$\pi_{qr} \overset{\text{i.i.d}}{\sim} \mathcal{B}eta(a, b).$$

Given the cluster memberships of the nodes Y and the probability matrix π, the connections between the nodes are distributed according to the following joint probability:

$$p(A \mid Y, \pi) = \prod_{i\neq j}^{M}\prod_{q,r}^{Q}\left(\pi_{qr}^{A_{ij}}(1-\pi_{qr})^{(1-A_{ij})}\right)^{Y_{iq}Y_{jr}}. \tag{3}$$

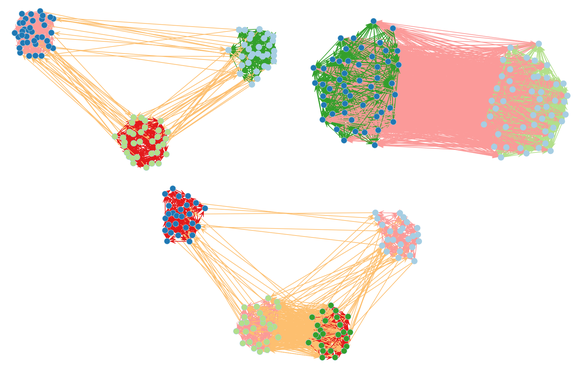

Fig. 2: Example of simulated graphs. The first one corresponds to communities, the second one is one community where people discuss of different topics. The last one is a combination of both.

Modelling the texts

Two nodes are connected if and only if they share textual data, for instance when two persons exchange an email. For the modelling of those texts, we make use of ETM, initially proposed by Dieng, Ruiz, and Blei 2019. We assume a logistic-normal prior distribution on the **topic proportion vector** θ_{qr} for the pair (q, r) of clusters, which is given by:

$$p(\delta_{qr}) = \mathcal{N}(0_K, \mathrm{I}_K), \quad \theta_{qr} = \mathrm{softmax}(\delta_{qr}),$$

where $\mathrm{softmax}(x) = \left(\frac{e^{x_1}}{\sum_{k=1}^K e^{x_k}}, \dots, \frac{e^{x_K}}{\sum_{k=1}^K e^{x_k}}\right)$ and K refers to the number of topics. If two nodes i, j are connected and they are respectively in the clusters q and r, the n-th word of the d-th documents is assumed to be distributed according to a **mixture of topics**:

$$p\left(w_{ij}^{dnv} = 1 \mid \theta_{qr}, Y_{iq}Y_{jr}A_{ij} = 1\right) = \sum_{k=1}^{K}\theta_{qrk}\beta_{kv}, \tag{4}$$

where, for all $k \in \{1, \dots, K\}$, $\beta_k = \mathrm{softmax}\left(\rho^\top \alpha_k\right)$ with $\rho \in \mathcal{M}_{L\times K}(\mathbb{R})$ **the embeddings of the vocabulary** in a vector space of dimension L, and α_k **the embedding of the topic** k in the same vector space. Finally, the following holds:

$$p(W \mid Y, A, \theta, \alpha, \rho) = \prod_{i\neq j}^{M}\prod_{d=1}^{D_{ij}}\prod_{n=1}^{N_{ij}^d}\prod_{q,r}^{Q}\prod_{v=1}^{V}\Big(\sum_{k=1}^{K}\theta_{qrk}\beta_{kv}\Big)^{w_{ij}^{dnv}A_{ij}Y_{iq}Y_{jr}}. \tag{5}$$

Inference

Quantity of interest :

$$\log p(A, W \mid \alpha, \rho) = \log\left(\sum_Y \int_\delta \int_\gamma \int_\pi p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho) d\pi d\delta d\gamma\right) \tag{6}$$

The quantity (6) is untractable since there are M^K configurations of Y and the $\mathrm{softmax}$ over δ makes it impossible to compute. For those reasons, we use a variational distribution $R(\cdot)$ allowing to compute the follwing:

$$\log p(A, W \mid \alpha, \rho) = \mathscr{L}(R(\cdot)) + \underbrace{\mathrm{KL}(R(\cdot)||p(Y, \pi, \gamma, \delta \mid A, W))}_{\geq 0} \tag{7}$$

with $R(\cdot)$ a distribution over Y, π, γ, δ.
$\mathscr{L}(R(\cdot))$ is the **expected lower bound (ELBO)** and is defined by :

$$\mathscr{L}(R(\cdot)) = \sum_Y \int_{\pi,\gamma,\theta} R(Y, \pi, \gamma, \delta) \log \frac{p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho)}{R(Y, \pi, \gamma, \delta)} d\pi d\delta d\gamma$$

Objective : maximise the ELBO over a set of distribution $\mathcal{F}$ with the following constraints:

$$R(Y, \pi, \gamma, \delta) = R(Y)R(\pi)R(\gamma)R(\delta) \quad \text{Mean-field hypothesis}$$
$$R(Y) = \prod_{i=1}^{M}\mathcal{M}(Y_i; \tau_i)$$
$$R(\pi) = \prod_{q,r}\mathcal{B}eta(\pi_{qr}; \tilde{\pi}_{qr1}, \tilde{\pi}_{qr2})$$
$$R(\gamma) = \mathcal{D}ir(\gamma; \tilde{\gamma})$$
$$R(\delta) = \prod_{q,r}^{Q}\mathcal{N}(\delta_{qr}; \mu_{qr}, \sigma_{qr}^2)$$

where $\mu_{qr}^\nu = f(\tilde{w}_{qr}; \nu_m)$ and $\sigma_{qr}^\nu = f(\tilde{w}_{qr}; \nu_v)$, $\tilde{w}_{qr} = \sum_{i\neq j}\tau_{iq}\tau_{jr}A_{ij}W_{ij} = \mathbb{E}_R\left[W_{qr}\right]$.
Since τ lies on the simplex, this constraint is added within the Lagrangian of $\mathscr{L}(\cdot)$ which can be computed and the first order conditions gives the updates of the parameters except for the ETM parameters which are optimised by stochastic gradient descent.

Results & Remarks

		Scenario A		Scenario B	
		Node	Edge	Node	Edge
Easy	STBM	0.98 ± 0.04	0.98 ± 0.04	1.00 ± 0.00	1.00 ± 0.00
	ETSBM	1.00 ± 0.00	0.97 ± 0.03	1.00 ± 0.00	1.00 ± 0.00
Hard 1	STBM	1.00 ± 0.00	0.90 ± 0.13	1.00 ± 0.00	1.00 ± 0.00
	ETSBM	1.00 ± 0.00	0.92 ± 0.06	1.00 ± 0.00	1.00 ± 0.00
Hard 2	STBM	0.99 ± 0.02	0.99 ± 0.01	0.59 ± 0.35	0.54 ± 0.40
	ETSBM	0.44 ± 0.21	0.95 ± 0.08	0.69 ± 0.43	0.98 ± 0.08

		Scenario C	
		Node	Edge
Easy	STBM	1.00 ± 0.00	1.00 ± 0.00
	ETSBM	0.99 ± 0.04	1.00 ± 0.00
Hard 1	STBM	1.00 ± 0.00	0.98 ± 0.03
	ETSBM	0.99 ± 0.04	0.94 ± 0.07
Hard 2	STBM	0.68 ± 0.07	0.62 ± 0.14
	ETSBM	0.30 ± 0.09	1.00 ± 0.00

Fig. 3: Results of the models compared with state of the art method STBM, presented in Bouveyron, Latouche, and Zreik 2016. In experience *hard 1*, the nodes between different clusters are connected with a 0.2 probability. In experience *Hard 2*, 40% of the topics are sampled from the wrong topics (noise in the topics).

This work is on going and **the very good results of ETM** suggest pursuing the efforts in that direction. Some irregularities in the **ELBO** require some extra care.
The next step of this work will be to find a good model selection criteria in order to find the right number of clusters Q and the of topics K.

This work was funded by a grant from the Ecole doctorale 386 and the University of Paris.

References

[BLZ16] C. Bouveyron, P. Latouche, and R. Zreik. "The stochastic topic block model for the clustering of vertices in networks with textual edges". In: *Statistics and Computing* 28.1 (Oct. 2016), pp. 11–31. ISSN: 1573-1375. DOI: 10.1007/s11222-016-9713-7. URL: http://dx.doi.org/10.1007/s11222-016-9713-7.

[DRB19] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. "Topic Modeling in Embedding Spaces". In: *CoRR* abs/1907.04907 (2019). arXiv: 1907.04907. URL: http://arxiv.org/abs/1907.04907.

[NS01] Krzysztof Nowicki and Tom A. B. Snijders. "Estimation and Prediction for Stochastic Blockstructures". In: *Journal of the American Statistical Association* 96 (2001), pp. 1077–1087.