

Clustering networks with textual edges by combining the Embedded Topic Model and the Stochastic Block model and derivation of a model selection criterion.

Rémi Boutin¹, Pierre Latouche¹ and Charles Bouveyron²

¹ MAP5 - Université de Paris Cité

² Maasai team - INRIA, Université Côte d'Azur

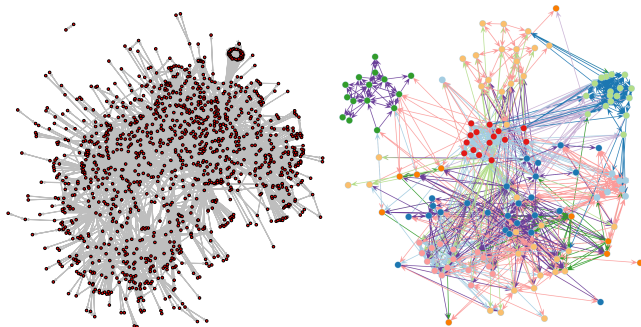
JDS 2022



Introduction

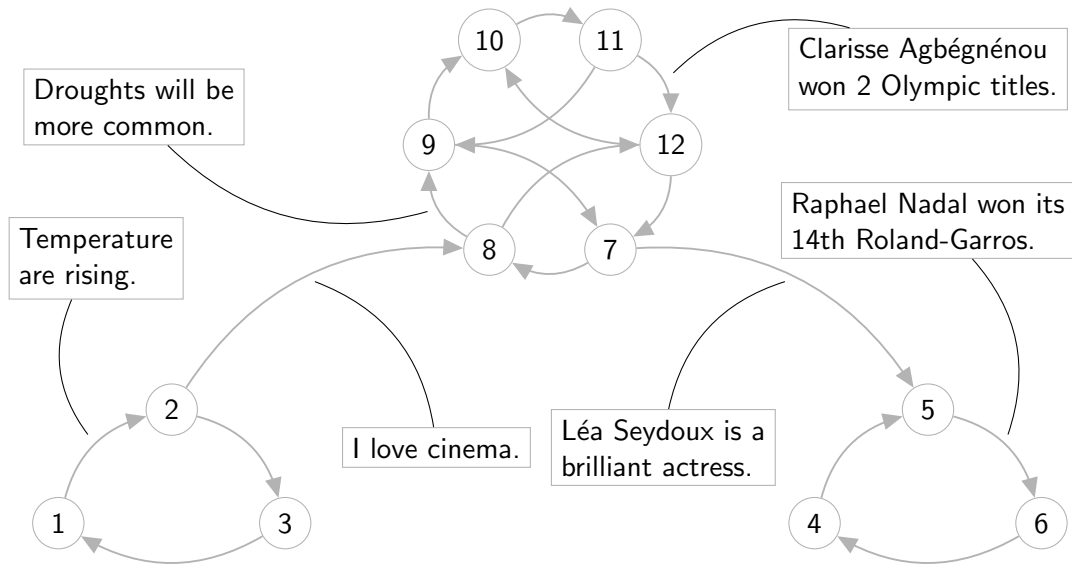
Networks can be observed directly or indirectly from a variety of sources:

- ▶ social websites (Facebook, Twitter, ...),
- ▶ emails (from your Gmail, Clinton's mails, Enron Email data ...),
- ▶ digital/numeric documents (Panama papers, co-authorships, ...),
- ▶ and even archived documents in libraries (digital humanities).

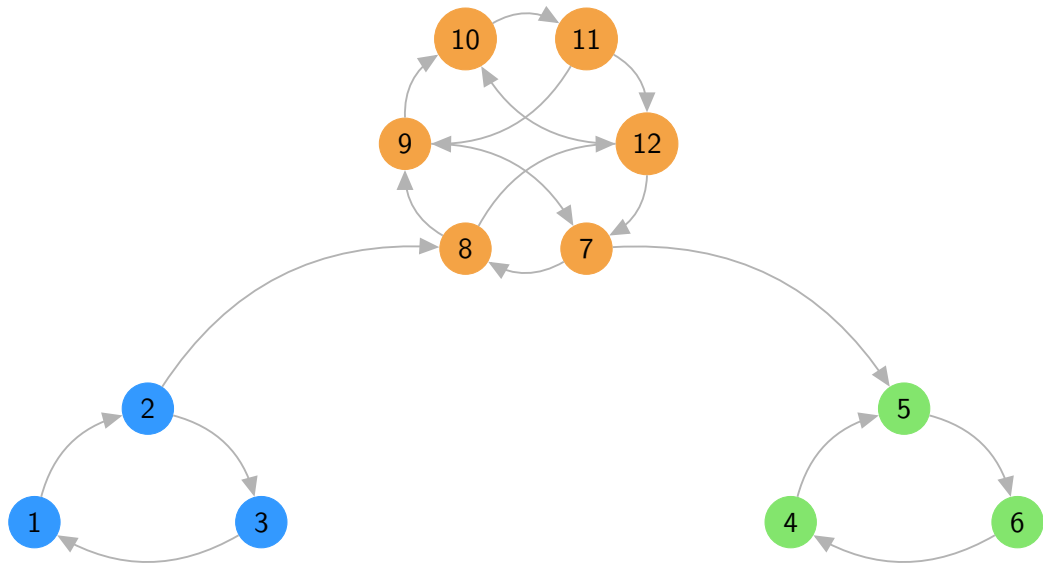


⇒ most of these sources involve text!

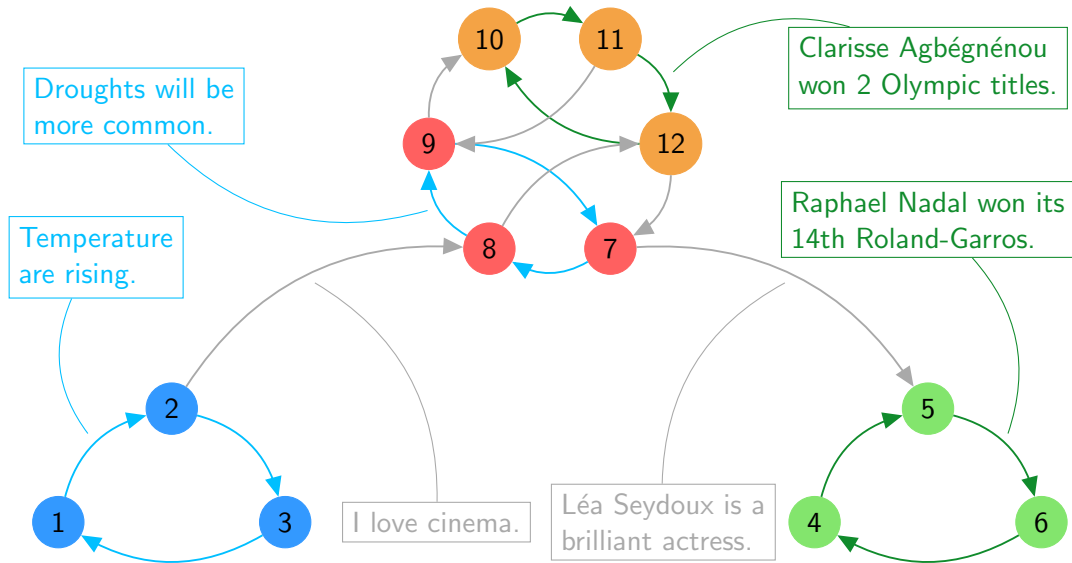
Observed data



Clustering of nodes



Our objective with ETSBM



Node generation

Based on the Stochastic Block Model (SBM), [HLL83] Y_i the cluster membership of node i for all $i \in \{1, \dots, M\}$

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

where γ , the cluster proportion, lives in Δ_{Q-1} , the simplex of dimension Q , the number of clusters. **The $(Y_i)_i$ are not observed, they are latent variables.**

We assume that two nodes are connected with a probability only depending on their clusters

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

Node generation

Since a **Bayesian framework** for SBM proved to provide an **efficient criterion for selecting the number of clusters** Q , see [LBA12] and it offers penalization that can be beneficial in a noisy setting, we assume a product of Dirichlet distributions as a prior for the cluster proportions,

$$\gamma \sim \text{Dir}_Q(\gamma_0).$$

We also assume a Beta distribution as a prior on the probability matrix $\pi \in \mathcal{M}_{Q \times Q}(\mathbb{R})$, such that for all $q, r \in \{1, \dots, Q\}$,

$$\pi_{qr} \stackrel{\text{i.i.d}}{\sim} \text{Beta}(a, b).$$

Edge generation in ETSBM

Based on the **Embedded Topic Model**, see [DRB20], if given that node i, j belongs to cluster q and r respectively, the document d sent from i to j is generated as follow:

1. $\delta_{qr} \sim \mathcal{N}(0, I_K)$
2. $\theta_{qr} = \text{softmax}(\delta_{qr})$, topic proportions of documents sent from q to r
3. for each word n in doc d :
 - (i) $z_{dn} \sim \mathcal{M}_K(1, \theta_{qr})$, draw a topic with this proportions
 - (ii) $w_{dn} \mid z_{dnk} = 1 \sim \mathcal{M}_V(1, \beta_k)$, draw a word following this topic

where

$$\beta_k = \text{softmax}(\rho^\top \alpha_k)$$

vocabulary $\in \mathbb{R}^{V \times L}$ topic vector $\in \mathbb{R}^L$

Edge generation in ETSBM

Based on the **Embedded Topic Model**, see [DRB20], if given that node i, j belongs to cluster q and r respectively, the document d sent from i to j is generated as follow:

1. $\delta_{qr} \sim \mathcal{N}(0, I_K)$
2. $\theta_{qr} = \text{softmax}(\delta_{qr})$, topic proportions of documents sent from q to r
3. for each word n in doc d :
 - (i) $z_{dn} \sim \mathcal{M}_K(1, \theta_{qr})$, draw a topic with this proportions
 - (ii) $w_{dn} \mid z_{dnk} = 1 \sim \mathcal{M}_V(1, \beta_k)$, draw a word following this topic

where

$$\beta_k = \text{softmax}(\rho^\top \alpha_k)$$

vocabulary $\in \mathbb{R}^{V \times L}$ topic vector $\in \mathbb{R}^L$

STBM, [BLZ18] is based Latent Dirichlet Analysis, which does not allow to use pre-trained embeddings to incorporate semantic meaning.

ETSBM graphical model

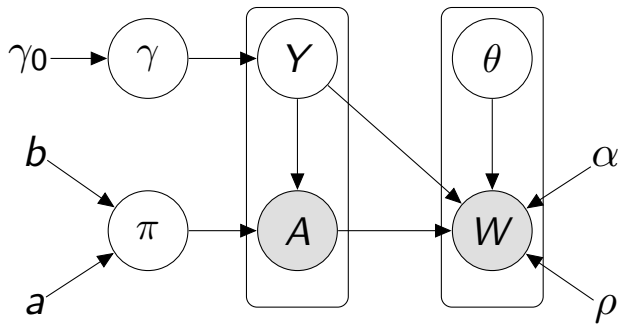


Figure: Graphical representation of ETSBM.

Inference & Model selection

We aim at maximizing this quantity w.r.t ρ and α

$$\log p(A, W \mid \alpha, \rho) = \log \left(\sum_Y \int_{\delta} \int_{\gamma} \int_{\pi} p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho) d\pi d\delta d\gamma \right). \quad (2)$$

This quantity is **intractable** because it would require computing M^Q terms. To keep the inference tractable to large graphs, we use **variational inference** (VI).

Inference & Model selection

Variational Inference consists in splitting (2) in two terms using a **surrogate distribution** on Y , π , γ and δ , denoted $R(Y, \pi, \gamma, \delta)$, restricted to a family distributions described in the next slide,

$$\log p(A, W \mid \alpha, \rho) = \mathcal{L}(R(\cdot)) + \text{KL}(R(\cdot) \parallel p(Y, \pi, \gamma, \delta \mid A, W)),$$

with

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

Inference & Model selection

We choose the same family distributions as the one involved in the generative model,

$$R(Y, \pi, \gamma, \delta) = R(Y)R(\pi)R(\gamma)R(\delta), \quad \textbf{Mean-field hypothesis}$$

$$R(Y) = \prod_{i=1}^M \mathcal{M}(Y_i; \tau_i), \quad R(\pi) = \prod_{\substack{q,r \\ Q}} \text{Beta}(\pi_{qr}; \tilde{\pi}_{qr1}, \tilde{\pi}_{qr2}),$$
$$R(\gamma) = \text{Dir}(\gamma; \tilde{\gamma}), \quad R(\delta) = \prod_{q,r} \mathcal{N}(\delta_{qr}; \mu_{qr}, \sigma_{qr}^2),$$

where $\mu_{qr}^\nu = f(\tilde{w}_{qr}; \nu_m)$ and $\sigma_{qr}^\nu = f(\tilde{w}_{qr}; \nu_v)$, $\tilde{w}_{qr} = \sum_{i \neq j} \tau_{iq} \tau_{jr} A_{ij} W_{ij} = \mathbb{E}_R[W_{qr}]$.

Inference & Model selection

Setting the derivatives of the ELBO with respect to $\tilde{\pi}, \tilde{\gamma}$ to zero gives the **classical updates of the variational Bayes SBM**, [LBA12],

$$\tilde{\gamma}_q = \gamma_{0q} + \sum_{i=1}^M \tau_{iq}, \quad (3)$$

$$\tilde{\pi}_{qr1} = \pi_{qr1}^0 + \sum_{i \neq j}^M \tau_{iq} \tau_{jr} X_{ij}, \quad \tilde{\pi}_{qr2} = \pi_{qr2}^0 + \sum_{i \neq j}^M \tau_{iq} \tau_{jr} (1 - X_{ij}). \quad (4)$$

Inference & Model selection

Setting the derivatives of the ELBO with respect to $\tilde{\pi}, \tilde{\gamma}$ to zero gives the **classical updates of the variational Bayes SBM**, [LBA12],

$$\tilde{\gamma}_q = \gamma_{0q} + \sum_{i=1}^M \tau_{iq}, \quad (3)$$

$$\tilde{\pi}_{qr1} = \pi_{qr1}^0 + \sum_{i \neq j}^M \tau_{iq} \tau_{jr} X_{ij}, \quad \tilde{\pi}_{qr2} = \pi_{qr2}^0 + \sum_{i \neq j}^M \tau_{iq} \tau_{jr} (1 - X_{ij}). \quad (4)$$

ETM parameters are optimised following [DRB20]. To update τ with gradient descent, **we switch from the simplex Δ_{Q-1} to the unconstrained vector space \mathbb{R}^{Q-1}** through the transformation:

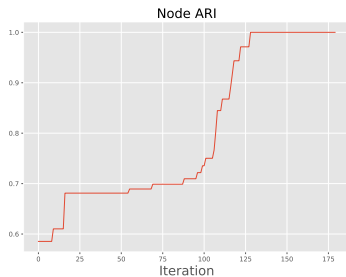
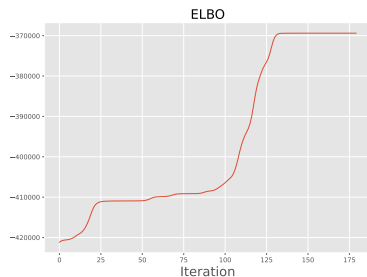
$$\xi_{iq} = \ln(\tau_{iq}) - \ln(\tau_{i,Q}), \quad \forall i \in \{1, \dots, M\}, \forall q \in \{1, \dots, Q-1\}$$

We then use auto differentiation with respect to ξ .

Inference & Model selection

As mentioned previously, the **Bayesian framework** allows the **ELBO to penalized models with a high number** of parameters and therefore to be a **natural and efficient selection model criteria**, [LBA12].

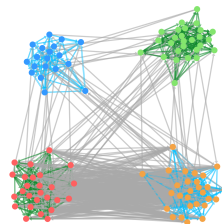
Simulations - Detailed example in Scenario C



(a)

Topics		
library	duke	lost
science	kensington	kingdom
credit	granddaughter	government
event	duchess	snp
shadow	palace	united
horizon	queen	resentment
light	cambridge	party
gravity	birth	political
hole	charlotte	david
black	princess	seats

Graph obtained by ETSBM algorithm



(b)

Simulations - Initialisation

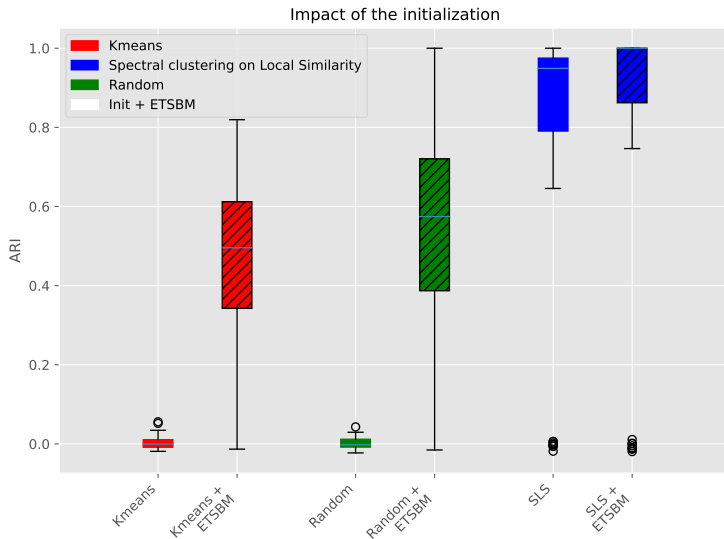


Figure: Our new initialization use textual information as well as graph information.

Simulations - Model selection

We choose to **only select the number of clusters Q** and to pick only the most used topics afterwards.

$\begin{array}{c c} & Q \\ \hline K & \end{array}$	2	3	4	5	10
2	0	78	12	8	2
3	0	76	12	12	0
4	0	82	14	4	0
5	0	78	12	10	0
10	0	76	14	10	0

Table: Scenario A: $Q_{\text{true}} = 3$,
 $K_{\text{true}} = 4$

$\begin{array}{c c} & Q \\ \hline K & \end{array}$	2	3	4	5	10
2	100	0	0	0	0
3	100	0	0	0	0
4	94	0	2	2	2
5	92	0	0	8	0
10	96	0	4	0	0

Table: Scenario B: $Q_{\text{true}} = 2$,
 $K_{\text{true}} = 3$



$\begin{array}{c c} & Q \\ \hline K & \end{array}$	2	3	4	5	10
2	0	2	98	0	0
3	0	0	96	0	4
4	0	2	96	0	2
5	0	0	98	0	2
10	0	2	96	0	2

Table: Scenario C: $Q_{\text{true}} = 4$,
 $K_{\text{true}} = 3$

Conclusion & further work

- ▶ Bayesian framework successfully select the number of clusters
- ▶ Automatic differentiation works fine
- ▶ “Soft” meta documents capture the textual information between clusters
- ▶ Our model captures meaningful clusters both in terms of connections and topics
- ▶ Improving our simulation scheme
- ▶ Try the model with pre-trained embeddings
- ▶ How to incorporate temporal information ?

Bibliography

-  Charles Bouveyron, Pierre Latouche, and Rawya Zreik, *The stochastic topic block model for the clustering of vertices in networks with textual edges*, Statistics and Computing **28** (2018), no. 1, 11–31.
-  Adji B Dieng, Francisco JR Ruiz, and David M Blei, *Topic modeling in embedding spaces*, Transactions of the Association for Computational Linguistics **8** (2020), 439–453.
-  Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, *Stochastic blockmodels: First steps*, Social networks **5** (1983), no. 2, 109–137.
-  Pierre Latouche, Etienne Birmele, and Christophe Ambroise, *Variational bayesian inference and complexity control for stochastic block models*, Statistical Modelling **12** (2012), no. 1, 93–115.

Feel free to send me an email if you have any question !

This is a joint work between:

- * **Rémi Boutin**, remi.boutin@u-paris.fr
- * Pierre Latouche,
- * Charles Bouveyron.