Embedded Topics in the Stochastic Block Model

Rémi Boutin, Pierre Latouche, Charles Bouveyron

Embedded Topics in the Stochastic Block Model

Rémi Boutin[a], Charles Bouveyron[b], Pierre Latouche[c,a]

[a] *Université Paris Cité, MAP5, CNRS, UMR 8145, Paris, France*
[b] *Université Côte d'azur, INRIA MASAI Team, Sophia-Antipolis, France*
[c] *Université Clermont Auvergne, CNRS, Laboratoire de Mathématiques Blaise Pascal, UMR 6620, Clermont-Ferrand, France*

Abstract

Communication networks such as emails or social networks are now ubiquitous and their analysis has become a strategic field. In many applications, the goal is to automatically extract relevant information by looking at the nodes and their connections. Unfortunately, most of the existing methods focus on analysing the presence or absence of edges and textual data is often discarded. However, all communication networks actually come with textual data on the edges. In order to take into account this specificity, we consider in this paper networks for which two nodes are linked if and only if they share textual data. We introduce a deep latent variable model allowing embedded topics to be handled called ETSBM to simultaneously perform clustering on the nodes while modelling the topics used between the different clusters. ETSBM extends both the stochastic block model (SBM) and the embedded topic model (ETM) which are core models for studying networks and corpora, respectively. The inference is done using a variational-Bayes expectation-maximisation algorithm combined with a stochastic gradient descent. The methodology is evaluated on synthetic data and on a real world dataset.

Keywords: Graph clustering, topic modelling, variational inference, generative model, probabilistic model, embedded topic model, stochastic block model

Email address: remi.boutin.stat@gmail.com (Rémi Boutin)

1. Introduction

Many real life interactions induce the exchange of texts, as in co-authorship networks, social networks or emails for instance. Since the storage capacity keeps increasing, networks with textual data on the edges become even more frequent. In order to make such networks, called communication networks, intelligible to humans, it is of great interest to gather information about the texts exchanged between the nodes and to summarise the connectivity structure. While those two questions have been studied independently, the work we propose aims at bridging the gap between the two by modelling the joint distribution of texts and edges. Indeed, this work aims at simultaneously modelling the connectivity of the nodes and the topics present in the exchanged texts to find meaningful clusters of nodes. To the best of our knowledge, the interest on making the two disciplines of topic modelling, when texts are present on the edges, and model-based graph clustering meets is recent and the methods that have been proposed only rely on the frequency of word within the documents without incorporating semantic meaning. In this paper, we propose to take advantage of pre-trained word embeddings in the topic-model as presented in Dieng et al. (2020) in order to incorporate semantic meaning of the words and to obtain topic-meaningful clusters. Figure 1, illustrates the necessity to combine graph clustering and topic modelling in order to distinguish all four clusters and to obtain more meaningful topics for each cluster.

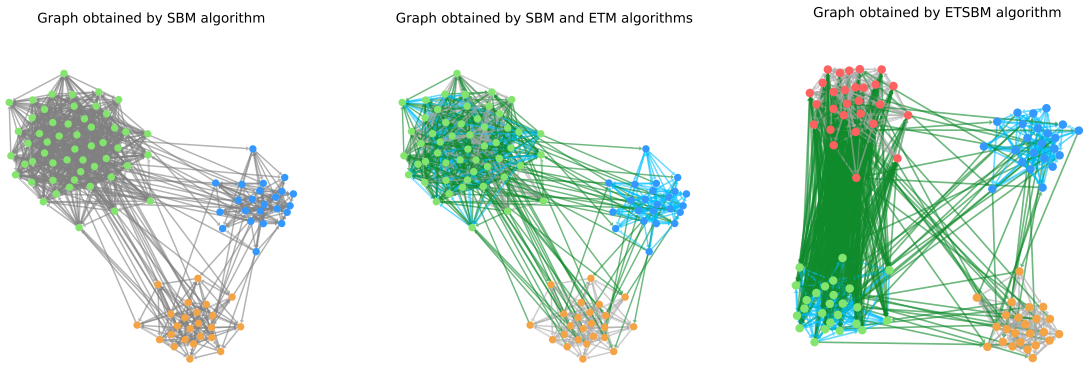


Figure 1: Comparison of results on a simulated network with SBM on the left, SBM and ETM in the middle and ETSBM on the right. The colours of the nodes and edges indicate the cluster assignment and the major topic used in the corresponding document respectively. Note that SBM alone does not provide edge information. Thus, the left network only has a single edge colour. On the left hand side, SBM clustering results uncover 3 clusters. In the middle, ETM edge information is added to the network through the 3 edge colours green, grey and blue. On the right hand side, ETSBM clustering results uncover 4 clusters. The cluster coloured in green, in the middle of the figure, is split into two cluster on the right hand side, the green one and the red one, each discussing of a different topic, the blue and grey topic respectively.

2. Related work

Both the topic modelling methods and the graph clustering techniques have first emerged as deterministic optimisation problems to progressively incorporate uncertainty which led to many developments in the statistical literature. The next part provides a brief summary of the advancements in those domains.

2.1. Probabilistic models for topic modelling

The statistical analysis of topics has emerged in the late 90s with Papadimitriou et al. (1998), developing statistical results for the latent semantic indexing (LSI), first proposed by Deerwester et al. (1990). LSI relies on a spectral analysis of the "term frequency - inverse document frequency" and successfully captures synonymy between words. To overcome the lack of probabilistic foundations of LSI, Hofmann (1999) introduced the probabilistic latent semantic index (pLSI) which models each word distribution as a mixture model such that each mixture component corresponds to a "topic". The topic membership of each word is modelled by a multinomial random variable in pLSI. Even though the topic membership of the words depends on the document, a major drawback of this model is the absence of model at the the document level. This was overcome by Blei et al. (2003) with the latent Dirichlet allocation (LDA) which for each document uses a Dirichlet random variable to model the proportion of each topic. However, the Dirichlet distribution makes the topics almost uncorrelated and does not directly model correlation. Blei & Lafferty (2006) then proposed to use a normal-logistic prior instead of a Dirichlet prior on the topic proportion to directly model the correlations. All these models require to derive the equations for any new generative model. In Srivastava & Sutton (2017), they bridged the gap between topic modelling and autoencoders, taking full advantage of gradient descent for those models. Nevertheless, all the former approaches do not incorporate semantic meaning to the words. Indeed, since the model is only based on the document term-frequency matrix, they loose the information provided by the order of the words. In the embedded topic model (ETM), Dieng et al. (2020) used the strength of word embeddings, such as the continuous bag of words (CBOW) or skipgram (Mikolov et al., 2013) as a part of the decoder of a variational autoencoder (VAE). The topics are also embedded into the same vector space which allows to easily measure similarities between words and topics. The optimisation is done using gradient descent,

as proposed in Rezende et al. (2014) or Kingma & Welling (2014). For a review of the former methods relying exclusively on the document term frequency matrix, the reader may refer to Vayansky & Kumar (2020).

2.2. Probabilistic models for graph analysis

Statistical network analysis first started with random graph theory, initiated by Erdos et al. (1960). They studied probabilistic properties of graphs with binary connections, and a unique probability for any connection to exist. However, real life datasets do not show such regularity. Therefore, more complex and realistic graph structures have been considered. Here, a structure designates a partition of the nodes such that nodes in a cluster present a homogeneous connectivity pattern. For example, a community is a group of nodes highly connected one to another but with few connections to the rest of the graph. If the graph is only composed of communities, reordering the adjacency matrix by group would output a block matrix. Another direction emerged with Fienberg & Wasserman (1981) who first introduced a probabilistic model that assumes that the probability for two nodes to be connected only depends on the group to which they belong to and applied it to Sampson's monastery dataset (Sampson, 1969). Introducing a latent representation of the nodes then became popular thanks to the latent position cluster model (Handcock et al., 2007) or the stochastic block model (SBM) (Wang & Wong, 1987; Nowicki & Snijders, 2001). Many extensions have been developed to incorporate valued edges, as in Mariadassou et al. (2010), as well as categorical edges in Jernite et al. (2014) or to add prior information in Zanghi et al. (2010). Some developments also focused on looking for overlapping clusters (Airoldi et al., 2008; Latouche et al., 2011) as well as dynamic networks (Matias & Miele, 2017; Zreik et al., 2017; Corneli et al., 2016). The inference of SBM-based model is often done either using Markov chain monte carlo (MCMC), variational expectation maximisation (VEM) as in Daudin et al. (2008) or variational Bayes expectation maximisation (VBEM) as in Latouche et al. (2012). The classification can either be deduced from the latent variable distribution or be incorporated in the optimisation strategy with a hard clustering, for instance using the classification variational expectation maximisation (CVEM) algorithm (Bouveyron et al., 2018). The choice of the number of cluster K can either be done through a model selection criterion (Daudin et al., 2008; Latouche et al., 2012), through a greedy search (Côme & Latouche, 2015) or through a non parametric schemes (Kemp et al., 2006).

Fore more insights about SBM developments, see Lee & Wilkinson (2019). For reviews on statistical network modelling, we also relate to Goldenberg et al. (2010) and Matias & Robin (2014).

2.3. Probabilistic models for the joint analysis of texts and networks

The rise of data combining networks with texts, such as emails, social networks or co-authors articles led to developing methods using both the network and the textual information. In that regard, Zhou et al. (2006) proposed the community-user topic model (CUT). This model relies on the author-topic model (AT) (Rosen-Zvi et al., 2004) and adds a latent variable to the bayesian hierarchical model for modelling the communities. Two versions are proposed in the paper, CUT1 hypothesises that a community is entirely defined as a group of users while CUT2 makes the assumption that a community is defined as a set of topics. This model is inferred using a Gibbs sampler to approximate the joint distribution of the communities, topics and users. Eventually, the community-author-recipient-topic (CART) model introduced in Pathak et al. (2008) makes use of communities both at the document generation level and at the author and recipient generation level which corresponds to the network generation. However, the high number of parameters combined with the inference based on a Gibbs sampler does not allow to scale this model to large datasets. The topic-link LDA presented in Liu et al. (2009) also offers a joint-analysis of texts and links in a unified framework by conditioning the generation of a link on both the topics within the documents and the community of authors. The inference relies on a variational EM approach which allows to scale to large datasets but this method only deals with undirected networks. Finally, the topic-user-community models (TUCM) was introduced in Sachan et al. (2012) and was able to discover topic-meaningful communities. The main feature of this model was its capacity to incorporate different types of interactions, well-suited for social networks applications (tweets, retweets, messages, comments, ...). The inference relies on Gibbs sampling which can be limiting when dealing with large datasets. The stochastic topic block model (STBM) presented in Bouveyron et al. (2018) was the first model to handle the simultaneous clustering of nodes and edges while keeping the inference tractable to large dataset thanks to a variational classification EM based inference. This model was extended in Bergé et al. (2019) for the simultaneous clustering of rows (observations) and columns (variables). It was also adapted for dynamic networks in Corneli et al. (2019).

Unfortunately, those models only rely on word counts and cannot use the position of words within a sentence or any form of context information.

Our contribution. This paper aims at bridging the gap between network analysis and topic modelling by performing a simultaneous clustering of nodes and edges. It distinguishes from former method by being able to learn a representation of topics using either the occurrence of word as before or incorporating a pre-trained representation of the vocabulary. The former allows to incorporate semantic meaning in our model-based clustering algorithm. To the best of our knowledge, it is the first strategy for model-based clustering that can perform both nodes and edges clustering simultaneously while incorporating semantic meaning in the analysis.

Organisation of the paper. The embedded topics for the stochastic block model (ETSBM) is presented in Section 3. The inference and the model selection are presented in Section 4. Eventually, the model is evaluated against state of the art algorithms on synthetic data and we present results for a real word example build from tweets during the last French presidential election in Sections 5 and 6, respectively. Section 7 presents some concluding remarks and further work.

3. The ETSBM Model

3.1. Background and notations

In this work, we focus on data represented by a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, such that $\mathcal{V} = \{1, \dots, M\}$ denotes the set of nodes and $\mathcal{E} := \{(i,j) : i,j \in \{1,\dots,M\}, i \rightsquigarrow j\}$ the set of edges, where $i \rightsquigarrow j$ indicates that i is connected to j. The connections, or edges, are represented by a binary matrix $A \in \mathcal{M}_{M\times M}(\{0,1\})$ such that i is connected to j, or $(i,j) \in \mathcal{E}$, if and only if $A_{ij} = 1$. In the applications we consider, this implies that node i sent textual information to j such as one or a series of emails for instance. These texts are denoted $W_{ij} = \{W_{ij}^1, \dots, W_{ij}^{D_{ij}}\}$ with D_{ij} the number of documents sent from i to j and are gathered in the collection $W = \{W_{ij}, (i,j) \in \mathcal{E}\}$. Each document d in W_{ij} is a collection of words of size N_{ij}^d, i.e $W_{ij}^d = \{w_{ij}^{d1}, \dots, w_{ij}^{dN_{ij}^d}\}$. The size of the vocabulary is denoted V and the words are identified by their index in the vocabulary: each word w is in $\{1, \dots, V\}$. Finally, only graphs without self loops are considered in this work, therefore $A_{ii} = 0$ for all $i \in \mathcal{V}$. Notice that all the present work can be extended to

undirected networks using $W_{ij} = W_{ji}$ for all pairs (i, j) such that $A_{ij} = A_{ji} = 1$. The directed case is more adequate to messages sent from i to j while the undirected case is better suited for co-authorships networks for instance.

The notation $\mathcal{M}_{d\times p}(\mathbb{F})$ will be used to denote the matrix space with matrix of dimension $d \times p$ and coefficients in $\mathbb{F}$ while the notation $\mathcal{M}_d(N, \omega)$ will be used to denote the multinomial distribution with parameters $N \in \mathbb{N}$ and $\omega \in \Delta_{d-1}$ where

$$\Delta_{d-1} \ =: \ \left\{ x \in \mathbb{R}^d : \forall i \in \{1, \dots, d\}, x_i \geq 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

3.2. Modelling the interactions

In this work, we assume that each node belongs to a single cluster. Moreover, we assume that the connexion probability between two nodes only depends on the cluster memberships. Indeed, let Y_i denotes the cluster membership of node i for any $i \in \{1, \dots, M\}$. All Y_i are assumed to follow a multinomial distribution and to be independent and identically distributed (*i.i.d*), given the cluster proportions $\gamma \in \Delta_{Q-1}$, lying in the simplex of dimension Q,

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

Thus, each node i is associated with cluster q with probability γ_q. Then, we define the cluster membership matrix Y by stacking the node cluster membership vectors $(Y_i)_i$ together such that $Y = (Y_1 \cdots Y_M)^\top \in \mathcal{M}_{M\times Q}(\{0, 1\})$. The probability of Y is given by

$$p(Y \mid \gamma) = \prod_{i=1}^{M} \prod_{q=1}^{Q} \gamma_q^{Y_{iq}}. \tag{1}$$

Besides, the connections between nodes are supposed to be independent given their cluster memberships. Moreover, if nodes i and j are respectively in clusters q and r, an edge is assumed to be present with probability π_{qr},

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

where $\mathcal{B}(\mu)$ denotes the Bernoulli distribution with probability μ. Thus, given the cluster memberships of the nodes Y and the probability matrix π, the probability of all node

connections is given by

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

Eventually, the joint-probability of the adjacency matrix A, and the cluster memberships vector Y, is obtained by multiplying Equations (1) and (3),

$$p(A, Y \mid \pi, \gamma) = p(A \mid Y, \pi) p(Y \mid \gamma). \tag{4}$$

Combining Equations (1), (3), and (4), we retrieve the SBM distribution (Daudin et al., 2008).

3.3. Modelling the texts

Our approach extends ETM to capture information of groups of texts. Essentially, texts are assumed to be generated according to a mixture of topics with latent topic vectors only depending on node clusters. More precisely, a text sent from node i in cluster q to node j in cluster r is assumed to have a logistic-normal topic proportion vector $\theta_{qr} = (\theta_{qr1}, \ldots, \theta_{qrK})^\top \in \Delta_{K-1}$, with the number of topics K fixed beforehand. It is obtained by applying the softmax function to a Gaussian random vector δ_{qr},

$$\begin{aligned} \delta_{qr} &\sim \mathcal{N}(0_K, I_K), \\ \theta_{qr} &= \text{softmax}(\delta_{qr}), \end{aligned}$$

where $\text{softmax}(x) = \left(\sum_{k=1}^{K} e^{x_k} \right)^{-1} (e^{x_1}, \ldots, e^{x_K})^\top$.

In the rest of this paper, the notation $\theta = (\theta_{qr})_{1 \leq q,r \leq Q}$ is used to refer to the topic proportions while $\delta = (\delta_{qr})_{1 \leq q,r \leq Q}$ will refer to the sampling of the random variable. If two nodes i and j are connected and if they are respectively in cluster q and r, the words in document W_{ij} are assumed to be *i.i.d.* Indeed, the n-th word of the d-th documents is assumed to be distributed according a mixture of topics conditionally on the node clusters,

$$W_{ij}^{dn} \mid Y_{iq} Y_{jr} A_{ij} = 1, \theta_{qr}, \alpha, \rho \sim \mathcal{M}_V(1, \theta_{qr}^\top \beta), \tag{5}$$

where the matrix $\beta = (\beta_1 \cdots \beta_K)^\top \in \mathcal{M}_{K \times V}(\mathbb{R})$ corresponds to the distribution over

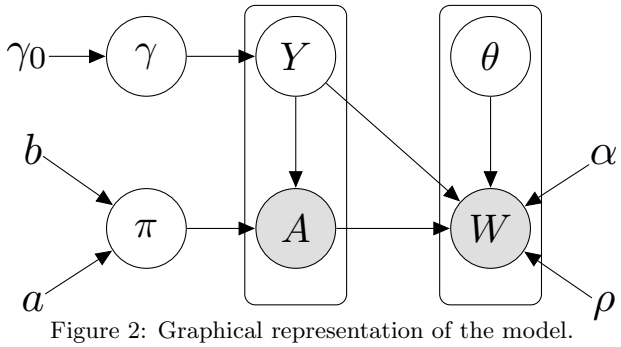


Figure 2: Graphical representation of the model.

the vocabulary for each topic such that $\beta_k = \text{softmax}\left(\rho^\top \alpha_k\right)$ for any $k \in \{1, \dots, K\}$. The matrix $\rho \in \mathcal{M}_{L\times V}(\mathbb{R})$ corresponds to the matrix of the vocabulary embedded into an L-dimensional vector space, and $\alpha = (\alpha_1 \cdots \alpha_K) \in \mathcal{M}_{L\times K}(\mathbb{R})$ the matrix of topics represented into the same vector space.

Therefore, the probability of texts can be computed as follow:

$$\begin{aligned} p(W \mid Y, A, \theta, \alpha, \rho) &= \prod_{i\neq j}^{M} \prod_{d=1}^{D_{ij}} p(W_{ij} \mid Y_i, Y_j, A_{ij} = 1, \theta, \alpha, \rho) \\ &= \prod_{i\neq j}^{M} \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^d} \prod_{q,r}^{Q} \prod_{v=1}^{V} \Big(\sum_{k=1}^{K} \theta_{qrk} \beta_{kv} \Big)^{W_{ij}^{dnv} A_{ij} Y_{iq} Y_{jr}} \\ &= \prod_{q,r}^{Q} \prod_{v=1}^{V} \Big(\sum_{k=1}^{K} \theta_{qrk} \beta_{kv} \Big)^{W_{qr}^{v}}. \end{aligned} \tag{6}$$

The number of time the word v of the dictionary is used in texts sent from cluster q to cluster r is denoted $W_{qr}^v = \sum_{i\neq j}^{M} \sum_{d=1}^{D_{ij}} \sum_{n=1}^{N_{ij}^d} W_{ij}^{dnv} A_{ij} Y_{iq} Y_{jr}$. Hereafter, $W_{qr} = (W_{qr}^1, \dots, W_{qr}^V)^\top \in \mathbb{N}^V$ shall be designated as meta-document (q, r). Moreover, we shall use the bag of words notations such that for any connected pair of nodes $(i, j) \in \mathcal{E}$, $W_{ij} = (W_{ij}^1, \dots, W_{ij}^V)^\top \in \mathbb{N}^V$ with for any $v \in \{1, \dots, V\}$, W_{ij}^v represents the total count of word v for all documents sent from i to j. The model is represented in Figure 2.

3.4. Distribution of the model and links with SBM and ETM.

Given a cluster configuration Y, the joint probability of the model is obtained using Equations (3) and (6)

$$p(A, W \mid Y, \alpha, \rho) = p(W \mid Y, A, \alpha, \rho) p(A \mid Y, \pi). \tag{7}$$

At this point, we emphasise that meta-documents between pairs of clusters of nodes

are constructed using the cluster memberships Y and the node connections A. Assuming that the cluster membership Y is available as well as all the network information holded by π and γ, the model we propose would simply correspond to ETM applied on the meta-documents $(W_{qr})_{1\leq q,r,\leq Q}$, computed with the available Y.

On the other hand, if no texts are exchanged between nodes or the texts are not available, the distribution would reduce to the second term of Equation 7. In that case, the conditional distribution of a standard SBM (Daudin et al., 2006) is recovered. It is also worth noticing that if a Dirichlet prior is assumed on the topic proportion instead of a logistic-normal, and no factorisation in a embedded latent space is considered, the model corresponds to STBM. By construction, ETSBM generalises SBM and ETM to incorporate both textual data and network information.

4. Inference

This section presents the Bayesian framework considered for inference. It also describes the variational-bayes EM algorithm used to maximise the integrated joint likelihood.

4.1. Bayesian framework for the graph modelling part

First, a Dirichlet distribution is assumed as a prior distribution on the proportions γ of nodes in each cluster,

$$\gamma \sim \mathcal{D}ir_Q(\gamma_0). \tag{8}$$

where γ_0 is set to $(1,\dots,1) \in \mathbb{R}^Q$, which corresponds to a uniform prior on the simplex. Moreover, each coefficient of the probability matrix $\pi \in \mathcal{M}_{Q\times Q}(\mathbb{R})$, is assumed to be sampled from from a Beta distribution, such that for any pair $(q,r) \in \{1,\dots,Q\}^2$,

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

In particular, a and b are set to 1. Thus, the Beta prior corresponds to a Uniform distribution between 0 and 1.

4.2. Variational inference

Eventually, the integrated joint log-likelihood is given by:

$$\log p(A, W \mid \alpha, \rho) = \log \left(\sum_{Y} \int_{\delta} \int_{\gamma} \int_{\pi} p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho) d\pi d\delta d\gamma \right). \tag{9}$$

Unfortunately, this quantity is intractable since it requires computing it for the Q^M configurations of Y, which is naturally computationally too demanding. Moreover, the integral with respect to δ is not tractable either because of the softmax function. Thus, it cannot be optimised directly. However, it is possible to overcome this issue using a variational-bayes expectation-maximisation algorithm (VBEM) Attias (1999). This comes handy as it makes the inference scalable to large datasets.

The variational approach consists in splitting Equation (9) in two terms using a surrogate distribution on Y, π, γ and δ, denoted $R(Y, \pi, \gamma, \delta)$.

Proposition 4.1. *Denoting $R(\cdot)$, a distribution on Y, π, γ and δ, the integrated joint log-likelihood can be decomposed as follow:*

$$\log p(A, W \mid \alpha, \rho) = \mathscr{L}(R(\cdot); \alpha, \rho) + \mathrm{KL}(R(\cdot) || p(Y, \pi, \gamma, \delta \mid A, W, \alpha, \rho)),$$

where

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

Proof. The proof is provided in Appendix A. □

To make $\mathscr{L}(R(\cdot); \alpha, \rho)$ tractable, we use the following mean-field assumption :

$$R(Y, \pi, \gamma, \delta) = R(Y) R(\pi) R(\gamma) R(\delta). \tag{10}$$

Following the optimality results of Latouche et al. (2012), we impose the following vari-

ational distributions:

$$
\begin{aligned}
R(Y) &= \prod_{i=1}^{M} R(Y_i) = \prod_{i=1}^{M} \mathcal{M}_Q(Y_i; 1, \tau_i), \\
R(\pi) &= \prod_{q,r=1}^{Q} R(\pi_{qr}) = \prod_{q,r=1}^{Q} \mathcal{B}eta(\pi_{qr}; \tilde{\pi}_{qr1}, \tilde{\pi}_{qr2}), \\
R(\gamma) &= \mathcal{D}ir_Q(\gamma; \tilde{\gamma}).
\end{aligned} \tag{11}
$$

Each vector τ_i is of size Q and encodes the (approximate) posterior probabilities for node i to be in each cluster. Given $\tau = (\tau_i)_i$, the set of posterior cluster membership probabilities, for any pair (q, r) the corresponding expected meta-document can be computed as

$$
\tilde{W}_{qr} = \sum_{i \neq j} \tau_{iq} \tau_{jr} W_{ij}. \tag{12}
$$

By construction, the v-th element of vector $\tilde{W}_{qr}$ is the expected pseudo count of word v for all documents sent from nodes in cluster q to nodes in cluster r. Finally, the variational distribution on latent topic proportions is assumed to be:

$$
R(\delta) = \prod_{q,r=1}^{Q} R(\delta_{qr}) = \prod_{q,r=1}^{Q} \mathcal{N}\big(\delta_{qr}; \mu_{qr}(\tau, \nu), \mathrm{diag}(\sigma^2_{qr}(\tau, \nu))\big), \tag{13}
$$

with $(\mu_{qr}(\tau, \nu), \sigma_{qr}(\tau, \nu))^\top = f(\tilde{W}^{norm}_{qr}(\tau); \nu)$ the output of a parametric function, typically a (deep) neural network, with parameters denoted ν. Hereafter, the ETM encoder will be used as the function f parametrised by ν. The normalised expected meta-documents $\tilde{W}^{norm}_{qr}(\tau) = \left(\sum_{v=1}^{V} \tilde{W}^{v}_{qr}(\tau)\right)^{-1} \tilde{W}_{qr}(\tau) \in \mathbb{R}^V$ are then given to the encoder which outputs the mean and variance vectors $(\mu_{qr}(\tau, \nu), \sigma_{qr}(\tau, \nu))^\top$ of the posterior distribution. Our inference strategy is inspired by Dieng et al. (2020) and finds its roots in the original work of Kingma & Welling (2014) for classical data. However, as we shall see, a critical property of our methodology is that the (approximate) posterior allocation probabilities τ will change through the updates and so are the inputs of the encoder. In all experiments we carried out, we used a 3-layer architecture with 800 units for the hidden layers, as originally proposed in Dieng et al. (2020). In order not to increase the number of parameters ν linearly with the number of pairs of groups, amortised inference

is used as advocated in Gershman & Goodman (2014) or Kingma & Welling (2014).

Proposition 4.2. *Using the assumptions describes in Equations* (10), (11) *and* (13), *the ELBO, which is a functional of the variational distribution, reduces to a function of the variational parameters and can be split in two terms associated with the network and with the texts respectively:*

$$\mathscr{L}(R(\cdot);\alpha,\rho) = \mathscr{L}(\tau,\tilde{\pi}_1,\tilde{\pi}_2,\tilde{\gamma},\nu;\alpha,\rho) \tag{14}$$

$$= \mathscr{L}^{net}(\tau,\tilde{\pi}_1,\tilde{\pi}_2,\tilde{\gamma};\alpha,\rho) + \mathscr{L}^{texts}(\tau,\nu;\alpha,\rho), \tag{15}$$

where $\tilde{\pi}_1 = (\tilde{\pi}_{qr1})_{qr}$, $\tilde{\pi}_2 = (\tilde{\pi}_{qr2})_{qr}$.

Proof. The proof and the exact value of the ELBO is detailed in Appendix A □

4.3. Optimisation and Algorithm

We now aim at maximising the ELBO with respect to the variational parameters $\tilde{\pi}, \tilde{\gamma}, \tau$ and ν and to the parameters ρ and α. On the one hand, following Latouche et al. (2012), the variational parameters $\tilde{\pi}$ and $\tilde{\gamma}$ only depend on τ and are updated as follow:

$$\begin{aligned} \tilde{\gamma_q} &= \gamma_{0q} + \sum_{i=1}^{M} \tau_{iq} \\ \tilde{\pi}_{qr1} = \pi^0_{qr1} + \sum_{i\neq j}^{M} \tau_{iq}\tau_{jr}X_{ij}, &\quad \tilde{\pi}_{qr2} = \pi^0_{qr2} + \sum_{i\neq j}^{M} \tau_{iq}\tau_{jr}(1-X_{ij}). \end{aligned} \tag{16}$$

On the other hand, the optimisation of the ETM parameters is done by stochastic gradient descent based on Dieng et al. (2020). Once both parts are done, we only need to update τ using the already up-to-date parameters. To do so, we switch from τ lying on the simplex Δ_{Q-1} to the unconstrained space $\mathbb{R}^{Q-1}$ using for any $i \in \mathcal{V}$ and $q \in \{1,\dots,Q-1\}$:

$$\xi_{iq} = \ln(\tau_{iq}) - \ln(\tau_{iQ}).$$

We then use auto differentiation for the optimisation algorithm with respect to ξ. It is worth emphasising that the ELBO is optimised over the whole set of allocation probability vectors $\tau = (\tau_i)_i$ contrary to STBM which looks for a hard allocation of nodes to clusters, one allocation being optimised at a time, all the others being fixed. Moreover, by optimising the entry of the encoder through τ, thus looking for an optimal allocation

of documents to pairs of clusters, the moves in τ aim at uncovering the optimal direction in the posterior distribution in $(\theta_{qr})_{qr}$ maximising the ELBO. In that regard, ETSBM has links with the quasi branching bound algorithm of Jouvin et al. (2021) for document clustering.

4.4. Model selection

Finally, the selection of the number of cluster Q is performed using the ELBO. It is useful to remind that the aim of the model is to select the number of clusters providing the more meaning. Therefore, relying on Latouche et al. (2012), we take advantage of the Bayesian framework that automatically penalises the complexity of the model with respect to Q. The best number of cluster Q is then selected by estimating the parameters for models with different number of cluster Q and keeping the one with the highest ELBO. Our experiment Section 2 confirms that this procedure provides a relevant model selection criterion. In this paper, the number of topics K is not selected. Indeed, we choose to keep a high K as advocated in Dieng et al. (2020). In practice, once the inference of the topics is done, a classical approach consists in focusing the interpretation on the results associated with the most frequent topics. As we shall see, in the experiment section, provided that the value of K chosen is large enough, the proposed procedure provides an accurate estimate of Q.

5. Numerical experiments

In this section, a series of experiments is presented to assess the proposed methodology. First, three scenarii used for benchmarking are described. Second, an illustration of the results provided by ETSBM on a simulated dataset from one of the scenarii is given. Then, results from experiments to evaluate the model selection criterion on the three scenarii considered are brought. Moreover, various strategies to initialise ETSBM are compared. Finally, an extensive set of experiments on the three scenarii with three levels of difficulty is carried out to evaluate the clustering performances of ETSBM against competitive algorithms.

5.1. Simulation setup

The networks with textual edges are generated following three scenarii A, B, C, as originally introduced in Bouveyron et al. (2018).

Sampling networks with textual edges.

- Scenario A is composed of three communities, each defining a cluster, and four topics. By definition, a community is defined such that more connections are present between nodes of the same community. For each cluster, a specific topic is employed to sample all the documents associated with the corresponding intra-cluster connections. Besides, an extra topic is considered to model documents exchanged between nodes from different clusters. Thus, by construction, the clustering structure can be retrieved either using the network or the texts only.

- Scenario B is made of a single community and three topics. Thus, all nodes connect with the same probability. Then, the community is split into two clusters with their respective topics. An extra topic is used to model documents exchanged between the two clusters. Therefore, in such a scenario, the network itself is not sufficient to find the two clusters but the documents are.

- Scenario C is composed of three communities and three topics. Two of the communities are associated with their respective topics, say t_1 and t_2. Moreover, following the previous scenario, the third community is split in two clusters, one being associated with topic t_1 and the other with t_2. Thus, considering both texts and topology, each network is actually made of four node clusters. Fundamentally, both textual data and the network itself are necessary to uncover the clusters. This scenario will be of major interest in this experiment section since it allows to ensure that the two sources of information are correctly used to retrieve partitions.

The edges holding the documents are constructed by sampling words from four BBC articles, focusing each on a given topic. The first topic deals with the UK monarchy, the second with cancer treatments, and the third with the political landscape in the UK. The last topic deals with astronomy. In the general setting, for all scenarii, the average text length for the documents is set to 150 words. The parameters used to sample data from the three scenarii are given in Table 1. Moreover, three examples of networks generated from A, B and C are presented in Figure 3.

Clustering performance evaluation. The main criterion used in the following to evaluate the clustering performances of the different strategies is the adjusted random index (ARI).

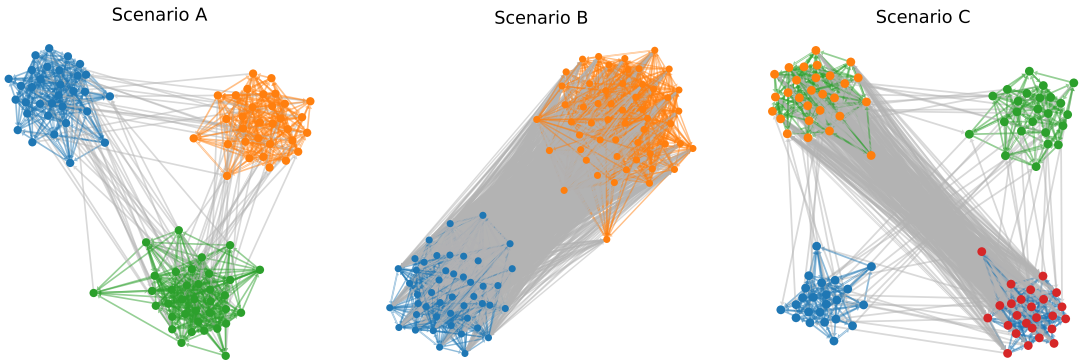


Figure 3: An example of each scenario is presented. The node colours denote the cluster memberships and the edge colours denote the major topic within the text. The Scenarii A, B and C are composed of 3, 1 and 3 communities respectively.

ARI measures how close two partitions are from one another. The closer ARI is to 1, the better the results are. A random cluster assignment leads to an ARI of 0, while a perfect retrieval of the cluster memberships gives an ARI of 1.

	Scenario A	Scenario B	Scenario C
Q (clusters)	3	2	4
K (topics)	4	3	3
Communities	3	1	3
π_{qr} (connection probabilities) $\eta = 0.25$, $\epsilon = 0.01$	$\begin{pmatrix} \eta & \epsilon & \epsilon \\ \epsilon & \eta & \epsilon \\ \epsilon & \epsilon & \eta \end{pmatrix}$	$\begin{pmatrix} \eta & \eta \\ \eta & \eta \end{pmatrix}$	$\begin{pmatrix} \eta & \epsilon & \epsilon & \epsilon \\ \epsilon & \eta & \epsilon & \epsilon \\ \epsilon & \epsilon & \eta & \eta \\ \epsilon & \epsilon & \eta & \eta \end{pmatrix}$
Topics between pairs of clusters (q, r)	$\begin{pmatrix} t_1 & t_4 & t_4 \\ t_4 & t_2 & t_4 \\ t_4 & t_4 & t_3 \end{pmatrix}$	$\begin{pmatrix} t_1 & t_3 \\ t_3 & t_2 \end{pmatrix}$	$\begin{pmatrix} t_1 & t_3 & t_3 & t_3 \\ t_3 & t_2 & t_3 & t_3 \\ t_3 & t_3 & t_1 & t_3 \\ t_3 & t_3 & t_3 & t_2 \end{pmatrix}$
Sufficient information to uncover the clusters	Network	Topics	Network & Topics

Table 1: Detail of the three simulation scenarii to evaluate our model.

Different levels of difficulties. To evaluate ETSBM against state of the art STBM in Sections 5.3 and 5.5, two levels of difficulty are introduced. The first one, named *Hard 1*, makes it particularly hard to distinguish connectivity patterns by using an intra-cluster connectivity probability of 0.2. In Table 1, it corresponds to $\epsilon = 0.2$ instead of 0.01. The

second one, named *Hard 2*, introduces difficulty on the text part by using smaller texts of 110 words on average instead of 150 and by adding noise. In our case, this translates into fixing:

$$\theta_{qr} = (1-\zeta)\theta^{\star}_{qr} + \zeta * \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)^{\top}, \tag{17}$$

with $\zeta = 0.7$. Thus, for each pair of clusters (q, r), the texts are sampled according to a mixture between a multinomial distribution with probability 1 on the corresponding topic and a uniform distribution over all topics considered. Finally, the intra-cluster connection probability is decreased from 0.2 to $\eta = 0.1$.

5.2. An introductory example

A first glimpse at the ETSBM results on a single network simulated with Scenario C is presented here. In Figure 4, the evolution of the ELBO and ARI values are monitored at each iteration of the inference of ETSM applied on this single simulated network.

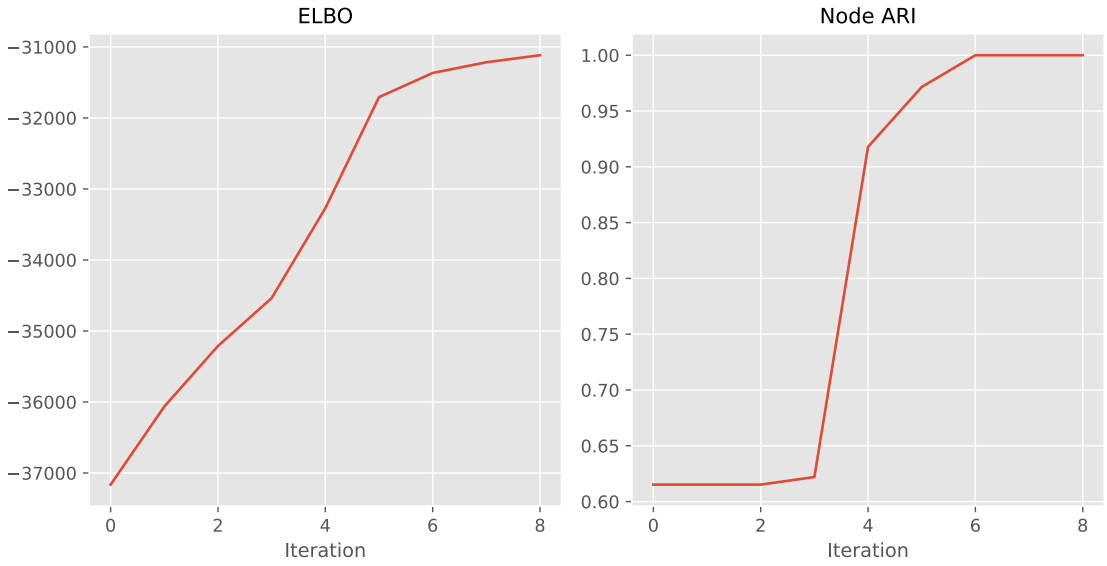


Figure 4: Evolution of ETSBM ELBO and ARI (y-axis) at each iteration (x-axis) on the Scenario C after an initialisation with the K-means algorithm.

As we can see, both the ELBO and the ARI increase after each iteration. In particular, starting from the clustering initialisation with an ARI value of 0.62, the algorithm converges to a value of 1, characterising a perfect cluster recovery. This figure illustrates the ability of the methodology proposed to retrieve the true node partition, by combining the textual and network data.

In addition, Figure 5 provides representations for the expected posterior estimates $\hat{\pi}$ and $\hat{\gamma}$ computed as follows $\hat{\pi}_{qr} = \tilde{\pi}_{qr1}/(\tilde{\pi}_{qr1}+\tilde{\pi}_{qr2})$ and $\hat{\gamma}_q = \tilde{\gamma}_q/(\sum_{r=1}^{Q}\tilde{\gamma}_r)$. We emphasise that the matrix characterises the connexion probabilities between clusters with a 10^{-2} rounding. It matches the expected connectivity structure described in Table 1.

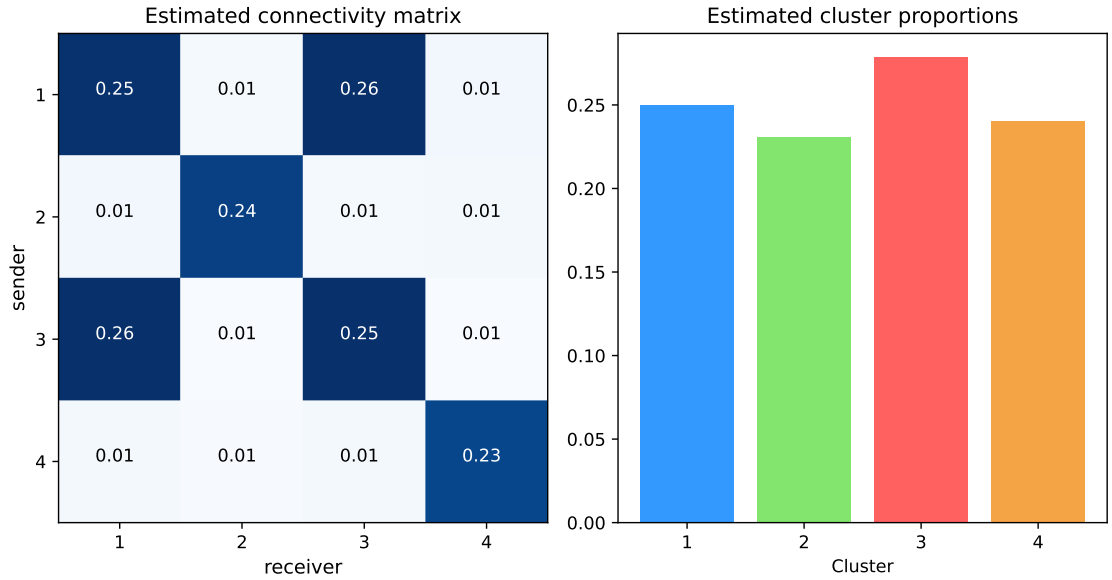


Figure 5: On the left hand side, the expected posterior estimate of the connectivity matrix π provided by ETSBM. On the right hand side, the expected posterior estimate of the cluster proportions γ. The graph was generated following Scenario C.

Eventually, the topics learnt as well as the clustering results on the network are presented in Figure 6. In the network representation, the node colours correspond to the cluster memberships while the edge colours indicate the most used topic in the corresponding documents. Moreover, for each topic t_k with $k \in \{1, 2, 3\}$, the 10 words with the highest probabilities, according to the corresponding topic vector β_k, are displayed. The three topics presented are well-separated and can be identified as the topics dealing respectively with astronomy, the political landscape in the UK, and the UK monarchy, as expected. In addition, four node clusters have been retrieved and the edge topics, or colours, match the description of the Scenario C setup. To conclude, ETSBM successfully render both the network topology and the edge topics.

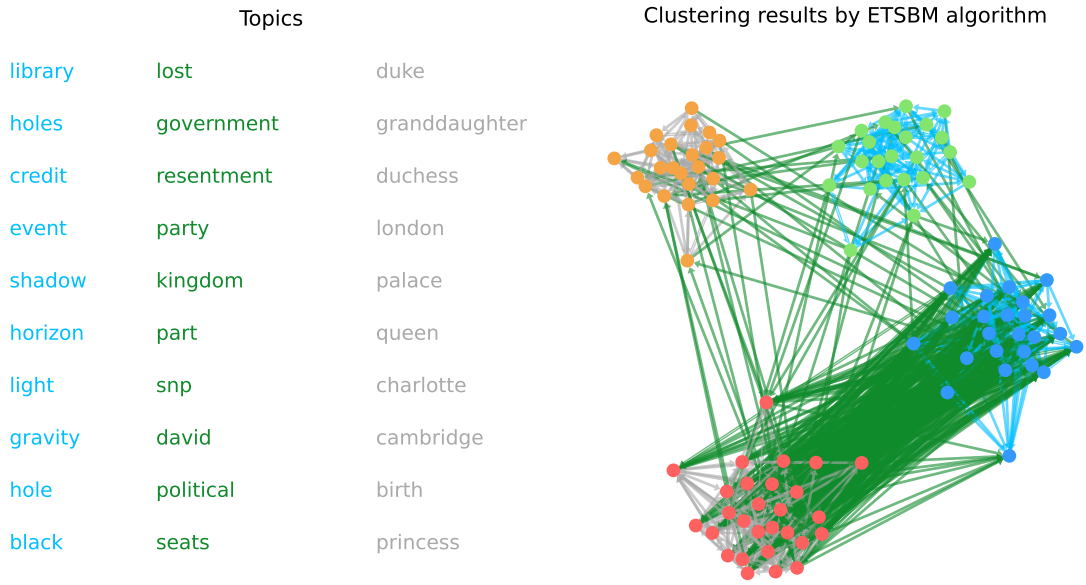


Figure 6: On the left hand side, the top 10 words of each topic according to ETSBM results. Thus, for each topic t_k with $k \in \{1, 2, 3\}$, the 10 words with the highest probability values, according to the corresponding topic vector β_k, are displayed. On the right hand side, ETSBM clustering result is illustrated. The node colours indicate the node clusters while the edge colours correspond to the most used topic within the document.

Finally, Figure 7 provides a high level representation of the results. On the one hand, the "meta-nodes" represent ETSBM clusters and their size is proportional to the number of nodes assigned to the corresponding clusters. Moreover, the "meta-node" colours are consistent with the colours in Figure 6. On the other hand, the edges represent the meta-documents. We recall that they correspond to the expected posterior estimate of a document for a given pair of clusters. The edge colours correspond to the most used topic within the meta-document. The edge widths are determined by the number of connexions between the corresponding clusters. This figure underlines ETSBM capability to produce intelligible and accurate data summary. We emphasise that graphs with thousands of edges, that sometimes cannot be represented because of memory issues, are here able to be summarised in easy-to-read meta-graphs.

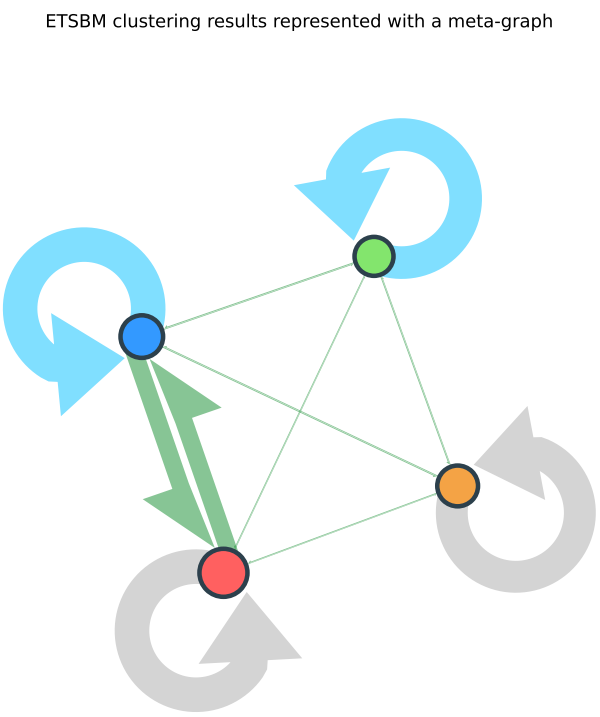


Figure 7: Meta representation of ETSBM results. On the one hand, the clusters are represented by the node colours, the node widths are proportional to the expected posterior estimate of the cluster proportions, and their colours correspond to the same cluster colours as in the network in Figure 6. On the other hand, the edges are coloured as the most used topic within the meta-document and the widths are proportional to the expected posterior estimate of the number of documents exchanged between the two.

To conclude, this introductory example showed the ETSBM capacity to render meaningful summaries by combining both network and text information. It is worth reminding that, since it comes from Scenario C, those results could not have been retrieved with models handling only network or texts as SBM, LDA or ETM.

5.3. Effect of the initialisation

This experiment aims to evaluate the impact of the initialisation on the final performance of our methodology. The networks are generated according to the *Hard 2* difficulty, to easily visualise the differences between the tested configurations. Moreover, the experiment is performed on Scenario C to ensure both the network and textual data are used. Three different initialisations are compared: clusters may be randomly assigned to the nodes (random), or initial clusters can be determined by a K-Means algorithm fitted on the adjacency matrix A. Finally, the dissimilarity procedure proposed in Bouveyron et al. (2018) is evaluated as the last initialisation strategy (dissimilarity). It uses both

network and textual information to build a similarity matrix based on the topics discussed between nodes. Then, a K-means algorithm is performed on this similarity matrix to find a cluster allocation for each node. This initialisation strategy requires to provide the topic proportion of each edge. Thus, ETM is trained on the texts and the estimated topic proportions $(\theta_{ij})_{(i,j)\in\mathcal{E}}$ are used for the dissimilarity initialisation. Figure 8 presents the ARI results with, for each initialisation strategy, a boxplot of the raw initialisation and of ETSBM clustering.

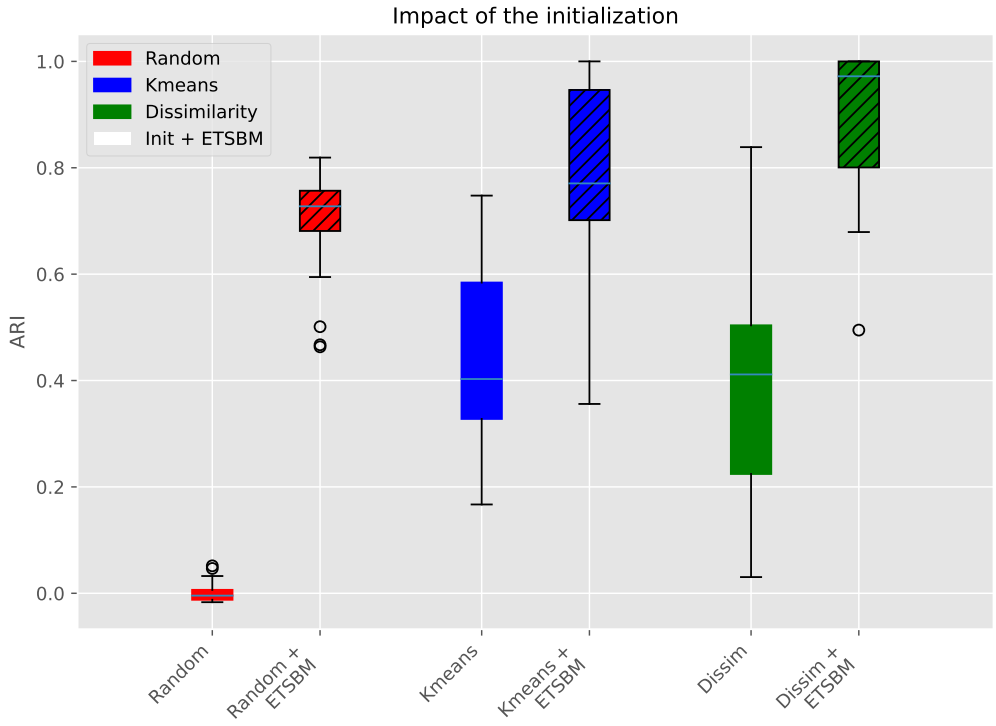


Figure 8: This figure displays the boxplots of the initialisation ARI (the boxplot without stripe) and of ETSBM clustering ARI with the same initialisation (the boxplot with stripes). This experiment was performed on 50 networks generated following Scenario C in the *Hard 2* setting.

While the random initialisation is close to 0 for ARI, both the K-means and the dissimilarity initialisation fluctuates in terms of ARI, with no clear advantage for one of the two strategies. However, ETSBM provides much better results with the dissimilarity initialisation than with K-means. It is also worth noticing that the gap between the random and K-Means initialisations has largely been closed by ETSBM algorithm. One possibility is that the model suffers the same flaws as SBM, which is for the ELBO to fall into local minimum. It is possible that the use of texts in the dissimilarity limits this effect. Therefore, we will only use the dissimilarity initialisation in the rest of the paper

Q / K	Scenario A					Scenario B					Scenario C				
	2	**3**	4	5	10	**2**	3	4	5	10	2	3	**4**	5	10
2	0	**94**	6	0	0	**74**	24	2	0	0	0	0	**92**	8	0
3	0	**90**	10	0	0	**78**	18	4	0	0	0	0	**90**	10	0
4	0	**78**	20	2	0	**76**	20	4	0	0	0	0	**94**	6	0
5	0	**86**	14	0	0	**68**	28	4	0	0	0	0	**84**	16	0
10	0	**88**	10	2	0	**82**	18	0	0	0	0	0	**86**	14	0

Table 2: This table presents the percentage of time a number of clusters have been selected on 50 simulated networks. The experiment is repeated for different values of K, and for Scenario A, B and C. For instance, in Scenario A with $K = 3$, the model with $Q = 3$ clusters was selected in 90% of cases.

as it provides the best results in most cases.

5.4. Model selection

This experiment aims to assess the efficiency of the model selection criterion, presented in Section 4.4. Let us remind that we do not aim at selecting the number of topics K since it is handled afterwards. As a consequence, the model selection criterion is evaluated for different values of K to ensure that the performances remain high, in all cases. For each scenario, 50 networks are sampled following the setup described in Section 5.1. For each network, ETSBM parameters are estimated taking the best initialisation out of 10. Table 2 presents the percentage of time a number Q is selected using the strategy proposed in Section 4.4 over the 50 networks, for each K value. It is worth noticing that the right model is selected more than 75% of the time, except for the Scenario B with $K = 5$, slightly bellow with 68%. In addition, as advocated before, for $K = 10$, the right model is selected more than 80% of the time in each scenario. This experiment illustrates the capacity of the model selection criterion to retrieve the number of clusters. Moreover, keeping a high value of K is confirmed to be compatible with an efficient cluster number selection.

5.5. Benchmark study

To end this section, ETSBM is evaluated against state of the art clustering algorithms for STBM, and SBM on the three levels of difficulty presented in Section 5.1. Results for LDA as well as ETM for text clustering are also provided. For each level of difficulty and each scenario, Table 3 displays the mean and the standard deviation of the ARI values obtained over 50 graphs. Both the node and edge clusters ARI are provided but

Table 3: Benchmark of our model against STBM, SBM and LDA. When a model does not provide an information, a line is displayed instead of the result. For instance, SBM does not provides edge information.

		Scenario A		Scenario B		Scenario C	
		Node ARI	Edge ARI	Node ARI	Edge ARI	Node ARI	Edge ARI
Easy	ETSBM	**1.00 ± 0.00**	**0.99 ± 0.03**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	STBM	0.98 ± 0.04	0.98 ± 0.04	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	SBM	1.00 ± 0.00	——	0.01 ± 0.01	——	0.69 ± 0.07	——
	LDA	——	0.97 ± 0.06	——	1.00 ± 0.00	——	1.00 ± 0.00
	ETM	——	0.96 ± 0.14	——	1.00 ± 0.00	——	1.00 ± 0.00
Hard 1	ETSBM	**1.00 ± 0.00**	**0.95 ± 0.03**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	0.97 ± 0.04
	STBM	**1.00 ± 0.00**	0.90 ± 0.13	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.98 ± 0.03**
	SBM	0.01 ± 0.01	——	0.01 ± 0.01	——	0.01 ± 0.01	——
	LDA	——	0.90 ± 0.17	——	1.00 ± 0.00	——	0.99 ± 0.01
	ETM	——	0.93 ± 0.07	——	1.00 ± 0.00	——	0.98 ± 0.03
Hard 2	ETSBM	**0.98 ± 0.06**	**0.83 ± 0.07**	**1.00 ± 0.00**	0.86 ± 0.03	**0.91 ± 0.12**	**0.84 ± 0.12**
	STBM	0.75 ± 0.27	0.82 ± 0.22	**1.00 ± 0.00**	**1.00 ± 0.00**	0.63 ± 0.19	0.77 ± 0.15
	SBM	0.96 ± 0.05	——	0.00 ± 0.00	——	0.63 ± 0.11	——
	LDA	——	0.77 ± 0.09	——	0.88 ± 0.02	——	0.84 ± 0.04
	ETM	——	0.83 ± 0.08	——	0.85 ± 0.03	——	0.86 ± 0.04

we recall that the main interest of this work concerns the node clustering performances. In the *Easy* and *Hard 1* settings, the ARI is always 1, which indicates that the true partitions are successfully retrieved by ETSBM and STBM. On the contrarty, SBM is not able to distinguish clusters in Scenario B since all nodes connect one another with the same probability. Identically, in Scenario C, SBM alone cannot differentiate the nodes highly connected but discussing of different topics. For instance, in the *Easy* case, this translates into an ARI of 0.01 and 0.69 respectively. In the *Hard 2* setting, ETSBM node clustering significantly outperforms STBM. In particular in Scenario C, *Hard 2*, ETSBM results reach an ARI of 0.91 against 0.63 for STBM. Even though it is not the main focus of this work, the edge ARI is always higher than 0.84, which is satisfactory, and is competitive when not higher than STBM. This significant gaps in the noisy settings highlight ETSBM clustering improvement upon STBM. To conclude, our experiments strongly indicates that ETSBM node clustering performances are either the same or significantly better than STBM.

6. Real World example: analysing a the French presidential election with a Twitter dataset

6.1. Context

This section presents a use case on a Twitter dataset dealing with the French presidential election of 2022. The election resulted in Emmanuel Macron being re-elected as President of France. The objective is to use ETSBM to capture the global trends on Twitter before the first round of the French presidential election in April 2022. The network has been constructed using tweets collected by the Linkfluence, a Meltwater company, during a collaboration between journalists of the French newspaper *Le Monde* and two authors of this article (Laurent, 2022). Newspapers such as *Le Monde* may be interested in having a good understanding of the global dynamics on social media during an electoral period, in order to understand the interest of the public opinion. Thus, interpretable topics and meaningful clusters may help them getting a grasp on the core factors interesting the elector. During the last 50 years, French political landscape has been split between two main parties, the left-democrat, mainly represented by the socialist party, and the right-liberal, represented by *Les Républicains* (formerly UMP). A shift occurred in 2017 when a three-way split between the far-left political families, the centrists, or liberals, and the far-right emerged. This analysis aims at capturing the major topics discussed prior to the election. In addition, we want to understand the way those topics shape user groups interactions. However, this study does not aim at making any form of prediction about the election.

6.2. Dataset construction and method

In the collected data, each node represent a Twitter account. An account i is connected to j if the former retweeted the later or if i "mentioned" j with an "@account_name" in a tweet. The text on the edges are the tweet themselves. Our database has been created by saving any tweet talking about one of the twelve candidates. If several tweets appear from i to j, the edge (i, j) holds all those tweets stack together. We only keep edges with text length greater than 100 characters. Then, a lemmatisation procedure is used to reduce the vocabulary size. The "stopwords", defined as non-informative words such as "and" or "it", are withdrawn, as well as numeric characters and words with a length inferior to 3 characters. In the end, we keep the largest connected component of

this graph. Our dataset holds $2,730$ nodes and $403,768$ edges. This means that the graph is sparse at 94.58%. We emphasise that this level of sparsity is quite high and makes the data analysis particularly challenging. The number of topics is set to $K = 20$. Also, for each Q value, the model is trained for 10 different initialisations and the best result among those 10, ELBO wise, is kept. Then, the number of clusters is selected using our model selection criterion. Figure 9 shows that the most appropriate model according to our criterion corresponds to a number of clusters $Q = 5$.

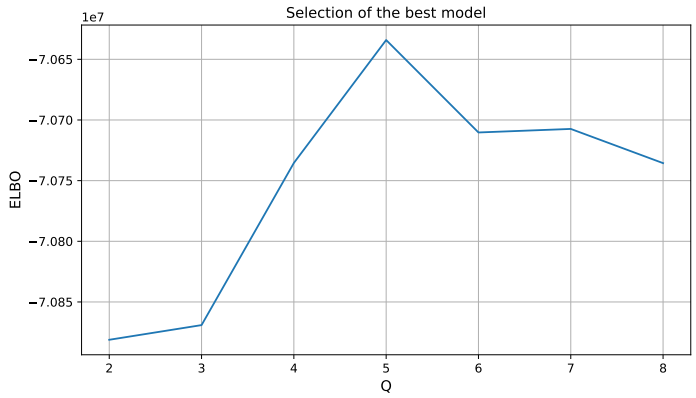
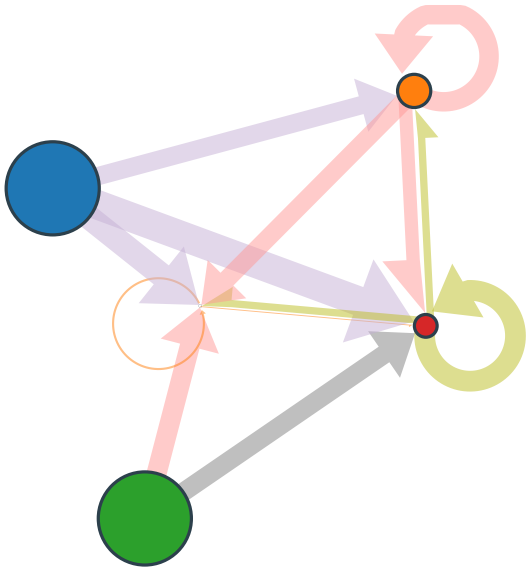


Figure 9: After running ETSBM with different number of clusters Q, the ELBO suggest to keep five clusters.

6.3. Results

The meta-graph presented in Figure 10a is a high-level representation of the network. The "meta-nodes" correspond to ETSBM clusters and the edges to the meta-documents as defined in Equation (12). A translation of the top words is provided in Appendix B.11. It is interesting to note the two types of clusters uncovered. The first one is composed of central accounts such as French politicians and their communication teams, for instance *Jean-Luc Mélenchon*, *Guillaume Peltier*, *En Marche #avecvous*, *les Républicains* or *Eléonore Lhéritier*. Some popular French media such as *BFMTV*, *Le Figaro*, *Valeurs actuelles*, *franceinfo* are also in this cluster. On average, the accounts in this cluster have been retweeted or mentioned 299 times against 12 times for the whole network. This cluster does not correspond to a political trend but to accounts with a high level of interactions with the rest of the graph. Despite the small size of this cluster, composed of 25 nodes, ETSBM is able to detect it and to render its central function as a relay of information to other parts of the graph. This is stressed by the main topic discussed within the

cluster. It regards the election as a democratic process: “round”, “vote”, “power”, “president”, “first” which we assume stands for “first round”. This core cluster is retweeted differently by the four clusters which on the contrary hold clear political trends. The orange and green clusters are interested in *Jean-Luc Mélenchon* (pink arrow) and left parties in general (grey arrow) but they seem to differ in terms of function. The orange cluster clearly relays information about *Jean-Luc Mélenchon* and is interacting with the red cluster, interested in *Eric Zemmour*, or by the blue cluster, interested in *Emmanuel Macron*. On the contrary, the green cluster seems to only relegate contents without being retweeted. Eventually, a cluster (the red one), interested in Eric Zemmour, appears to relegate few contents from the central accounts but creates and shares many of its own content. This dynamic differs to the blue cluster interested in *Emmanuel Macron*, which mainly retransmits informations without many self interactions. To conclude, the three-way split of the French political landscape is rightfully captured. ETSBM is also able to detect subtleties such as a split within the left-wing, with the orange cluster interested only in *Jean-Luc Mélenchon* and the biggest one exchanging about different left-political front runners, *Jean-Luc Mélenchon*, *Yannick Jadot*, *Fabien Roussel* and *Anne Hidalgo*. ETSBM combines the connection information, for instance all clusters are connected to the small one, and the topics information, for instance the green and orange should be separated, to provide relevant insights about the information organisation within the social network. This level of detail is promising and highlights how ETSBM gives a better comprehension of the complex dataset at our disposal.



(a) ETSBM allows to make high-level representation. Each node corresponds to a cluster and the edge colours indicate which topics is discussed between the clusters. The *blue cluster* reacts or retweets mainly about *Emmanuel Macron*, the French president, the *orange cluster* mainly reacts or retweets about *Jean-Luc Mélenchon*, a far-left candidate, the reactions or retweet of the *red cluster* mainly concerns *Eric Zemmour*, a far-right candidate, while the *green cluster* reacts or retweets information either about *Jean-Luc Mélenchon* or more generally about the left-wing candidates (*Yannick Jadot, Jean-Luc Mélenchon, Fabien Roussel* or *Anne Hidalgo*).

Topics

tour	heure	macron	melenchon	zemmour	faire
tout	melenchonvagagner	candidat	jadot	eric	aller
faire	monde	emmanuel	jlm	jevotezemmour	non
aller	erepublique	campagne	roussel	soutenir	dire
voter	melenchon	zemmour	voter	jevotezemmourle	savoir
pouvoir	meeting	presidentielle	gauche	hdelareconquete	bon
vote	unionpopulaire	journaliste	vote	zemmourpresident	comme
president	programme	debat	tour	partager	quand
merci	marchepourla	direct	hidalgo	zemmourvsmacron	plus
premier	melenchontf	via	droite	maintenant	tout

(b) The most important words of the topics present in the meta-graph. A translation is provided in Figure B.11.

Figure 10: ETSBM results on the Twitter dataset for $Q = 5$ clusters. The node and edge sizes are proportional to the cluster size and the number of documents within the meta-documents respectively.

7. Conclusion and discussion

The embedded topics for the stochastic block model (ETSBM) is well suited to simultaneously find meaningful node and edge clusters. In addition, ETSBM provides an intelligible high-level representation of the graph. It can be used both on directed and undirected graphs and is suited for large datasets thanks to the variational inference. The numerical experiments showed that the ELBO is a relevant model selection criterion to estimate the number of node clusters Q in this Bayesian framework. Moreover, this criterion keeps provide a good estimate of Q for a high number of topics K. In the end, a use case on a Twitter dataset proved the usefulness of the method. ETSBM clustering results were both meaningful and humanly intelligible. Further work may be directed in the study of theoretical foundations of the model selection criterion proposed. Adding temporal information concerning the connectivity patterns and the topics modelling could also contribute to obtain useful information on the data.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, .

Attias, H. (1999). A variational baysian framework for graphical models. *Advances in neural information processing systems*, *12*.

Bergé, L. R., Bouveyron, C., Corneli, M., & Latouche, P. (2019). The latent topic block model for the co-clustering of textual interaction data. *Computational Statistics & Data Analysis*, *137*, 247–270.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bouveyron, C., Latouche, P., & Zreik, R. (2018). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, *28*, 11–31.

Côme, E., & Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, *15*, 564–589. doi:10.1177/1471082X15577017.

Corneli, M., Bouveyron, C., Latouche, P., & Rossi, F. (2019). The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, *29*, 677–695.

Corneli, M., Latouche, P., & Rossi, F. (2016). Block modelling in dynamic networks with non-homogeneous poisson processes and exact icl. *Social Network Analysis and Mining*, *6*, 1–14.

Daudin, J.-J., Picard, F., & Robin, S. (2006). *A mixture model for random graphs*. Research Report RR-5840 INRIA.

Daudin, J.-J., Picard, F., & Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, *18*, 173–183.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*, 391–407.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Erdos, P., Rényi, A. et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*, 17–60.

Fienberg, S. E., & Wasserman, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociological methodology*, *12*, 156–192.

Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*. volume 36.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M. et al. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, *2*, 129–233.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*, 301–354.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *UAI*.

Jernite, Y., Latouche, P., Bouveyron, C., Rivera, P., Jegou, L., & Lamassé, S. (2014). The random subgraph model for the analysis of an ecclesiastical network in merovingian gaul. *The Annals of Applied Statistics*, *8*, 377–405.

Jouvin, N., Latouche, P., Bouveyron, C., Bataillon, G., & Livartowski, A. (2021). Greedy clustering of count data through a mixture of multinomial pca. *Computational Statistics*, *36*, 1–33.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI* (p. 5). volume 3.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. `arXiv:1312.6114`.

Latouche, P., Birmelé, E., & Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, (pp. 309–336).

Latouche, P., Birmele, E., & Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, *12*, 93–115.

Laurent, S. (2022). Comment la gauche sociale-démocrate a perdu la bataille des réseaux sociaux. *Le Monde*, . URL: `https://www.lemonde.fr/politique/article/2022/03/31/comment-la-gauche-sociale-democrate-a-perdu-la-bataille-des-reseaux-sociaux_6119986_823448.html`.

Lee, C., & Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, *4*, 1–50.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009). Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning* (pp. 665–672).

Mariadassou, M., Robin, S., & Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, *4*, 715–742.

Matias, C., & Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1119–1141.

Matias, C., & Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, *47*, 55–74.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, .

Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American statistical association*, *96*, 1077–1087.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., & Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. (pp. 159–168). ACM press.

Pathak, N., Delong, C., Erickson, K., & Banerjee, A. (2008). Social topic models for community extraction.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning* (pp. 1278–1286). PMLR.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* UAI '04 (p. 487–494). AUAI Press.

Sachan, M., Contractor, D., Faruquie, T. A., & Subramaniam, L. V. (2012). Using content and interactions for discovering communities in social networks. In *Proceedings of the 21st international conference on World Wide Web* (pp. 331–340).

Sampson, S. F. (1969). *Crisis in a cloister*. Ph.D. thesis Ph. D. Thesis. Cornell University, Ithaca.

Srivastava, A., & Sutton, C. (2017). Autoencoding variational inference for topic models. In *ICLR*.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wang, Y. J., & Wong, G. Y. C. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, *82*, 8–19.

Zanghi, H., Volant, S., & Ambroise, C. (2010). Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, *31*, 830–836. doi:10.1016/j.patrec.2010.01.026.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., & Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web* (pp. 173–182).

Zreik, R., Latouche, P., & Bouveyron, C. (2017). The dynamic random subgraph model for the clustering of evolving networks. *Computational Statistics*, *32*, 501–533.

Appendix A. Inference

Proof of Proposition 4.1. The ELBO can be decomposed as follow:

$$\begin{aligned}
\log p(A, W \mid \alpha, \rho) &= \mathbb{E}_R \left[\log p(A, W \mid \alpha, \rho)\right] \\
&= \mathbb{E}_R \left[\log \frac{p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho)}{p(Y, \pi, \gamma, \delta \mid A, W, \alpha, \rho)}\right] && \text{applying Bayes rule} \\
&= \mathbb{E}_R \left[\log \frac{p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho)}{R(Y, \pi, \gamma, \delta)} + \log \frac{R(Y, \pi, \gamma, \delta)}{p(Y, \pi, \gamma, \delta \mid A, W, \alpha, \rho)}\right] \\
&= \mathscr{L}(R(\cdot); \alpha, \rho) + \mathrm{KL}(R(\cdot) || p(Y, \pi, \gamma, \delta \mid A, W, \alpha, \rho)).
\end{aligned}$$

□

Proof of Proposition 4.2.

$$\begin{aligned}
\mathscr{L}(R(\cdot); \alpha, \rho) &= \overbrace{\mathbb{E}_R \left[\log \frac{p(W \mid Y, A, \theta, \alpha, \rho) p(\theta)}{R(\theta)}\right]}^{\mathscr{L}^{net}(\tau, \tilde{\pi}_{qr1}, \tilde{\pi}_{qr2} \tilde{\gamma}; \alpha, \rho) :=} + \overbrace{\mathbb{E}_R \left[\log \frac{p(A \mid Y, \pi) p(Y \mid \gamma) p(\pi) p(\gamma)}{R(Y) R(\pi) R(\gamma)}\right]}^{\mathscr{L}^{texts}(\tau, \nu; \alpha, \rho) :=} \\
&= \mathbb{E}_R \left[\log p(W \mid Y, A, \theta, \alpha, \rho)\right] + \mathbb{E}_R \left[\log p(\theta)\right] - \mathbb{E}_R \left[\log R(\theta)\right] \\
&+ \mathbb{E}_R \left[\log p(A \mid Y, \pi)\right] + \mathbb{E}_R \left[\log p(Y \mid \gamma)\right] + \mathbb{E}_R \left[\log p(\pi)\right] + \mathbb{E}_R \left[\log p(\gamma)\right] \\
&- \mathbb{E}_R \left[\log R(Y)\right] - \mathbb{E}_R \left[\log R(\pi)\right] - \mathbb{E}_R \left[\log R(\gamma)\right] \\
&= \sum_{i \neq j}^{M} \sum_{q,r}^{Q} A_{ij} \tau_{iq} \tau_{jr} \mathbb{E}_R \left[\underbrace{\log p(w_{ij} \mid \delta_{qr}, \alpha, \rho)}_{T_{ij}^{\delta_{qr}}}\right] - \sum_{q,r} \mathrm{KL}(\mathcal{N}(\mu_{qr}(\tau, \nu), \sigma_{qr}(\tau, \nu)) || \mathcal{N}(0, I)) \\
&+ \sum_{i \neq j}^{M} \sum_{q,r}^{Q} \tau_{iq} \tau_{jr} A_{ij} \left(\psi(\kappa_{qr1}) - \psi(\kappa_{qr2})\right) + \sum_{i \neq j}^{M} \sum_{q,r}^{Q} \tau_{iq} \tau_{jr} (\psi(\kappa_{qr2}) - \psi(\kappa_{qr1} + \kappa_{qr2})) \\
&+ \sum_{i=1}^{M} \sum_{q=1}^{Q} \tau_{iq} \left(\psi(\gamma_q) - \psi\left(\sum_q \gamma_q\right)\right) + \log \mathcal{B}(1_Q) + \log(\mathcal{B}(a, b)) \\
&- \sum_{i=1}^{M} \sum_{q=1}^{Q} \tau_{iq} \log(\tau_{iq}) - \sum_{q,r} \log \mathcal{B}(\kappa_{qr1}, \kappa_{qr2}) - \log \mathcal{B}(\gamma). \qquad \text{(A.1)}
\end{aligned}$$

where,

$$T_{ij}^{\delta_{qr}} = \sum_{d=1}^{D_{ij}} \sum_{n=1}^{N_{id}^d} \sum_{v=1}^{V} w_{ij}^{dnv} \log \left(\sum_{k=1}^{K} \theta_{qrk} \beta_{kv}\right). \qquad \text{(A.2)}$$

and $\theta_{qr} = \mu_{qr}(\tau, \nu) + \sigma_{qr}(\tau, \nu)\epsilon$, $\epsilon \sim \mathcal{N}(0_K, \mathrm{I}_K)$.

The Kullback-Leibler divergence between two Gaussian variables has a close form and is easy to compute. All the terms can be computed except for the expectation of $T_{ij}^{\delta_{qr}}$ that can be approximated using a Monte-Carlo estimator, by drawing S samples for each pair (q, r), such that:

$$\epsilon^s \sim \mathcal{N}(0, I_K), \quad \delta_{qr}^s = \mu_{qr}(\tau, \nu) + \sigma_{qr}(\tau, \nu) \odot \epsilon^s, \quad \theta_{qr}^s = \text{softmax}(\delta_{qr}^s).$$

with $\odot$ denoting the Hadamard product. Thus, for each pair of nodes (i, j) and pair of clusters (q, r), the estimate is given by:

$$\hat{T}_{ij}^{qr} = S^{-1} \sum_{s=1}^{S} T_{ij}^{\delta_{qr}^s}.$$

Plugging $\hat{T}_{ij}^{qr}$ in the Equation (A.1) gives the final estimator of the ELBO.

□

Appendix B. Real data

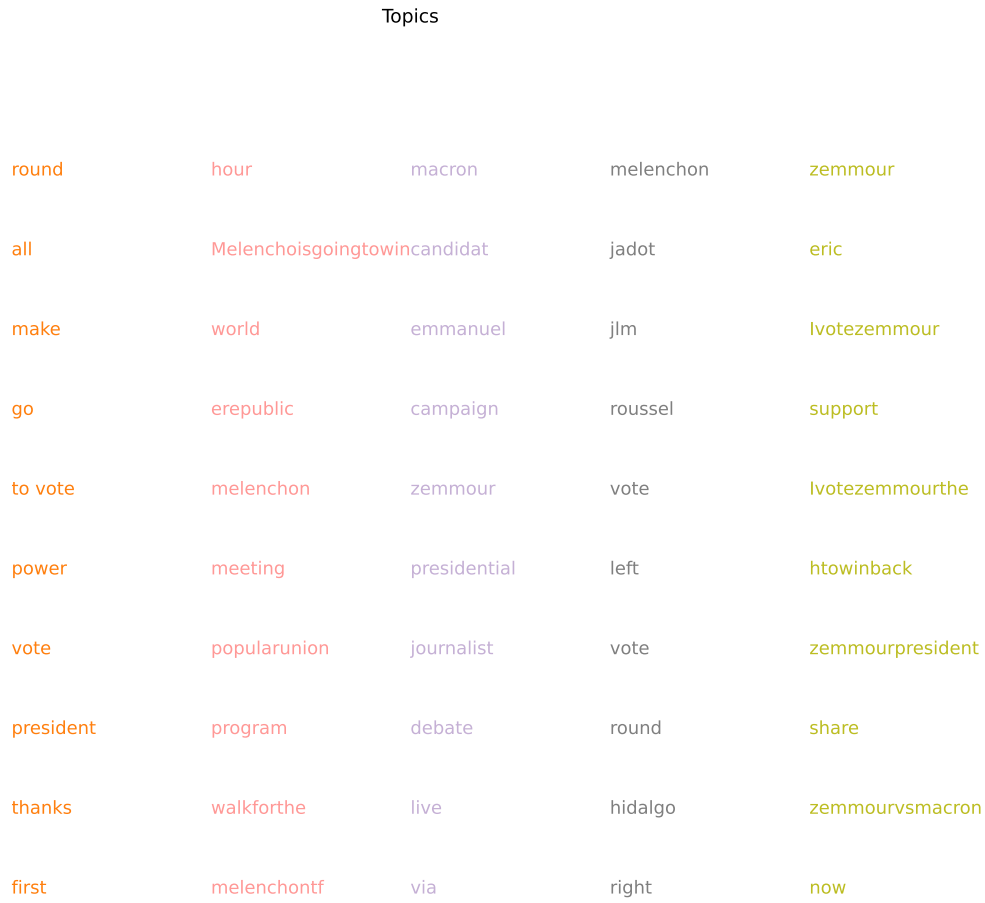


Figure B.11: The most important words of each topic present in the meta-graph translated in English.