Supplementary Material for Dynamic Mixed Membership Stochastic Block Model for Weighted Labeled Networks

Anonymous Author(s)

ABSTRACT

Most real-world networks evolve over time. Existing literature proposes models for dynamic networks that are either unlabeled or assumed to have a single membership structure. On the other hand, a new family of Mixed Membership Stochastic Block Models (MMSBM) allows to model static labeled networks under the assumption of mixed-membership clustering. In this work, we propose to extend this later class of models to infer dynamic labeled networks under a mixed membership assumption. Our approach takes the form of a temporal prior on the model's parameters. It relies on the single assumption that dynamics are not abrupt. We show that our method significantly differs from existing approaches, and allows to model more complex systems −dynamic labeled networks. We demonstrate the robustness of our method with several experiments on both synthetic and real-world datasets. A key interest of our approach is that it needs very few training data to yield good results. The performance gain under challenging conditions broadens the variety of possible applications of automated learning tools −as in social sciences, which comprise many fields where small datasets are a major obstacle to the introduction of machine learning methods.

CCS CONCEPTS

• **Information systems** → **Clustering**; *Social networks*; *Social recommendation*; **Data mining**; *Collaborative filtering*; *Network data models*; • **Mathematics of computing** → *Probabilistic inference problems*; **Bayesian networks**.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:

Anonymous Author(s). 2018. Supplementary Material for Dynamic Mixed Membership Stochastic Block Model for Weighted Labeled Networks. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 EXPLICIT DERIVATION OF THE E-STEP

1.1 Short derivation

This demonstration can be found in [1, 4, 5]. We recall the log-likelihood as defined in the main paper:

$$\begin{aligned}
\log P(\theta, p|R^\circ) &\propto \log P(R^\circ|\theta, p) \prod_t \prod_i P(\theta_i^{(t)}) \prod_k P(p_k^{(t)}) \\
&= \sum_{(i,o,t)\in R^\circ} \log \sum_{k\in K} \theta_{i,k}^{(t)} p_k^{(t)}(o) \\
&\quad + \sum_t \sum_i \log P(\theta_i^{(t)}) \sum_k \log P(p_k^{(t)}) \\
&\geq \sum_{(i,o,t)\in R^\circ} \sum_{k\in K} \omega_{i,o}^{(t)}(k) \log \frac{\theta_{i,k}^{(t)} p_k^{(t)}(o)}{\omega_{i,o}^{(t)}(k)} \\
&\quad + \sum_t \sum_i \log P(\theta_i^{(t)}) \sum_k \log P(p_k^{(t)})
\end{aligned} \tag{1}$$

In Eq.1, we introduced a proposal distribution $\omega_{i,o}^{(t)}(k)$, that represents the probability of one cluster allocation k given the observation (i, o, t). The last line followed from Jensen's inequality assuming that $\sum_k \omega_{i,o}^{(t)}(k) = 1$. We notice that Jensen's inequality holds as an equality when:

$$\omega_{i,o}^{(t)}(k) = \frac{\theta_{i,k}^{(t)} p_k^{(t)}(o)}{\sum_{k'} \theta_{i,k'}^{(t)} p_{k'}^{(t)}(o)} \tag{2}$$

which provides us with the expectation formula. The prior terms $P(\theta_i^{(t)})$ and $P(p_k^{(t)})$ have no effect on the result as they cancel in the inequality 1.

1.2 Full derivation

The derivation presented in this section follows a well-known general derivation of the EM algorithm, which can be found in C.M. Bishop's *Pattern Recognition and Machine Learning*-p.450 for instance.

We recall that one entry of the dataset R° takes the form of a tuple (i, o, t), where i is the input item and o an associated label at time t. $k \in K$ denotes the latent variable accounting for cluster allocation among K possible values. The total log-likelihood is the sum of each observation's log-likelihood. Without loss of generality, we focus on a single observation (i, o, t). The expression of the log-posterior distribution for one observation reads:

$$\begin{aligned}
&\log P(\theta^{(t)}, p^{(t)}|(i, o, t)) \\
&\propto \log P(R^\circ|\theta^{(t)}, p^{(t)}) P(\theta_i^{(t)}) \prod_k P(p_k^{(t)}) \\
&= \log P^{(t)}(i, o|\theta^{(t)}, p^{(t)}) + \log P(\theta_i^{(t)}) + \sum_k \log P(p_k^{(t)})
\end{aligned} \tag{3}$$

For an observation $(i, o, t) \in R^{\circ}$, we assume a distribution $Q_{i,o}^{(t)}(k)$ on the latent variables associated to it; this distribution is yet to be defined. Because k takes values among K possible ones, we have $\sum_{k \in K} Q_{i,o}^{(t)}(k) = 1$. Given this normalization condition, we can decompose each summed term in Eq.3 for any distribution $Q_{i,o}^{(t)}(k)$ as:

$$\begin{aligned}
&\log P^{(t)}(i,o|\theta^{(t)},p^{(t)}) \\
&= \underbrace{\log P^{(t)}(i,o,k|\theta^{(t)},p^{(t)}) - \log P^{(t)}(k|i,o,\theta^{(t)},p^{(t)})}_{\text{Does not depend on } k} \\
&= \sum_{k\in K} Q_{i,o}^{(t)}(k) \log P^{(t)}(i,o,k|\theta^{(t)},p^{(t)}) \\
&\quad - \sum_{k\in K} Q_{i,o}^{(t)}(k) \log P^{(t)}(k|i,o,\theta^{(t)},p^{(t)}) \\
&= \sum_{k\in K} Q_{i,o}^{(t)}(k) \log \frac{P^{(t)}(i,o,k|\theta^{(t)},p^{(t)})}{Q_{i,o}^{(t)}(k)} \\
&\quad - \sum_{k\in K} Q_{i,o}^{(t)}(k) \log \frac{P^{(t)}(k|i,o,\theta,p^{(t)})}{Q_{i,o}^{(t)}(k)}
\end{aligned} \tag{4}$$

We note that the term in the last line of Eq.4, $-\sum_{k \in K} Q_{i,o}^{(t)}(k) \log \frac{P^{(t)}(k|i,o,\theta^{(t)},p^{(t)})}{Q_{i,o}^{(t)}(k)}$ is the Kullback-Leibler (KL) divergence between $P^{(t)}$ and $Q_{i,o}^{(t)}$, noted $KL(P^{(t)}||Q_{i,o}^{(t)})$. The KL divergence obeys $KL(P^{(t)}||Q_{i,o}^{(t)}) \geq 0$, and is null iif $P^{(t)}$ equals $Q_{i,o}^{(t)}$. Therefore, the term in the before-last line of Eq.4, $\sum_{k \in K} Q_{i,o}^{(t)}(k) \log \frac{P^{(t)}(i,o,k|\theta^{(t)},p^{(t)})}{Q_{i,o}^{(t)}(k)}$, is interpreted as a lower bound on the log-likelihood $\log P^{(t)}(i,o|\theta^{(t)},p^{(t)})$.

The aim of the E-step is to find the expression of $Q_{i,o}^{(t)}(k)$ that maximizes the lower bound of the log-likelihood with respect to the latent variables k. Given that the log-likelihood does not depend on $Q_{i,o}^{(t)}(k)$ and $KL(P^{(t)}||Q_{i,o}^{(t)}) \geq 0$, the lower-bound is maximized when $KL(P^{(t)}||Q_{i,o}^{(t)}) = 0$, which occurs when $Q_{i,o}^{(t)}(k) = P^{(t)}(k|i,o,\theta^{(t)},p^{(t)})$. In this case, the lower-bound on the log-likelihood equals the likelihood itself and thus reaches a global maximum with respect to the latent variables k for fixed parameters $\theta^{(t)}$ and $p^{(t)}$.

Given the definition of our simple model, the derivation of $P(k|i,o,\theta^{(t)},p^{(t)})$ is straightforward. The probability of one combination of clusters k among K possible combinations given an input features vector and an output o is proportional to $p_k^{(t)}(o)\theta_{i,k}$. Therefore:

$$P^{(t)}(k|i,o,\theta^{(t)},p^{(t)}) = \frac{p_k^{(t)}(o)\theta_{i,k}^{(t)}}{\sum_{k'\in K} p_{k'}^{(t)}(o)\theta_{i,k'}^{(t)}} \tag{5}$$

which is the expression of $\omega_{i,o}^{(t)}(k)$ in the main article.

2 EXPLICIT DERIVATION OF THE M-STEP FOR P

$$\frac{\partial\left(\log P(\theta,p|R^{\circ}) - \sum_{k',t'} \psi_{k'}^{(t')}(\sum_o p_{k'}^{(t')}(o) - 1)\right)}{\partial p_k^{(t)}(o)} = 0$$

$$\leftrightarrow \sum_{(i,t)\in\partial o} \frac{\omega_{i,o}^{(t)}(k)}{p_k^{(t)}(o)} + \frac{\beta\langle p_k^{(t)}(o)\rangle}{p_k^{(t)}(o)} - \psi_k^{(t)} = 0$$

$$\leftrightarrow \sum_{(i,t)\in\partial o} \omega_{i,o}^{(t)}(k) + \beta\langle p_k^{(t)}(o)\rangle = \psi_k^{(t)} p_k^{(t)}(o)$$

$$\leftrightarrow \sum_{(i,t)\in\partial o} \sum_o \omega_{i,o}^{(t)}(k) + \beta \underbrace{\sum_o \langle p_k^{(t)}(o)\rangle}_{=1} = \psi_k^{(t)}$$

$$\leftrightarrow \frac{\sum_{(i,t)\in\partial o} \omega_{i,o}^{(t)}(k) + \beta\langle p_k^{(t)}(o)\rangle}{\sum_{(i,o,t)\in R^{\circ}} \omega_{i,o}^{(t)}(k) + \beta} = \theta_{i,k}^{(t)} \tag{6}$$

3 USING THE PRIOR ON SIMILAR WORKS

Throughout this section, we highlight the changes brought by our method to the EM equations derived in the mentioned papers. In summary, we see that out method allows to make these works dynamic with minimal changes of the original models.

3.1 Bi-MMSBM [1]

In [1], the authors apply a MMSBM to a labeled bipartite network. The nodes on each side of the bipartite network are associated to their own membership matrix; membership of nodes $i \in I$ over K clusters is encoded into $\theta \in \mathbb{R}^{I\times K}$, and membership of nodes $j \in J$ over L clusters is encoded into $\eta \in \mathbb{R}^{J\times L}$. The block-interaction matrix for the label $o \in O$ is noted $p(o) \in \mathbb{R}^{K\times L}$.

Assuming a temporal version, items i and j to be linked by a label o at time t reads:

$$P^{(t)}(o|i,j) = \sum_{k\in K}\sum_{l\in L} \theta_{i,k}^{(t)} \eta_{j,l}^{(t)} p_{k,l}^{(t)}(o) \tag{7}$$

Given the same set of observation s R° as in the main article, the posterior distribution follows:

$$\begin{aligned}
P(\theta,\eta,p|R^{\circ}) &= P(R^{\circ}|\theta,\eta,p) \\
&\times \prod_t \left(\prod_i P(\theta_i^{(t)}) \prod_j P(\eta_j^{(t)}) \prod_{k,l} P(p_{k,l}^{(t)}) \right)
\end{aligned} \tag{8}$$

such that:

$$P(R^{\circ}|\theta,\eta,p) = \prod_{(i,j,t,o)\in R^{\circ}} \sum_{k\in K}\sum_{l\in L} \theta_{i,k}^{(t)} \eta_{j,l}^{(t)} p_{k,l}^{(t)}(o) \tag{9}$$

$$P(\theta_i^{(t)}) \propto \prod_k \theta_{i,k}^{(t)\,\beta\langle\theta_{i,k}^{(t)}\rangle} \tag{10}$$

$$P(\eta_j^{(t)}) \propto \prod_l \eta_{j,l}^{(t)\,\beta\langle\eta_{j,l}^{(t)}\rangle} \tag{11}$$

$$P(p_{k,l}^{(t)}) \propto \prod_o p_{k,l}^{(t)}(o)^{\beta\langle p_{k,l}^{(t)}(o)\rangle} \tag{12}$$

where $\langle x^{(t)} \rangle = \frac{\sum_{t' \neq t} \kappa(t,t') x^{(t')}}{\sum_{t' \neq t} \kappa(t,t')}$. The expectation step is not influenced by the priors choice and is the same as in [1] for each temporal slice. The new maximization steps are:

$$\theta_{i,k}^{(t)} = \frac{\sum_l \sum_{(o,j) \in \partial(i,t)} \omega_{i,j,o}^{(t)}(k,l) + \beta \langle \theta_{i,k}^{(t)} \rangle}{N_{i,t} + \beta}$$

$$\eta_{j,l}^{(t)} = \frac{\sum_k \sum_{(o,i) \in \partial(j,t)} \omega_{i,j,o}^{(t)}(k,l) + \beta \langle \eta_{j,l}^{(t)} \rangle}{N_{j,t} + \beta}$$

$$p_{k,l}^{(t)}(o) = \frac{\sum_{(i,j,t) \in \partial o} \omega_{i,j,o}^{(t)}(k,l) + \beta \langle p_{k,l}^{(t)}(o) \rangle}{\sum_{(i,j,o,t) \in R^\circ} \omega_{i,j,o}^{(t)}(k,l) + \beta}$$

Here again, β is set fixed for demonstration purposes, but can be tuned at will by the user. This allows to choose the extent to which dynamics shall be smoothed, or ignored.

3.2 T-MBM [5]

The T-MBM is essentially the same model as [1] but with one type of entry that can appear twice in one observation. Both entries share the same membership matrix θ. The probability of a label of type o given entries h; i and j at time t is now written:

$$P(o|h,i,j,t) = \sum_{k \in K} \sum_{l \in L} \sum_{m \in M} \theta_{h,k}^{(t)} \theta_{i,l}^{(t)} \eta_{j,m}^{(t)} p_{k,l,m}^{(t)}(o) \tag{13}$$

The posterior distribution follows the same expression as in Eq.8. The expectation step is left unchanged by the choice of the priors, and the new maximization equations are given below:

$$\theta_{h,k}^{(t)} = \frac{\sum_{l,m} \sum_{(o,i,j) \in \partial(h,t)} \omega_{h,i,j,o}^{(t)}(k,l,m) + \beta \langle \theta_{h,k}^{(t)} \rangle}{N_{h,t} + \beta}$$

$$\eta_{i,l}^{(t)} = \frac{\sum_{k,m} \sum_{(o,h,j) \in \partial(i,t)} \omega_{h,i,j,o}^{(t)}(k,l,m) + \beta \langle \eta_{i,l}^{(t)} \rangle}{N_{i,t} + \beta}$$

$$p_{k,l,m}^{(t)}(o) = \frac{\sum_{(h,i,j,t) \in \partial o} \omega_{h,i,j,o}^{(t)}(k,l,m) + \beta \langle p_{k,l,m}^{(t)}(o) \rangle}{\sum_{(h,i,j,o,t) \in R^\circ} \omega_{h,i,j,o}^{(t)}(k,l,m) + \beta}$$

4 JOINTLY INFERRING θ AND p

In the main article, p is provided to the model and only θ has to be inferred. Doing so, we can confront inferred membership vectors to the ground truth while avoiding label-switching issues [2, 3]. When p is also to be inferred, finding a correspondence between the inferred clusters and the ground-truth is not a trivial task, and cannot be performed in unbiased ways. However, the good results yielded by the model, presented in Fig.1, when also inferring p hints that the membership vectors are correctly inferred.

5 CLUSTERS COMPOSITION FOR THE EPIGRAPHY EXPERIMENT

- Cluster 0
  - Roma (98.0%)
- Cluster 1
  - Latium et Campania (32.0%)
  - Venetia et Histria (14.0%)
  - Samnium (11.0%)
  - Umbria (10.0%)
  - Apulia et Calabria (8.0%)
- Cluster 2
  - Pannonia superior (15.0%)
  - Dalmatia (11.0%)
  - Noricum (10.0%)
  - Hispania citerior (7.0%)
  - Gallia Narbonensis (6.0%)
- Cluster 3
  - Dacia (24.0%)
  - Pannonia inferior (17.0%)
  - Moesia inferior (15.0%)
  - Syria (6.0%)
  - Numidia (6.0%)
  - Pannonia superior (5.0%)
- Cluster 4
  - Germania superior (24.0%)
  - Mauretania Caesariensis (11.0%)
  - Asia (11.0%)
  - Etruria (11.0%)
  - Galatia (9.0%)

REFERENCES

[1] Antonia Godoy-Lorite, Roger Guimerà, Cristopher Moore, and Marta Sales-Pardo. 2016. Accurate and scalable social recommendation using mixed-membership stochastic block models. *Proc. Natt. Acad. Sci.* 113, 50 (2016), 14207–14212. https://doi.org/10.1073/pnas.1606316113
[2] Clement Lee and Darren J. Wilkinson. 2019. A review of stochastic block models and extensions for graph clustering. *Applied Network Science* 4 (2019), 1–50. https://doi.org/10.1007/s41109-019-0232-2
[3] Catherine Matias and Vincent Miele. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *J. Roy. Stat. Soc.: Series B (Stat. Meth.)* 79, 4 (2017), 1119–1141. https://doi.org/10.1111/rssb.12200
[4] Gaël Poux-Médard, Julien Velcin, and Sabine Loudcher. 2021. Information Interactions in Outcome Prediction: Quantification and Interpretation Using Stochastic Block Models. *Proc. 15th ACM Conf. on Rec. Sys. (RecSys)* (2021), 199–208. https://doi.org/10.1145/3460231.3474254
[5] Marc Tarrés-Deulofeu, Antonia Godoy-Lorite, Roger Guimerà, and Marta Sales-Pardo. 2019. Tensorial and bipartite block models for link prediction in layered networks and temporal networks. *Phys. Rev. E* 99 (2019), 032307. Issue 3. https://doi.org/10.1103/PhysRevE.99.032307

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

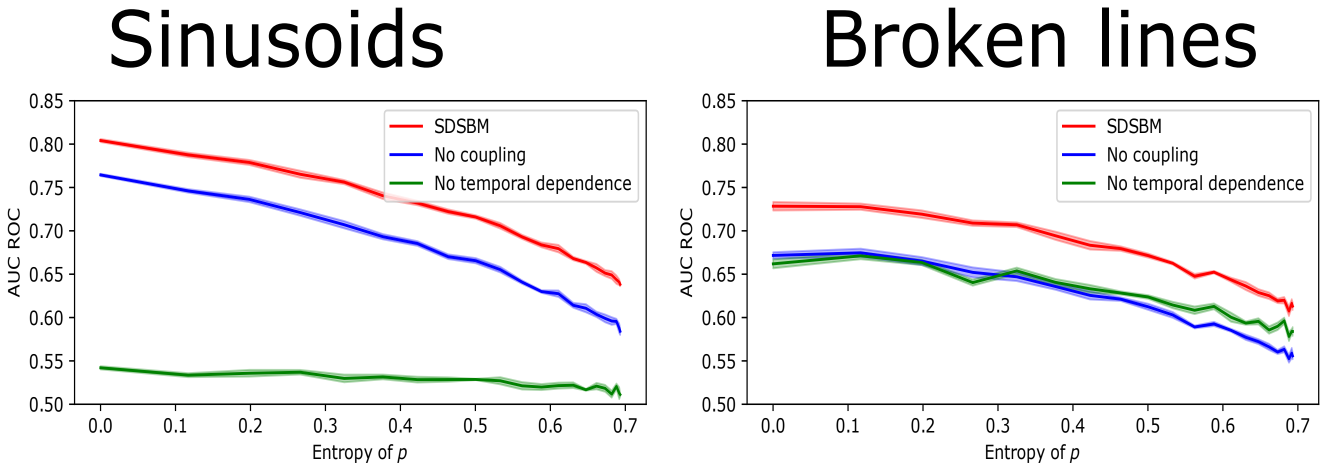

Figure 1: Experimental results when inferring both θ and p jointly. The AUC-ROC is as good as when p is provided to the model.