

---

# Temporal model for Enron email dataset

---

**Leman Akoglu and Seungil Huh**

School of Computer Science

Carnegie Mellon University

Pittsburgh, PA 15213

lakoglu@cs.cmu.edu & seungilh@cs.cmu.edu

## Abstract

The e-mail database for the Enron Corp. linked to the prosecution of a number of its senior executives poses an interesting challenge for researchers interested in network modeling. In this paper we adapt the mixed membership stochastic blockmodel (MMSB) approach to a time-varying setting and apply both the original MMSB and this extension to a version of the Enron database.

## 1 Introduction

Enron email dataset is one of the important datasets for social network analysis as a unique large-scale real email corpus that is generally accessible to the public. The expectation to discover a clue about bankruptcy of the Enron Corporation has interested researchers in the social network analysis. In addition, the fact that this dataset is a temporal record over a period of 3.5 years has also contributed to the value of the dataset. For these reasons, researchers have applied various methods such as automatic categorization methods, [2], graph theoretical analysis [3], data mining techniques [4], and language content analysis [6] to this Enron email dataset. Although various approaches have been exploited for Enron email dataset, statistical models concerned with pair-wise relations have seldom been applied.

In this paper, we first analyze group membership of the employees in the Enron Corporation applying the mixed membership stochastic blockmodels (MMSB) [1]. In addition, we work toward a statistical model for dynamic network analysis based on the MMSB. Ideally we should be thinking in terms of the number of latent groups and the mixed membership scores in the MMSB changing over time. In our simplified modeling strategy we break the data into separate time intervals, for each of which we employ a MMSB model, and thus we structure our model to maintain interactions over between consecutive time periods, using a Markovian-like formulation.

In the next section we describe this model and then we return to the Enron data and our analyses based on the original MMSB and our model. We conclude with some thoughts on how to expand the model and further develop our analyses of the Enron database.

## 2 Model

In this section, we briefly describe the mixed membership stochastic blockmodel (MMSB) [1] and introduce our model based on the MMSB. In the MMSB, relational data in the social network is represented as a graph  $\mathcal{G} = (\mathcal{N}, \mathcal{R})$ , which is drawn from the following procedure:

For each actor  $p \in \mathcal{N}$ :

- Draw a  $K$  dimensional mixed membership vector  $\vec{\pi}_p \sim \text{Dirichlet}(\vec{\alpha})$ .

For each pair of nodes  $(p, q) \in \mathcal{N} \times \mathcal{N}$ :

- Draw membership indicator for the initiator,  $\vec{z}_{p \rightarrow q} \sim \text{Multinomial}(\vec{\pi}_p)$ .
- Draw membership indicator for the receiver,  $\vec{z}_{q \rightarrow p} \sim \text{Multinomial}(\vec{\pi}_q)$ .
- Sample the value of their interaction,  $R(p, q) \sim \text{Bernoulli}(\vec{z}_{p \rightarrow q}^T B \vec{z}_{p \leftarrow q})$ .

where  $\mathcal{N}$  is the set of actors and  $\mathcal{R}$  is the positive relation between all pairs of the actors.

We propose to frame our analyses by dividing the temporally evolving internal Enron e-mail network into  $T$  sub-networks with respect to time line and represented as a graph  $\mathcal{G} = (\mathcal{N}^{(1:T)}, R^{(1:T)})$ , where  $\mathcal{N}^{(t)}$  is the set of actors at epoch  $t$  and  $R^{(t)}$  corresponds to a binary positive relation representing communication between pairs of the actors at epoch  $t$ . The basic notion is to have separate, but linked MMSBs for the time periods.

For the MMSB applied to the  $t$ -th subnetwork, the relation between two actors  $p$  and  $q$  is sampled according to the following procedure:

$$R^{(t)}(p, q) \sim \text{Bernoulli}(I^{(t)}(p, q))$$

$$I^{(t)}(p, q) = \vec{z}_{p \rightarrow q}^{(t)T} B^{(t)} \vec{z}_{p \leftarrow q}^{(t)} \quad (1)$$

where  $R^{(t)}(p, q)$  and  $I^{(t)}(p, q)$  respectively denote a positive relation and an interaction score between  $p$  and  $q$ ,  $\vec{z}_{p \rightarrow q}^{(t)}$  and  $\vec{z}_{p \leftarrow q}^{(t)}$  are the membership indicators of  $p$  and  $q$ , and  $B^{(t)}$  is the group interaction matrix at epoch  $t$ . Our model updates (1) using the previous interaction score that links successive time periods as follows:

$$I^{(t)}(p, q) = \left( \vec{z}_{p \rightarrow q}^{(t)T} B^{(t)} \vec{z}_{p \leftarrow q}^{(t)} \right)^{1-\tau^{(t)}} I^{(t-1)}(p, q)^{\tau^{(t)}} \quad (2)$$

where  $\tau^{(t)}$  is the influence rate of the previous interaction at epoch  $t$ , which ranges between 0 and 1. The main idea of this model is that the interactions between actors tend to be maintained in the consecutive time period although the number of latent groups and the mixed membership scores on the MMSB may change. If  $\tau^{(t)} = 0$ ,  $I^{(t)}(p, q)$  is independent to  $I^{(t-1)}(p, q)$  so that this model exactly represents separate MMSBs for each time period. The main process is illustrated in Figure 1.

To date, all experiments have been performed on the simplified version of this model. In the simplified model, we assume that  $I^{(t-1)}$  is given as

$$\left( \vec{\pi}_p^{(t-1)T} B^{(t-1)} \vec{\pi}_q^{(t-1)} \right)^{1-\tau^{(t)}} I^{(t-2)}(p, q)^{\tau^{(t)}} \quad \text{at epoch } t \geq 3$$

$$\left( \vec{\pi}_p^{(t-1)T} B^{(t-1)} \vec{\pi}_q^{(t-1)} \right) \quad \text{at epoch } t = 2$$

In other words, we estimate parameters and perform posterior inference sequentially as the network and time evolve. We update the variational approximation used in [1] for the linked multinomial parameters as follows:

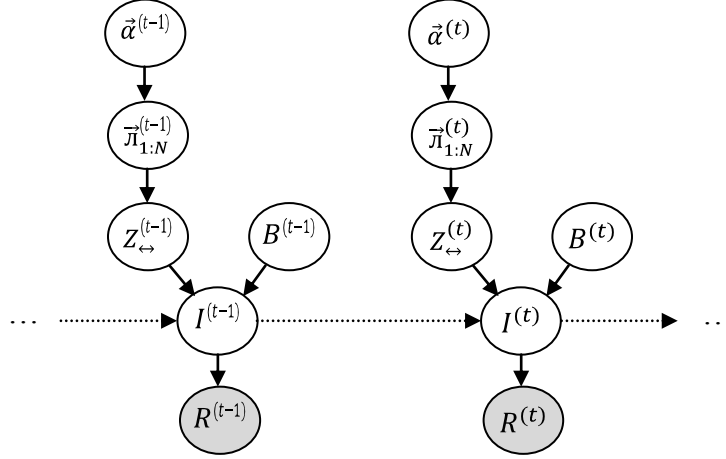


Figure 1: The graphical model of the proposed model.  $N$  denotes the number of actors and  $Z^{\leftrightarrow(t)}$  denotes the set of group indicators, or  $\{\vec{z}_{p \rightarrow q}^{(t)}, \vec{z}_{p \leftarrow q}^{(t)} : p, q \in \mathcal{N}\}$ . Dashed lines represent the dependency of pair-wise interaction scores between two consecutive epochs.

$$\hat{\phi}_{p \rightarrow q, g}^{(t)} \propto e^{E_q[\log \pi^{(t)}_{p, g}]} \cdot \prod_h \left( (B^{(t)}(g, h)^{1-\tau^{(t)}} I^{(t-1)}(p, q)^{\tau^{(t)}})^{R^{(t)}(p, q)} (1 - B^{(t)}(g, h)^{1-\tau^{(t)}} I^{(t-1)}(p, q)^{\tau^{(t)}})^{1-R^{(t)}(p, q)} \right)^{\phi_{p \rightarrow q, h}^{(t)}}$$

$$\hat{\phi}_{p \leftarrow q, h}^{(t)} \propto e^{E_q[\log \pi^{(t)}_{q, h}]} \cdot \prod_g \left( (B^{(t)}(g, h)^{1-\tau^{(t)}} I^{(t-1)}(p, q)^{\tau^{(t)}})^{R^{(t)}(p, q)} (1 - B^{(t)}(g, h)^{1-\tau^{(t)}} I^{(t-1)}(p, q)^{\tau^{(t)}})^{1-R^{(t)}(p, q)} \right)^{\phi_{p \leftarrow q, g}^{(t)}}$$

for  $g, h = 1, \dots, K$  where  $K$  is the number of latent groups. Also, the approximate MLE of  $B$  is

$$\hat{B}^{(t)}(g, h) = \text{Min} \left( \frac{\sum_{p, q} \left( R^{(t)}(p, q) / I^{(t-1)}(p, q)^{\tau^{(t)}} \right)^{\frac{1}{1-\tau^{(t)}}} \cdot \phi_{p \rightarrow q, g}^{(t)} \phi_{p \leftarrow q, h}^{(t)}}{\sum_{p, q} \phi_{p \rightarrow q, g}^{(t)} \phi_{p \leftarrow q, h}^{(t)}}, 1 \right)$$

for every index pair  $(g, h) \in [1, K] \times [1, K]$ . The remainder of our estimation procedure is same as for the MMSB as described by Airolidi et al. [1].

In order to find the most appropriate number of latent clusters, we employed the approximated Bayesian information criteria used by Airolidi et al [1]. Also, we use  $\tau^{(t)} = 0.1$  for all  $t$ .

### 3 Social network construction from Enron Database

Among several versions of Enron email dataset, we chose a version provided by Shetty and Adibi from ISI that has been cleaned by removing a large number of duplicate, junk, blank, and returned messages.<sup>1</sup> The dataset contains 252,759 emails from 151 Enron employees distributed in approximately 3000 user defined folders [6]. Among them, we use header information to verify that both the sender and at least one of the recipients are among the

<sup>1</sup> The Enron email corpus has several versions. The original one is posted as part of the investigation by the Federal Energy Regulatory Commission (FERC) in May of 2002. This dataset consists of 619,449 emails from 158 employees. However, since a significant part of the original dataset is shown to be repetitive, we use the cleaned version provided by Jitesh Shetty and Jafar Adibi from ISI.

151 known employees. Recipients are retrieved from 'to', 'cc', and 'bcc' fields without distinction.

To extract positive relations among Enron employees from the email dataset, a basic threshold scheme is used; that is, if the number of emails from one employee to another employee is greater than or equal to the threshold, a directed edge is constructed from the sender to the recipient. Our choice of the value of the threshold is one. As a result of thresholding, 2149 positive relations among 150 employees can be extracted (a single employee without connections is removed).

In order to analyze the dynamics of the communication network, we construct five communication networks corresponding to five time chunks, 1)  $t_1$ : before July 2000, 2)  $t_2$ : July 2000~ December 2000, 3)  $t_3$ : January 2001~ June 2001, 4)  $t_4$ : July 2001~ December 2001, and 5)  $t_5$ : after December 2001. These networks are shown in Figure 2. The numbers of employees and positive relations for each time period are shown in Table 1. Note that the fourth network has the most edges. This phenomenon is due to the series of events related to the financial crisis at Enron. On Aug. 2001 the CEO Jeffery Skilling resigned and Kenneth Lay was named as CEO again. The crisis fully broke out on Oct. 2001, which was followed by a federal investigation [4].

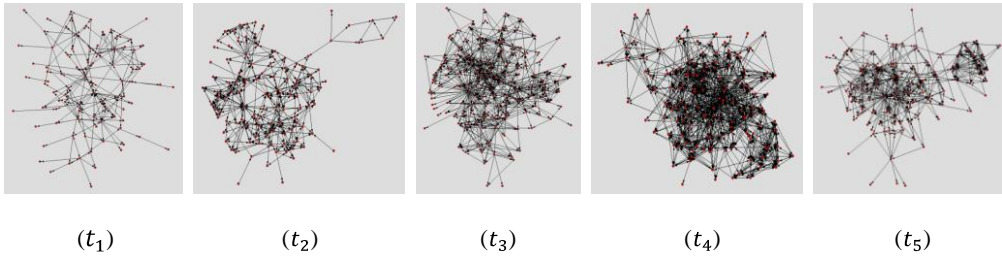


Figure 2: Temporal communication networks for five time chunks.

Table 1. Temporal summary of communication networks to five time chunks

Time Period	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$
# of employees	87	110	142	142	119
# of positive relations	201	480	819	1380	610

## 4 Experimental Result

### 4.1 Group membership analyses

Before the analyses based on the suggested model, we describe the result obtained by applying the original MMSB to the social network extracted from the whole Enron e-mail database. We find 6 latent groups in the social network using the approximated BIC scheme. Figure 3 shows the membership scores of 150 Enron employees for all time periods. As can be seen in the figure, many employees display mixed membership.

Since the group information of Enron Corporation is not available, we try to find meaningful pair-wise relations from the group membership scores. Based on the real status information of the employees introduced by [7], we describe some interesting pairs of employees as follows.

*James Steffes and Richard Shapiro*: Steffes was the Vice President of Government Affairs and Shapiro was the Vice President of Regulatory Affairs. These two people have high

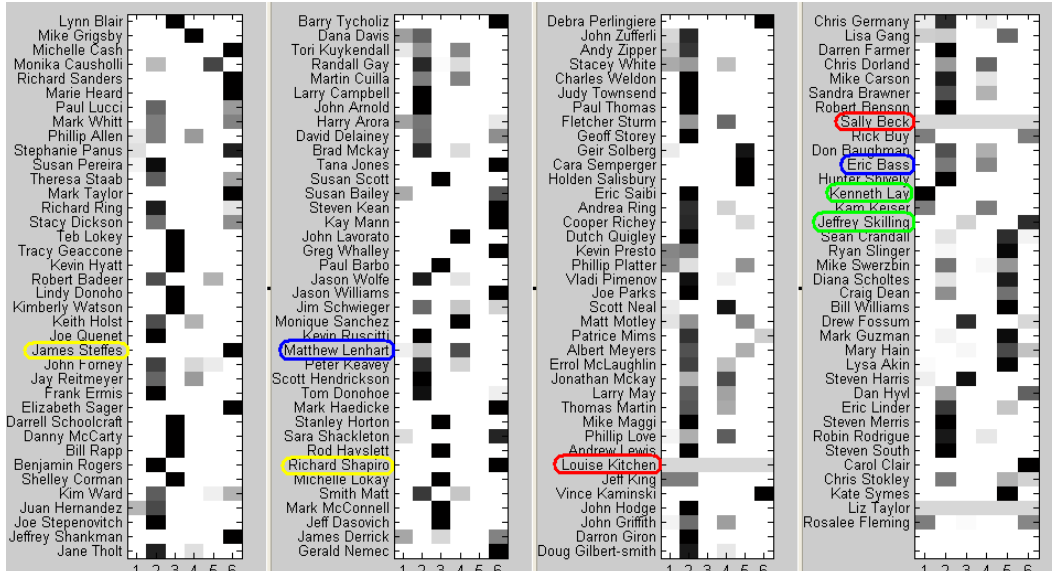


Figure 3. The posterior mixed membership scores of the MMSB model for all time periods. The 6 groups are displayed on the  $X$  axis and the 150 Enron employees are displayed on the  $Y$  axis. Darker shading indicates higher membership score. Several meaningful relations we found are circled on the figure with the same color. (This figure is best viewed in color.)

membership score to the 6<sup>th</sup> group, but no other group, which means that not only their affairs are highly related, but also they served for a distinct specialty in the company.

*Eric Bass and Matthew Lenhart:* Bass was the Coordinator of a fantasy basketball league and Lenhart had received many emails about paintball and bowling. They both have probability of belonging to two same groups. Based on this information, we conjecture that one of the two groups is related to their role in social activities and the other is due to their employment.

*Sally Beck and Louise Kitchen:* Beck was the Chief Operating Officer and Kitchen was the President of Enron Online. They have similar probability of belonging to all groups. They might be in contact with many people in the company which makes them have connection to all groups.

*Kenneth Lay and Jeffrey Skilling* must be the most important people for the financial crisis of the Enron Corporation. However, in the Figure 3, we cannot find anything about their relation in terms of the group membership; that is, they belong to the different groups. This unreasonable result is because we do not consider the change of their relation along the time line. We will examine their relation based on time-varying models in the next section.

## 4.2 Social Dynamics over time

First we apply the MMSB to each of sub-social networks extracted from 5 time epochs. (Separate MMSBs). We also apply the simplified version of the suggested model and compare these two models. Basically, while the Separate MMSBs provide static snapshots of the network in terms of membership scores and group interaction, our model links consecutive time periods and smoothes the change between the periods.

Figure 4 shows the latent group numbers for separate MMSBs fit to the data and for the suggested model. In both models, the numbers of latent groups increase till  $t_4$  and decrease at  $t_5$ . This phenomenon accounts for the expansion and reduction of the social network in terms of actors and their e-mail communications. With separate MMSBs, we find 3, 3, 5, 7 and 3 latent groups at each time epoch. On the other hand, our model which links these MMSBs dampens the overall change across the epoch boundaries, and finds 3, 3, 3, 6 and 3

latent groups at each time epoch. Our model discovered relatively small number of latent groups across the time intervals of interest compared to the separate MMSBs because it extracts latent groups not only from the current e-mail relations, but also based on the previous interactions among the actors.

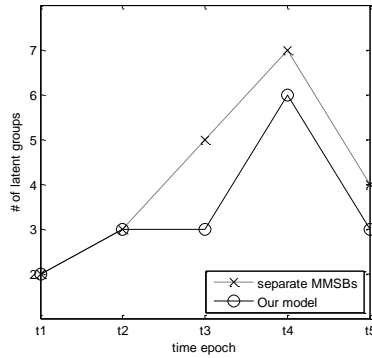


Figure 4. Comparison of the two models in terms of the latent group numbers.

Figure 5 shows the membership scores of the two key people in Enron Corporation. While MMSB applied to the entire social network did not discover anything about their relation, these two time-varying models show that they belong to same group at epoch  $t_1$ ,  $t_2$ , and  $t_5$ . In addition, they also share group membership at epoch  $t_3$ . The phenomenon that they have different group membership at epoch  $t_4$  may be cause by the Enron Corp. financial crisis. Before the crisis fully broke out on Oct. 2001, Skilling resigned from the CEO on Aug. 2001 and Lay tried to handle the crisis after him. Therefore, at the epoch  $t_4$  which contains these events, the roles of Skilling and Lay may be significantly different, which account for their membership difference.

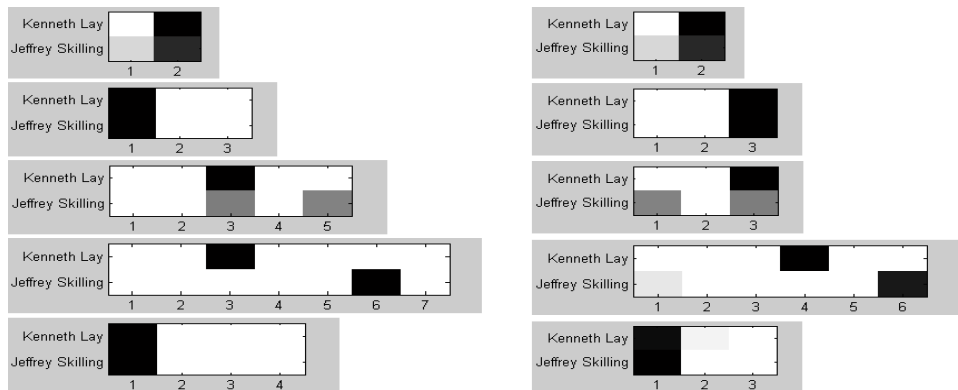


Figure 5. the membership scores of *Kenneth Lay* and *Jeffrey Skilling* at each time epoch based on Separate MMSBs (Left) and Our model (Right). Each line from top to bottom corresponds to each time epoch from  $t_1$  to  $t_5$ . Each group is displayed on the X axis. Darker shading indicates higher membership score.

The approximated BIC scores at each time epoch for both models are also shown in Table 2. The reason that we use the BIC score instead of log likelihood is that the number of latent group may be different for two models. The BIC scores of our model tend to be greater than those of the separate MMSBs for every epoch (The BIC scores at  $t_1$  are same because our

model applies the original MMSB due to absence of the previous time epoch). Although it is not definite that our model is more expressive than the separate MMSBs because the differences are not significant, we can conclude that the information accumulated in the previous time periods may be useful to improve the following models.

Table 2. BIC scores of separate MMSBs and our model for each time epoch

ime epoch	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$
Separate MMSBs	-1795	-4050	-6540	-8468	-4230
Our Model	-1795	-3972	-6330	-8286	-4132

## 5 Conclusion and Discussion

In this paper, we apply the MMSB and its extension to the social network from the Enron email database and show the results. Our contributions are in particular the following:

1. We apply a model based clustering method concerned with pair-wise relations to the Enron network, which has seldom been applied.
2. We extend the original MMSB to the temporal models that can discover meaningful relationship between the key people in Enron Corporation in terms of group membership.
3. We introduce the new model that links consecutive models and dampens the change between the time periods. Also we show the expression power of our model in terms of BIC scores though experiments.

Graph representation can visually show the difference of our model and separate MMSBs. We leave this work as one of the future works. Also we plan to exploit inference methods for our original suggested model. Lastly, we would like to apply our model to another dataset such as DBLP dataset [8].

## References

- [1] Airoldi, E.M., Blei, D.M., Fienberg, S.E. & Xing, E.P. (2007) Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*.
- [2] Bekkerman, R., McCallum, A. & Huang, G. (2004) Automated Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora, *CIIR Technical Report IR-418*.
- [3] Chapanond, A., Krishnamoorthy, M.S., & Yener, B. (2005) Graph Theoretic and Spectral Analysis of Enron Email Data, *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, SIAM International Conference on Data Mining, pp. 15-22.
- [4] Diesner, J. and Carley, K.M. (2005) Exploration of Communication Network from the Enron Email Corpus, *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, *SIAM International Conference on Data Mining*, pp. 3-14.
- [5] Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. (1999) Introduction to variational methods for graphical models. *Machine Learning*, 37:183-233.
- [6] McCallum, A., Corrada-Emmanuel, A. & Wang X. (2004) The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email. Technical Report UM-CS-2004-096.
- [7] Shetty, J. & Adibi, J. The Enron email dataset database schema and brief statistical report.
- [8] <http://kdl.cs.umass.edu/data/dblp/dblp-info.html>

## A Derivation of Variational Algorithm

Under the our model, the joint probability of the relational data  $R^{(t)}$  and the latent variables  $\{\vec{\pi}_{1:N}^{(t)}, Z_{\rightarrow}^{(t)}, Z_{\leftarrow}^{(t)}\}$  can be written in the following form,

$$\begin{aligned} & p(R^{(t)}, \vec{\pi}_{1:N}^{(t)}, Z_{\rightarrow}^{(t)}, Z_{\leftarrow}^{(t)} | \vec{\alpha}^{(t)}, B^{(t)}) \\ &= \prod_{p,q} P(R^{(t)}(p,q) | \vec{z}_{p \rightarrow q}^{(t)}, \vec{z}_{p \leftarrow q}^{(t)}, B^{(t)}, I^{(t-1)}) P(\vec{z}_{p \rightarrow q}^{(t)} | \vec{\pi}_p^{(t)}) P(\vec{z}_{p \leftarrow q}^{(t)} | \vec{\pi}_q^{(t)}) \prod_p P(\vec{\pi}_p^{(t)} | \vec{\alpha}^{(t)}). \end{aligned}$$

where  $Z_{\rightarrow}^{(t)}$  and  $Z_{\leftarrow}^{(t)}$  respectively denote  $\{\vec{z}_{p \rightarrow q}^{(t)} : p, q \in N^{(t)}\}$  and  $\{\vec{z}_{p \leftarrow q}^{(t)} : p, q \in N^{(t)}\}$ .

Based on the mean-field theory [5], we can approximate this intractable distribution by a fully factored distribution introduced in [1] and compute the approximate lower bound for the likelihood by minimizing the Kulback-Leibler divergence between the distributions. The approximate lower bound of our model is same as that of [1] except  $f(R^{(t)}(p,q), B^{(t)}(g,h))$  which is defined in our model as follows.

$$\begin{aligned} & f(R^{(t)}(p,q), B^{(t)}(g,h)) = \\ & R^{(t)}(p,q) \log \left( B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right) + \left( 1 - R^{(t)}(p,q) \right) \log \left( 1 - B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right) \end{aligned}$$

This difference leads to the following updates for the variational parameters  $(\vec{\phi}_{p \rightarrow q}^{(t)}, \vec{\phi}_{p \leftarrow q}^{(t)})$ , for a pair of nodes  $(p,q)$ .

$$\begin{aligned} \hat{\phi}_{p \rightarrow q, g}^{(t)} &\propto e^{E_q[\log \pi_{p,g}^{(t)}]} \cdot e^{\sum_h \phi_{p \rightarrow q, h}^{(t)} \cdot E_q[f(R^{(t)}(p,q), B^{(t)}(g,h))]} \\ &= e^{E_q[\log \pi_{p,g}^{(t)}]} \cdot \prod_h \left( \left( B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right)^{R^{(t)}(p,q)} \left( 1 - B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right)^{1-R^{(t)}(p,q)} \right)^{\phi_{p \rightarrow q, h}^{(t)}} \\ \hat{\phi}_{p \leftarrow q, h}^{(t)} &\propto e^{E_q[\log \pi_{q,h}^{(t)}]} \cdot e^{\sum_g \phi_{p \rightarrow q, g}^{(t)} \cdot E_q[f(R^{(t)}(p,q), B^{(t)}(g,h))]} \\ &= e^{E_q[\log \pi_{q,h}^{(t)}]} \cdot \prod_g \left( \left( B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right)^{R^{(t)}(p,q)} \left( 1 - B^{(t)}(g,h)^{1-\tau^{(t)}} I^{(t-1)}(p,q)^{\tau^{(t)}} \right)^{1-R^{(t)}(p,q)} \right)^{\phi_{p \rightarrow q, g}^{(t)}} \end{aligned}$$

for  $g, h = 1, \dots, K$  where  $K$  is the number of latent groups.

Also, isolating terms containing  $B$  from the approximate lower bound we obtain

$$\mathcal{L}_B^{(t)} = \sum_{p,q} \sum_{g,h} \phi_{p \rightarrow q, g}^{(t)} \phi_{p \leftarrow q, h}^{(t)} \cdot f(R^{(t)}(p,q), B^{(t)}(g,h))$$

whose approximate MLE of  $B$  is

$$\hat{B}^{(t)}(g,h) = \text{Min} \left( \frac{\sum_{p,q} \left( R^{(t)}(p,q) / I^{(t-1)}(p,q)^{\tau^{(t)}} \right)^{\frac{1}{1-\tau^{(t)}}} \cdot \phi_{p \rightarrow q, g}^{(t)} \phi_{p \leftarrow q, h}^{(t)}}{\sum_{p,q} \phi_{p \rightarrow q, g}^{(t)} \phi_{p \leftarrow q, h}^{(t)}}, 1 \right)$$