Dynamic Mixed Membership Blockmodel for Evolving Networks

Wenjie Fu WENJIEF@CS.CMU.EDU
Le Song LESONG@CS.CMU.EDU
Eric P. Xing EPXING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

In a dynamic social or biological environment, interactions between the underlying actors can undergo large and systematic changes. Each actor can assume multiple roles and their degrees of affiliation to these roles can also exhibit rich temporal phenomena. We propose a *state space mixed membership stochastic blockmodel* which can track across time the evolving roles of the actors. We also derive an efficient variational inference procedure for our model, and apply it to the Enron email networks, and rewiring gene regulatory networks of yeast. In both cases, our model reveals interesting dynamical roles of the actors.

1. Introduction

Over the course of a temporal process, such as the development of a company, there may exist multiple underlying themes that determine the functions of the actors and their relations to each other, and such themes are dynamic and stochastic. As a result, the networks of interactions between the actors at each time point are context-dependent and can undergo systematic rewiring, rather than being invariant over time. Furthermore, the actors involved in such dynamic interactions can also exhibit multiple functionalities in order to meet the requirement of the evolving themes.

It is our goal to dissect the evolving functional composition of the actors based on their dynamic interactions, which we call the *dynamic network tomography*. Dynamic network tomography can lead to important insights to the robustness of network structures, the cause and consequence of information diffusion, and the mechanism of hierarchy and organization formation. By appropriately modeling network tomography, a network analyst can also simulate and reason about the generative mechanisms of networks, and discover changing themes in networks, which will be relevant for activity and anomaly detection. Therefore, methods for finding the evolving functional and semantic underpinnings of dynamic networks are essential for understanding the organization and re-organization of complex relational networks.

Recently there is a surge of interest in applying latent space model for network modeling and analysis (Hoff et al., 2002; Li & McCallum, 2006; Handcock et al., 2007; Erosheva et al., 2004; Airoldi et al., 2008). However, an important aspect has not been addressed so far: none of these models considered the dynamic nature of the networks. Rather than having a single and invariant role, an actor in dynamic networks can undergo multiple sources of influence, and it can actively undertake a specific role depending on who it is interacting with and when this interaction occurs.

Here we propose a *dynamic mixed membership stochastic block model* (dMMSB) based on a latent space model proposed recently by Airoldi et al. (2008). To deal with the dynamic nature of the networks, we apply ideas similar to the correlated topic model (Blei & Lafferty, 2006a) and dynamic topic model (Blei & Lafferty, 2006b) for text document analysis. The major contributions of our paper are:

- Our model incorporates a state space model for tracking the mixed memberships of network actors in the latent space. This allows us to study the dynamic evolution of the functional roles of network actors. First and foremost, it is a useful tool for analyzing network topology and visualizing the property of entities in dynamic networks.
- Our application of logistic normal prior to the mixed membership vectors is novel in the context of network modeling. This prior provides us the extra ability to capture the relational structure

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

between latent functional roles.

- We also develop a Laplace variational EM algorithm for performing efficient learning and inference. This algorithm enables us to study real world dynamic networks such as the Enron email networks and yeast gene regulatory networks.

The remainder of the paper is organized as follow. In Section 2, we present the dMMSB model in detail. The Laplace variational EM algorithm will be described briefly in Section 3. In Section 4, we experiment on synthetic data, and two real world networks. We conclude our paper in Section 5.

2. Modeling Time Evolving Networks

Consider a temporal sequence of networks $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$ where each $\mathcal{G}^{(t)} \equiv \{\mathcal{V}, \mathcal{E}^{(t)}\}$ is the network observed at time t; and we assume that the set of network actors $\mathcal{V}$ is invariant over time ($N = |\mathcal{V}|$), but $\mathcal{E}^{(t)} \equiv \{e_{ij}^{(t)}\}_{i,j=1}^{N}$, the set of links between actors, is possibly transient and can evolve over time.

Our model for time evolving networks builds upon a mixed membership stochastic blockmodel (MMSB) for *static* networks (Airoldi et al., 2008). The key idea is to superimpose a state space model on top of the MMSB, and connect the two via a logistic normal prior, such that temporal dynamics of the networks are captured. We next introduce the MMSB for static networks.

2.1. Mixed membership stochastic blockmodel

In the mixed membership stochastic blockmodel, we assume that each actor $v_i \in V$ can undertake up to K different roles from a latent tomographic space; and its degrees of affiliation to these roles can be represented as a normalized vector π_i (of dimension K) drawn from a prior distribution $p(\pi)$. We also refer to π as the *mixed membership vector*.

We further assume that a directed interaction from actors v_i to v_j are instantiated stochastically according to a *compatibility function* $C(z_{i\rightarrow j}, z_{i\leftarrow j})$ over the roles, $z_{i\rightarrow j}$ and $z_{i\leftarrow j}$, undertaken by the actor-pair in question. $z_{i\rightarrow j} \sim p(z|\pi_i)$ denotes the latent role of actor v_i when it is to interact with v_j, and $z_{i\leftarrow j} \sim p(z|\pi_j)$ denotes the role of actor v_j when it is approached by v_i. Here $z_{i\rightarrow j}$ and $z_{i\leftarrow j}$ are unit indicator vectors (of dimension K); $z_{i\rightarrow j,k} = 1$ (or $z_{i\leftarrow j,k} = 1$) means v_i (or v_j) understakes role k in this particular interaction.

We will focus on compatibility functions of a bilinear form: $C(z, z') = z^\top B z'$, where $B \equiv \{\beta_{kl}\}_{k,l=1}^{K}$ is called a role compatibility matrix. Then a link, $e_{ij} \in \{0, 1\}$, from v_i to v_j is drawn from a Bernoulli distribution with parameter $C(z_{i\rightarrow j}, z_{i\leftarrow j})$. In summary, the mixed membership stochastic blockmodel (MMSB) posits the following generative process for links in a static network:

1. $\{\pi_i\}_{i=1}^{N} \sim p(\pi)$, sample a mixed membership vector for each actor v_i.
2. For each actor v_j that actor v_i possibly interacts with:
   - $z_{i\rightarrow j} \sim$ Multinomial$(z|\pi_i)$, sample a role indicator vector for donor v_i;
   - $z_{i\leftarrow j} \sim$ Multinomial$(z|\pi_j)$, sample a role indicator vector for receiver v_j;
   - $e_{ij} \sim$ Bernoulli$(e|z_{i\rightarrow j}^\top B z_{i\leftarrow j})$, sample a link indicator between actor v_i and v_j.

The generative model above defines a conditional probability distribution of the relations $\mathcal{E} = \{e_{ij}\}_{i,j=1}^{N}$ between N actors in a way that reflects naturally interpretable latent roles of the actors. Note that each vertex v_i is associated with not just one but a set of latent membership indicators $\{z_{i\rightarrow\cdot}, z_{\cdot\leftarrow i}\}$ (if the links are undirected, then we can ignore the asymmetry of "$\rightarrow$" and "$\leftarrow$"). Thus the semantic underpinning of each interaction between actors is captured by a pair of instantiated memberships unique to this interaction; the nature and strength of this interaction is controlled by the compatibility function determined by this pair of membership instantiations.

The role compatibility matrix B encodes the affinity between functional roles. If we want to model assortative relations where actors of the same role are more likely to connect to each other, we can use a B with diagonally dominant entries; If we want to model differential preference between different roles, we can encode stronger off-diagonal entries and rich block patterns in B. The flexibility of choosing this B matrix provides the MMSB with strong expressive power for dealing with complex relational patterns. If necessary, a prior distribution, $\beta_{kl} \sim p(\beta)$, over the elements of B can be introduced, which can offer desirable smoothing or regularization effects.

Crucial to the MMSB, is the *mixed membership vector* π of in the above generative model, which represents the overall function spectrum of an actor and succinctly captures the probabilities of an actor involving in different roles when it interacts with others. The expressive power of the MMSB also arises from the choice of the prior distribution $p(\pi)$ for the mixed membership vector π. For instance, Airoldi et al. (2008) employed a simple Dirichlet prior because it is conjugate to the multinomial distribution generating the actual role indicator z. To capture non-trivial correlations

between different roles (*i.e.* different dimensions of π), and to model temporal dynamics of the actor roles, we can also employ a logistic normal distribution over a simplex as in Ahmed and Xing (2007):

$$\pi_k = \exp(\gamma_k - C(\gamma)), \quad \gamma \sim \text{Normal}(\mu, \Sigma) \qquad (1)$$

where $C(\gamma) = \log(\sum_{k=1}^{K} \exp(\gamma_k))$ is a normalization constant that insures entries of π sum to 1.

2.2. Dynamic logistic normal mixed membership stochastic blockmodel

In time evolving networks the mixed membership vectors $\pi^{(t)}$ can change over time; so are the priors, $p^{(t)}(\pi)$ and $p^{(t)}(\beta)$, and the role compatibility matrix $B^{(t)}$. Conditioning on the observed network sequence $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$, our goal is to infer the trajectories of the mixed membership vectors $\pi^{(t)}$ in the latent function (or role) space. In the following, we present a generative model that augments the MMSB with a state space model for linear dynamic systems for modeling time evolving networks.

We can superimpose temporal dynamics on both the prior, $p^{(t)}(\pi)$, for the mixed membership vectors, and the prior, $p^{(t)}(\beta)$, for entries of the role compatibility matrix. Specifically, we use a logistic normal distribution $\mathcal{LN}(\mu^{(t)}, \Sigma^{(t)})$ for $p^{(t)}(\pi)$, and another logistic normal distribution $\mathcal{LN}(\eta^{(t)}, S^{(t)})$ for $p^{(t)}(\beta)$[1]. We assume that the two means, $\mu^{(t)}$ and $\eta^{(t)}$, are evolving over time according to a linear Gaussian model, but the covariances, $\Sigma^{(t)}$ and $S^{(t)}$, which captures time specific topic correlations are independent across time.

The way we model temporal dynamics is based on the well-known state space model (SSM) popular in object tracking and trajectory modeling. It defines a linear transition between states of adjacent time points: for $\mu^{(t)}$, we have $\mu^{(t)} = A\mu^{(t-1)} + w^{(t)}$ where A is a transition matrix, $w^{(t)} \sim \mathcal{N}(0, \Phi)$ a normal transition noise, and $\mu^{(0)} := \nu$; and similarly for $\eta^{(t)}$, $\eta^{(t)} = b\eta^{(t-1)} + \xi^{(t)}$ where b is a scalar, $\xi^{(t)} \sim \mathcal{N}(0, \psi)$, and $\eta^{(0)} := \iota$. Now the MMSB model functions as an emission model within SSM, and it is connected to the SSM via the logistic normal priors $\mathcal{LN}(\mu^{(t)}, \Sigma^{(t)})$ and $\mathcal{LN}(\eta^{(t)}, S^{(t)})$. With such a construction, our model is able to track the functional changes of network actors, and sense the emergence and termination of "functional themes" in time evolving networks.

Overall, our dynamic mixed membership stochastic blockmodel (dMMSB) consists of three components: an SSM for mixed membership vector, an SSM for role compatibility matrix and a logistic normal mixed membership stochastic blockmodel (LNMMSB) for the networks. The first two components model the temporal dynamics while the third component model the generative process of the network at each time point. The following is an outline of the generative process of our model. A graphical model representation of this model is illustrated in Figure 1.

1. **State Space Model for Mixed Membership Vectors**:
   - $\mu^{(1)} \sim \text{Normal}(\nu, \Phi)$, sample the mean of the mixed membership prior at $t = 1$;
   - $\mu^{(t)} \sim \text{Normal}(A\mu^{(t-1)}, \Phi)$, sample the mean of the mixed membership prior for $t > 1$;
2. **State Space Model for Role Compatibility Matrix**:
   - $\eta^{(1)} \sim \text{Normal}(\iota, \psi)$, sample the mean of the role compatibility matrix prior at $t = 1$;
   - $\eta^{(t)} \sim \text{Normal}(b\eta^{(t-1)}, \psi)$, sample the mean of the role compatibility matrix prior for $t > 1$;
   - $\{\beta_{kl}^{(t)}\}_{k,l=1}^{K} \sim \mathcal{LN}(\eta^{(t)}, S^{(t)})$, sample entries of the role compatibility matrix.
3. **Logistic Normal Mixed Membership Stochastic Blockmodel (LNMMSB)**:
   - $\{\pi_i^{(t)}\}_{i=1}^{N} \sim \mathcal{LN}(\mu^{(t)}, \Sigma^{(t)})$, sample a mixed membership vector for each actor v_i at each time point t.
   - For each actor v_j that actor v_i possibly interacts with, at each time point t:
     - $z_{i\to j}^{(t)} \sim \text{Multinomial}(z|\pi_i^{(t)})$, sample a role indicator vector for donor v_i;
     - $z_{i\leftarrow j}^{(t)} \sim \text{Multinomial}(z|\pi_j^{(t)})$, sample a role indicator vector for receiver v_j;
     - $e_{ij}^{(t)} \sim \text{Bernoulli}(e|z_{i\to j}^{(t)\top} B^{(t)} z_{i\leftarrow j}^{(t)})$, sample a link indicator between actor v_i and v_j.

Note that one can also introduce dynamics directly on the (pre-transformed) mixed membership vectors γ rather than their prior. However, this leads to an SSM for each network actor and dramatically increase the model size. Alternatively, one can also design an intermediate model by introducing SSMs on clusters of actors; there is a tradeoff between model expressiveness and computational complexity. For clarity, we focus on the model with dynamics only on the priors.

Another thing to note is the state space model for the mixed membership vectors generates not a single, but N emissions each time, each corresponding to a (pre-transformed) mixed membership vector $\gamma_i^{(t)}$. In order to apply directly the Kalman filter and Rauch-Tung-Striebel smoother for posterior inference and parameter estimation, we introduce an intermediate random variable $Y^{(t)} = \frac{1}{N}\sum_i \gamma_i^{(t)}$; it is easy to see that $Y^{(t)}$ follows a standard SSM reparameterized from the orig-

[1]Note that β is a scalar, the logistic normal distribution is as follows: $\alpha \sim \text{Normal}(\eta, S), \beta = \exp(\alpha)/(\exp(\alpha) + 1)$.

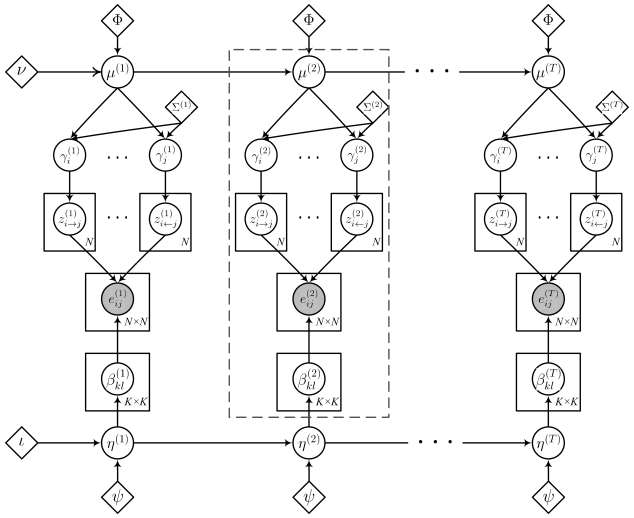

Figure 1. A graphical model representation of the dynamic mixed membership stochastic blockmodel. Enclosed by the dotted lines is a logistic normal MMSB.

inal dMMSB, *i.e.* $Y^{(t)} \sim \text{Normal}(\mu^{(t)}, \frac{\Sigma^{(t)}}{N})$.

In principle, we can use the dMMSB to capture not only membership correlation within and between actors at a specific time (as did in Blei and Lafferty (2006a)), but also dynamic coupling (or co-evolution) of the overall functional roles in the whole network via the covariance matrix Φ. In the simplest scenario, when $A = I$ and $\Phi = \sigma I$, this model reduces to random walk in the latent role space. In most realistic time evolving networks, both the role coupling and role compatibility matrix are unlikely to be invariant over time, we expect that even a random walk mixed membership stochastic blockmodel can provide a better fit to the data than a static model that ignores the time stamps of the networks.

3. Learning and Inference for DMMSB

Under dMMSB, exact posterior inference of the latent variables of interest (z, π and γ), and direct EM estimation of the model parameters (μ, Σ and B) are computationally intractable. This is due to difficulties in both the marginalization over the latent variables z[2], and in the integration of the mixed membership vector π[3]. In this section, we present a variational EM algorithm for inferring the latent variables and estimating the model parameters. Our algorithm is based on a generalized means field (GMF) approximation proposed by Xing et al. (2003).

[2]The state space of all z is super-exponential.

[3]No closed-form integration exists under a logistic normal prior.

Generalized mean field approximates the joint posterior $p(\{z^{(t)}, \gamma^{(t)}, \mu^{(t)}, B^{(t)}\}_{t=1}^T | \Theta, \{\mathcal{G}^{(t)}\}_{t=1}^T)$[4] by a product of three marginals, $q = q_1\big(\{z^{(t)}, \gamma^{(t)}\}_{t=1}^T\big)\ q_2\big(\{\mu^{(t)}\}_{t=1}^T\big)\ q_3\big(\{B^{(t)}\}_{t=1}^T\big)$. q_1 corresponds to the marginal distribution of $\{z^{(t)}, \gamma^{(t)}\}_{t=1}^T$ under a reparameterized LNMMSB; q_2 and q_3 corresponds to SSMs over $\{\mu^{(t)}\}_{t=1}^T$ and $\{B^{(t)}\}_{t=1}^T$ respectively, with emission models related to the expectation of $\{z^{(t)}, \gamma^{(t)}\}_{t=1}^T$ under q_1[5].

The computation of the variational parameters for the approximate marginal q_1, q_2 and q_3 leads to the coupling between them (more details in the following sections). Once the variational parameters are obtained, inference on any latent variable under the joint distribution p which is intractable, can be approximated by a much simpler inference in one of the q_is that contains the variable of interest.

For simplicity, we drop the SSM superimposed in the role compatibility matrix B, and assume that it is invariant over time. Thus we do not need to consider the parameters η and S in our exposition. Furthermore, we focus on random walk dynamics and hence set the state transition matrix A to I. In the following sections, we will discuss two key components of our learning and inference algorithm in more details.

3.1. Variational inference for LNMMSB

Under a LNMMSB, the posterior of the latent variables γ and z can be written as:

$$p(\gamma, z | \mu, \Sigma, B, \mathcal{G}) \propto \prod_i p(\gamma_i | \mu, \Sigma) \prod_{i,j} p(z_{i\to j}, z_{i\leftarrow j} | \gamma_i, \gamma_j) p(e_{ij} | z_{i\to j}, z_{i\leftarrow j}, B). \quad (2)$$

Directly marginalizing over all but one hidden variables, say γ_i, is intractable in the above posterior. Again, we resort to generalized mean field (GMF) for apprxoimating p with $q_\gamma q_z$. Xing et al. (2003) showed that under GMF approximation, the optimal solution to the marginal over a cluster of variables is isomorphic to the conditional distribution of this cluster given its *expected Markov Blanket*. That is,

$$q_\gamma(\gamma_i) = p(\gamma_i | \mu, \Sigma, \langle z_{i\to\cdot} \rangle_{q_z}, \langle z_{\cdot\leftarrow i} \rangle_{q_z}), \text{ and}$$
$$q_z(z_{i\to j}, z_{i\leftarrow j}) = p(z_{i\to j}, z_{i\leftarrow j} \mid e_{ij}, B, \langle \gamma_i \rangle_{q_\gamma}, \langle \gamma_j \rangle_{q_\gamma}),$$

where $\langle \cdot \rangle_q := \mathbb{E}_q[\cdot]$. These equations define a fixed point for q_γ and q_z. To obtain this fixed point, the marginal for variables in one cluster is updated when

[4]Θ denotes other model parameters.

[5]This can be shown by minimizing the KL-divergence between p and q over arbitrary choices of q_1, q_2 and q_3 (Xing et al., 2003).

we fix the marginals of all the other variables. Such iterative updates continue until convergence.

More specifically, the update formulae for $q_\gamma(\gamma_i)$ and $q_z(z_{i\to j}, z_{i\leftarrow j})$ are as follows:

$$q_\gamma(\gamma_i) \propto \text{Normal}(\tilde{\gamma}_i, \tilde{\Sigma}_i), \quad (3)$$

$$q_z(z_{i\to j}, z_{i\leftarrow j}) \propto p(z_{i\to j}|\langle\gamma_i\rangle_{q_\gamma})\; p(z_{i\leftarrow j}|\langle\gamma_j\rangle_{q_\gamma})$$

$$p(z_{i\to j}, z_{i\leftarrow j}, B) \sim \text{Multinomial}(\delta_{(ij)}), \quad (4)$$

where $\tilde{\gamma}_i$ and $\tilde{\Sigma}_i$ are obtained via a Laplace approximation; $\delta_{(ij)kl} = \frac{1}{Z}\exp(\langle\gamma_{ik}\rangle_{q_\gamma} + \langle\gamma_{jl}\rangle_{q_\gamma})\beta_{kl}^{e_{ij}}(1-\beta_{kl})^{1-e_{ij}}$, and Z is a normalization constant.

3.2. Parameter estimation for dMMSB

We use variational EM algorithm for updating the parameters (Ghahramani & Beal, 2001). The E-step uses Kalman Filter and Rauch-Tung-Striebel (RTS) Smoother for estimating the hidden states $\mu_{t|T}$, and the LNMMSB update in section 3.1 for computing sufficient statistics $\tilde{\gamma}_i^{(t)}$, $\tilde{\Sigma}_i^{(t)}$ and $\delta_{(ij)uv}^{(t)}$. In the M-step, the model parameters are updated by maximizing the log-likelihood obtained from the E-step. The update formulae are as follows:

$$\beta_{kl} \leftarrow \frac{\sum_{t,i,j} e_{ij}^{(t)} \delta_{(ij)kl}^{(t)}}{\sum_{t,i,j} \delta_{(ij)kl}^{(t)}}, \quad \mu^{(t)} \leftarrow \mu_{t|T} \quad (5)$$

$$\hat{\Sigma}^{(t)} \leftarrow \sum\nolimits_i (\mu_{t|T} - \tilde{\gamma}_i^{(t)})^2 + \sum\nolimits_i \tilde{\Sigma}_i^{(t)}, \quad (6)$$

The overall algorithm is summarized in Algorithm 1. Note that the variational cluster marginals $q(z_{i\to j}, z_{i\leftarrow j})$, $q(\gamma_i)$ and $q(\{\mu^{(t)}\}_{t=1}^N)$ each dependent on variational parameters defined by other cluster marginals. Thus the overall algorithm is essentially a fixed point iteration that will converge to a local optimum. We use multiple random restarts to obtain a near global optimum.

Algorithm 1 Learning dMMSB

Input: a temporal sequence of networks $\{\mathcal{G}^{(t)}\}_{t=1}^T$.
Output: latent variables $z^{(t)}$ and $\gamma^{(t)}$, and model parameters $\mu^{(t)}$, $\Sigma^{(t)}$, $B = \{\beta_{kl}\}_{k,l=1}^K$.

Initialize $z^{(t)}$, $\gamma^{(t)}$ $\mu^{(t)}$, $\Sigma^{(t)}$ and B.
repeat
  repeat
    for $t = 1 \dots T$ **do**
      $q_z(z_{i\to j}^{(t)}, z_{i\leftarrow j}^{(t)}) \sim \text{Multinomial}(\delta_{(ij)}^{(t)})$.
      $q_\gamma(\gamma_i^{(t)}) \sim \text{Normal}(\tilde{\gamma}_i^{(t)}, \tilde{\Sigma}_i^{(t)})$.
    end for
    Update $\beta_{kl} \leftarrow \frac{\sum_{t,i,j} e_{ij}^{(t)} \delta_{(ij)kl}^{(t)}}{\sum_{t,i,j} \delta_{(ij)kl}^{(t)}}$.
  until convergence.
  Filtering plus smoothing $\mu_{t|T}$ using $\{Y^{(t)}\}_{t=1}^T$, update $\mu^{(t)} \leftarrow \mu_{t|T}$.
  Update $\Sigma^{(t)} \leftarrow \sum_i (\mu_{t|T} - \tilde{\gamma}_i^{(t)})^2 + \sum_i \tilde{\Sigma}_i^{(t)}$.
until convergence.

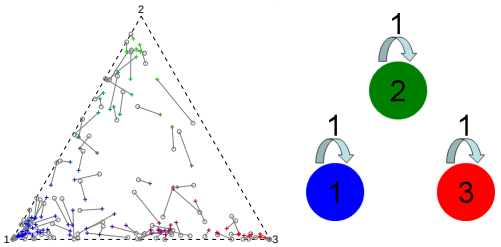

Figure 2. **Left**: the true mixed membership vectors (circle) and the estimates by LNMMSB (cross) visualized in a 2-simplex; each truth-estimate pair is connected by a grey line. **Right**: the true role compatibility matrix is the identity matrix. The estimate by LNMMSB is exactly the same as the ground truth and hence not displayed. Three roles are represented by blue, green and red respectively corresponding to the color of the vertices in the 2-simplex.

4. Experiments

We will first use synthetic data to investigate the performance of our learning algorithms and demonstrate the advantage of the dMMSB model. Then we will apply dMMSB to two real world datasets.

4.1. Experiments on synthetic data

We compare the logistic normal MMSB (LNMMSB) to a Dirichlet MMSB proposed by Airoldi et al. (2008), and then to dMMSB. We investigate their differences in two major aspects: (*i*) for a static network, does LNMMSB provides a better fit to the data when different roles are correlated? and (*ii*) for dynamic networks, does dMMSB provides a better fit to the data?

To investigate the first question, we generate a static network containing 100 actors and 3 latent roles. We use a covariance matrix of $4I$ between the latents roles. Furthermore, to introduce dependency structure between roles, we set the off-diagonal entries corresponding to role 2 and 3 to -3.9, which means that these two roles are negatively correlated. For one instantiation of the dataset, the true the mixed membership vectors are plotted as circles in Figure 2(left). It can be observed that very few actors have components for both role 2 and 3 which reflects the negative correlation we introduced in the covariance matrix. The role compatibility matrix is a identity matrix.

To evaluate the fitness of a model to the data, we use the log-likelihood of the model parameter as our mea-

Table 1. Dirichlet vs. Logistic Normal Prior for MMSB

Prior	Avg. ℓ_2 distance	Log-likelihood
Dirichlet	0.091	-5755.8
Logistic Normal	0.092	-5691.7

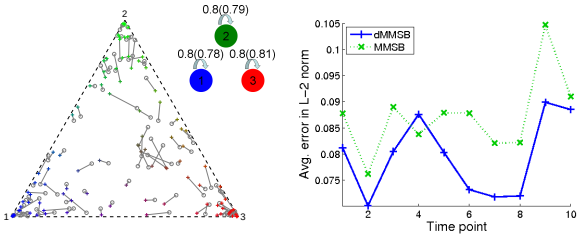

Figure 3. **Left**: the true mixed membership vectors (circle) and the estimates by dMMSB (cross) at time point 6 visualized in a 2-simplex; each truth-estimate pair is linked by a grey line. **Middle**: the learned role compatibility matrix, whose non-zero entries are shown by arcs with values; values outside the brackets are the truths and the values inside the brackets are estimates. **Right**: Average ℓ_2 errors of mixed membership vectors for MMSB and dMMSB.

sure. The results for LNMMSB and Dirichlet MMSB are listed in Table 1. We can see that the log-likelihood for LNMMSB is 64 smaller than that of the Dirichlet MMSB. Since no simple form of the log-likelihood can be derived for both methods, the log-likelihoods were obtained via importance sampling. As a second measure, we also used the ℓ_2 norm to measure the distance between the inferred mixed membership vectors and the ground truth. Both methods recovers the ground truth quite well (Table 1) and we illustrated the recovered mixed membership vectors in Figure 2.

To answer our second question, we generate dynamic networks consisting of 10 time points. The number of actors remains 100 and the number of roles remains 3. Furthermore, we generate the networks in such a way that networks between adjacent time points show certain degree of similarity. As an illustration, the true role compatibility matrix and the mixed membership vectors at time point 6 are displayed in Figure 3.

In Figure 3(right), we compare dMMSB to an LNMMSB learning a static network for each time point separately. We measure the performance in terms of the average ℓ_2 distance between the estimates of the mixed membership vectors and their true values. It can be seen that the error of dMMSB is lower than the error of MMSB in most cases and about 10 percent lower on average. This suggests that dMMSB can indeed integrate information across temporal domain and better models the networks. More settings of model parameters have been tested on both LNMMSB and dMMSB; they confirm that dMMSB is more effective in modeling dynamic networks.

4.2. Gene regulatory networks of yeast

In this section, we use our dMMSB as a visualization and exploration tool to study the dynamic gene regulatory networks of yeast collected by Luscombe et al. (2004). This dynamic network is built upon a static network assembled from known regulatory interactions from genetic, biochemical and ChIP-chip experiments. The dynamic aspect of the network is augmented using gene expression data. In our study, we focused on a subset of 116 genes and their dynamic networks over 5 stages of the cell cycle–early G1, late G1, S, G2, and M. We learned a dMMSB of 6 latent functional roles selected by BIC. The composition of functional roles from each gene is plotted in Figure 4, and the role compatibility matrix is visualized in Figure 5.

The functional roles we discovered for each gene correspond nicely to the structural importance of these genes. In particular, Role 1 (blue) corresponds to those genes that are inactive or disconnected from the rest of the regulatory network. For example, gene #8 is inactive in the 4 later cell stages.

Role 2 (light blue) and Role 3 (cyan) correspond to a pair of regulator-regulatee relation. These regulatees are active only for a few time points (e.g. gene #77-84 and #110-115). Moreover, the top 4 regulators in role 3 (gene #35, #41, #48 and #34) regulate a large number of genes in role 2, and the out-degree of these 4 regulators are much larger than other regulators in role 3. Therefore, we see these 4 genes are dominated by role 3 across time. They corresponds to essential genes needed across all stages of the cell cycle.

Interestingly, there is a second pair of regulator-regulatee relation between Role 4 (yellow) and Role 5 (orange). Note that this is different from the regulator-regulatee relation between Role 2 and Role 3. An example within this relation is gene #44 a persistent regulator assuming role 4, and gene #32 a target gene being regulated and assuming role 5.

Role 6 (dark red) corresponds to a cluster of genes with tight within-cluster interactions. They form cliques indicating that these genes are tightly interact with each to accomplish the underlying biological process. Gene #16 is one of this kind.

Another important feature is that genes may assume multiple functional roles and their functional roles may evolve as the cell proceeds to different stages of the cycle. For instance, #32 assumes role 3 and 5, which makes it sometimes a donor and sometimes a receiver. Another example is #9, which is of role 1 in time 2

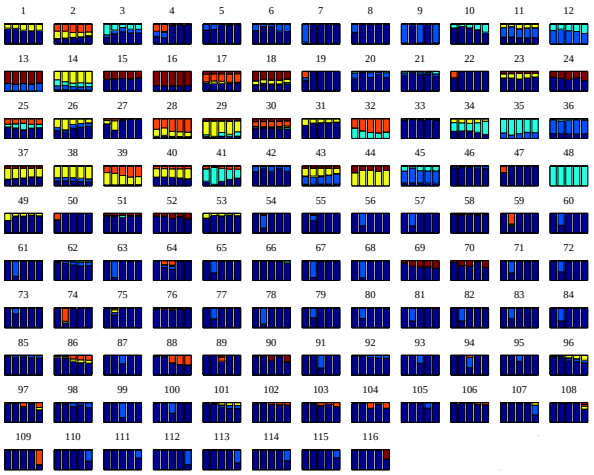

Figure 4. Temporal changes in mixed membership vectors grouped by individual. The horizontal axes of each subplot is time, and the vertical axes visualize the components of each mixed membership vector. Six roles are represented by six different colors.

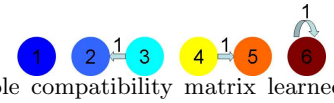

Figure 5. Role compatibility matrix learned by dMMSB. Each role is represented by a circle with the same color as in Figure 4. Each non-zero entry of the matrix corresponds to the weight on an edge.

and 4, but switches to role 2 in other times.

It is worth noting that although the role compatibility matrix is simple and separable (to components), a few actors that assume multiple active roles make the relationship network complicated and connected, which make it difficult for traditional methods to uncovering such structures.

4.3. Enron email networks

In this section, we study the Enron email communication networks. The email data was processed by Shetty and Adibi (2004). We further extract email sender and recipients in order to build email networks. We have processed the data such that numerous email aliases are properly corresponded to actual persons.

There are 151 persons in the dataset. We used emails from 2001, and built an email network for each month, so the dynamic network has 12 time points. We learn a dMMSB of 5 latent roles. The composition of roles of each member and the role compatibility matrix are depicted in Figure 6.

Similar to the discussion on gene networks, the first role (blue) stands for inactivity. The other roles are active. Actors with Role 2 (cyan) only send email to persons of the same role, therefore they form a clique. So is Role 4 (orange), which leads to another clique. Persons #6, 9, 48, 67, etc mainly assume this role, and they communicate with many others in the same role. They appear to be normal employees according to available information and the underlying meaning of the clique is yet discovered.

Role 5 (red) is within the functional composition of many people. Persons in Role 5 sends emails to persons with either Role 5 or Role 3 (green). They form a large clique, where Role 3 corresponds to receivers and Role 5 to both senders and receivers.

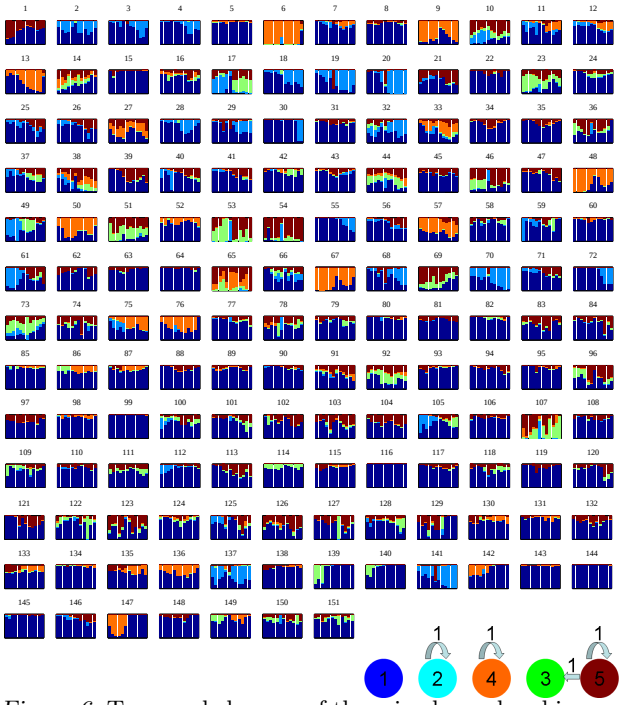

Figure 6. Temporal changes of the mixed membership vectors for each actor; and the visualization for role compatibility matrix.

Of special interest are individuals that are frequently dominated by multiple active roles (especially those falling into separate cliques), because they have strong connection with different groups and may serve important positions in the company. By scanning Figure 6, person #65 and #107 fit best to this category. According to external sources, *Mark Haedicke* (#65) was the Managing Director of the Legal Department, and *Louise Kitchen* (#107) was the President of Enron Online, which supports the finding of our method.

We also zoom into *Kenneth Lay* (#127), the Chairman and CEO of Enron at the time. His mixed membership vector in August is abnormally dominated by Role 3, which stands for a receiver. It is exactly the time when Enron's financial flaws were first publicly disclosed by an analyst, which might lead to a massive increase in

enquiry emails from the internal employees.

With respect to systematic changes in temporal space, mixed membership vectors of most actors are smooth over time. However, a few people experiences a large increase in the weight of the inactivity role in December (i.e., persons $\#6, 13, 36, 67, 76$). This is the time when Enron filed for bankruptcy.

Finally, we can also visualize the mixed membership vectors of the network entities and track the trajectory of the mixed membership vector for an individual as shown in Figure 7. They can help us understand the network as a whole and how each individual evolve in their roles. Based on these examples, we believe dMMSB can provide a useful visual portal for exploring the stories behind Enron.

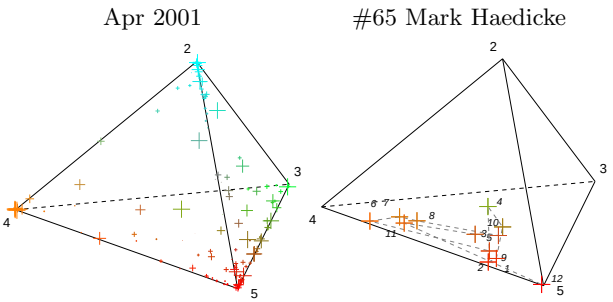

Figure 7. **Left**: visualization of mixed membership vectors of network actors in 3-simplex at one time point. Each vertex of the tetrahedron corresponds to a role marked by its ID. A mixed membership vector is represented by a cross whose location and color are the weighted average of its active roles and whose size is proportional to the sum of the weights from the active roles. **Right**: we track the trajectory of the mixed membership vector for an actor across time. Numbers in italics show time stamps.

5. Conclusion

DMMSB analysis for real datasets presented in the paper is merely a preliminary study. One can examine various properties of the mixed membership vectors along with the role compatibility matrix. Each finding should be verified through external sources; the model only serves as a tool for exploring dynamic networks.

The main purpose of dMMSB is to facilitate the process of uncovering interesting patterns underlying dynamic social or biological networks. In many cases, there can be hundreds and thousands of actors; examining them one by one can be extremely tedious. DMMSB offers an effective way for unveiling the block structure of the networks, which is very different from traditional analysis (*e.g.* degree distribution).

In term of the model and the algorithm, there are many dimensions where we can extend our current work. For instance, the current model does not explicitly take hubs of networks into account. Incorporating these elements will be interesting future research.

Acknowledgement

This work is under a support from NSF DBI-0546594 and DBI-0640543 awarded to EPX; LS is also supported by a Ray and Stephenie Lane Fellowship.

References

Ahmed, A., & Xing, E. P. (2007). On tight approximate inference of logistic-normal admixture model. *Artificial Intelligence and Statistics*.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed-membership stochastic blockmodels. *JMLR*, *9*, 1981–2014.

Blei, D., & Lafferty, J. (2006a). Correlated topic models. *Neural Information Processing Systems*.

Blei, D. M., & Lafferty, J. D. (2006b). Dynamic topic models. *Intl. Conf. Machine Learning*.

Erosheva, E. A., Fienberg, S. E., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *PNAS*, *97*, 11885–11892.

Ghahramani, Z., & Beal, M. (2001). Propagation algorithms for variational Bayesian learning. *Neural Information Processing Systems*.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *J. R. Statist. Soc. A*, *170*, 1–22.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. American Statistical Association*, *97*, 1090–1098.

Li, W., & McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. *Intl. Conf. Machine Learning*.

Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, *431*, 308–312.

Shetty, J., & Adibi, J. (2004). *The enron email dataset database schema and brief statistical report* (Technical Report). Information Sciences Institute, University of Southern California.

Xing, E. P., Jordan, M. I., & Russell, S. (2003). A generalized mean field algorithm for variational inference in exponential families. *Uncertainty in AI*.