Evolving Cluster Mixed-Membership Blockmodel for Time-Varying Networks

Qirong Ho
Carnegie Mellon University

Le Song
Carnegie Mellon University

Eric P. Xing
Carnegie Mellon University

Abstract

Time-evolving networks are a natural representation for dynamic social and biological interactions. While latent space models are gaining popularity in network modeling and analysis, previous works mostly ignore networks with temporal behavior and multi-modal actor roles. Furthermore, prior knowledge, such as division and grouping of social actors or biological specificity of molecular functions, has not been systematically exploited in network modeling. In this paper, we develop a network model featuring a state space mixture prior that tracks complex actor latent role changes through time. We provide a fast variational inference algorithm for learning our model, and validate it with simulations and held-out likelihood comparisons on real-world time-evolving networks. Finally, we demonstrate our model's utility as a network analysis tool, by applying it to United States Congress voting data.

1 INTRODUCTION

Social and biological systems can often be represented as a series of temporal networks over actors, and these networks may undergo systematic rewiring or experience large topological changes over time. The dynamics of these *time-evolving networks* pose many interesting questions. For instance, what are the roles played by these networked actors? How will these roles dictate the way two actors interact? How do actors play multiple roles (multi-functionality) in different social and biological contexts, and how does an actor's set of roles evolve over time? Knowledge of actor roles provides insight into how social or biological communities form in networks. In particular, we might elucidate how actors with diverse role compositions group together, and how these groupings change over time.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

There is increasing interest in employing latent space models for network analysis (Hoff, Raftery, and Handcock 2002; Handcock, Raftery, and Tantrum 2007). However, most of these models assume static networks and a single, fixed role for each actor. Hence they cannot model actor multi-functionality and role evolution over time, making them unsuitable for analyzing complex temporal networks. Airoldi *et al.* (2008) proposed a Mixed Membership Stochastic Blockmodel (MMSB) that captures actor multi-functionality, but it applies to static data only.

Recently, Xing *et al.* (2010) have addressed temporal evolution in networks with a dynamic extension of MMSB, which they call dMMSB. The dMMSB places a time-evolving, unimodal prior on all network actors; specifically, it employs a time-evolving logistic normal distribution similar to a state-space model. Although an important first step towards dynamic network analysis, dMMSB offers very weak modeling power — because it employs a unimodal logistic normal for the role distribution of all actors, it is only applicable to networks where the multi-functionalities of all actors follow similar, unimodal dynamics. A direct solution might be to introduce a separate dynamic process for each actor, but not only is this computationally impractical for large networks with many actors, it is also statistically unsatisfactory from a Bayesian standpoint as the actors no longer share any common pattern and coupling, leaving the model prone to over-fitting and unable to support activity and anomaly detection.

This challenge naturally leads us to explore "evolving clusters" of actors — by modeling dynamic processes on clusters, rather than on individuals or on the whole network, we can increase inferential power while retaining a common, yet much more expressive multimodal mixture model prior, for each actor. With such a prior, we can accommodate the actors' potentially non-stationary and heterogeneous behaviors.

Thus, in order to model both the temporal evolution and multi-modal nature of networks, we propose an evolving cluster of mixed membership stochastic blockmodels. Our model employs the vanilla MMSB as the basic building block, but augments it with a multi-modal mixture prior to capture both the multi-functionality and the multi-modality of the actor trajectories. We conjoin the mixture MMSB

with a set of state space models, one over each mixture component, allowing the model to follow as many trajectories as there are mixture components. Each state space trajectory corresponds to the average evolution of the multi-functionality of a *group* of actors.

This evolving mixture prior over vanilla MMSB presents additional challenges to parameter learning and latent variable inference. We overcome these difficulties by developing a variational EM algorithm inspired by ideas from Ghahramani & Hinton (2000) and dMMSB (Xing, Fu, and Song 2010). Our algorithm performs approximate inference and learning efficiently. Moreover, it is fundamentally different from dMMSB's algorithm — the latter's M-step equations lack second order moments found in ours.

In our experiments, we validate our model on synthetic data, and compare our held-out likelihood on real networks to that of dMMSB. Finally, we analyze voting data from the United States Congress using our model.

2 TIME-EVOLVING NETWORK MODEL

We condsider a sequence of interaction networks or graphs, denoted by $\{\mathcal{G}^{(t)}\}_{t=1}^{T}$, where each $\mathcal{G}^{(t)} \equiv \{\mathcal{V}, \mathcal{E}^{(t)}\}$ represents the network observed at time t. We assume the set of actors $\mathcal{V} = \{1, \ldots, N\}$ is constant. Furthermore, we permit $\mathcal{E}^{(t)} \equiv \{e_{ij}^{(t)}\}_{i,j=1}^{N,N}$, the set of interactions between actors, to evolve with time. We ignore self edges $e_{ii}^{(t)}$.

Our goal is to infer the underlying multi-functionalities and clusters that give rise to this network sequence. We approach this problem by extending the mixed membership stochastic blockmodel (MMSB) (Airoldi, Blei, Fienberg, and Xing 2008), a static network model. The idea is to place a time-evolving (i.e. dynamic) model on top of the MMSB, allowing it to account for temporally-evolving network dynamics. An earlier approach, the dynamic MMSB (dMMSB) (Xing, Fu, and Song 2010), used a single dynamic model to account for all network actors. Because dMMSB learns just one dynamic process for all actors' multi-functionalities, it is a poor statistical fit when the multi-functionalities follow a multimodal distribution. At the other extreme, one might contemplate placing a separate dynamic model on every actor, but then the multi-functionalities would no longer share a common prior.

We resolve these conflicting goals by generalizing the prior on actors to a *mixture* of time-evolving logistic normal distributions. This mixture prior is *multi-modal* and captures *correlations* between roles, allowing it to fit complex data densities that the unimodal Gaussian prior of dMMSB or the uncorrelated Dirichlet prior of MMSB cannot.

2.1 Mixed Membership Stochastic Blockmodel (MMSB)

We begin by describing the Mixed Membership Stochastic Blockmodel (Airoldi, Blei, Fienberg, and Xing 2008), which serves as the foundation for our model. The MMSB assumes that each actor $v_i \in \mathcal{V}$ possesses a latent mixture of K roles, which determine observed network interactions. This role mixture formalizes the notion of actor multi-functionality, and we denote it by a normalized $K \times 1$ vector π_i, referred to as a *mixed membership* or MM vector. We assume these vectors are drawn from some prior $p(\pi)$.

Given MM vectors π_i, π_j for actors i and j, the network edge e_{ij} is stochastically generated as follows: first, actor i (the *donor*) picks one role $z_{\rightarrow ij} \sim p(z|\pi_i)$ to *interact* with actor j. Next, actor j (the *receiver*) also picks one role $z_{\leftarrow ij} \sim p(z|\pi_j)$ to *receive the interaction* from i. Both $z_{\rightarrow ij}, z_{\leftarrow ij}$ are $K \times 1$ unit indicator vectors. Finally, the chosen roles of i, j determine the network interaction $e_{ij} \sim p(e|z_{\rightarrow ij}, z_{\leftarrow ij})$, where $e_{ij} \in \{0, 1\}$. The specific distributions over $z_{\rightarrow ij}, z_{\leftarrow ij}, e_{ij}$ are:

- $z_{\rightarrow ij} \sim \text{Multinomial}(\pi_i)$
- $z_{\leftarrow ij} \sim \text{Multinomial}(\pi_j)$
- $e_{ij} \sim \text{Bernoulli}(z_{\rightarrow ij}^{\top} B z_{\leftarrow ij})$

where B is a $K \times K$ *role compatibility matrix*. Intuitively, the bilinear form $z_{\rightarrow ij}^{\top} B z_{\leftarrow ij}$ selects a single element of B; the indicators $z_{\rightarrow ij}, z_{\leftarrow ij}$ behave like indices into B.

This generative model has two noteworthy features. First, observed relations $\mathcal{E}$ result from actor latent roles interacting. In the case of social networks, the latent roles are naturally interpretable as social functions, e.g. political party affiliations. Note that actor i's latent membership indicators $\{z_{\rightarrow i\cdot}, z_{\leftarrow \cdot i}\}$ are *unique to each interaction*; he/she may assume different roles for interacting with each actor.

Second, the role compatibility matrix B completely determines the affinity between latent roles. For example, a diagonally-dominant B signifies that actors of the same role are more likely to interact. Conversely, off-diagonal entries in B suggest interactions between actors of different roles. The MMSB's expressive power lies in its ability to control the interaction strength between any pair of roles, by specifying the corresponding entries of B.

2.2 Mixture of MMSBs (M^{3}SB)

The actor MM prior $p(\pi)$ significantly affects MMSB's expressive power. Airoldi *et al.* originally used a Dirichlet prior in MMSB (2008), allowing their variational inference algorithm to exploit Dirichlet conjugacy with the multinomial role indicator distribution $p(z|\pi)$. Later, Xing *et al.* employed a *logistic normal prior* in dMMSB (2010) to capture correlations between roles, which the Dirichlet prior cannot. However, the logistic normal prior is unimodal and cannot fit complex, multi-modal data densities.

As a step towards our final model, we extend the MMSB by making $p(\pi)$ a *logistic normal mixture prior*:

- $c_i \sim \text{Multinomial}(\delta)$
- $\gamma_i | c_i \sim \text{Normal}(\mu_{c_i}, \Sigma_{c_i})$
- $\pi_i | c_i = \text{Logistic}(\gamma_i)$, $[\text{Logistic}(\gamma)]_k = \frac{\exp\{\gamma_k\}}{\sum_{l=1}^{K} \exp\{\gamma_l\}}$

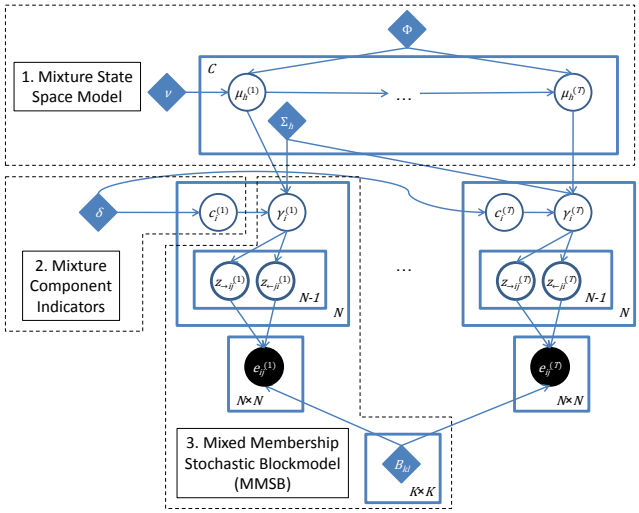

Figure 1: Graphical model representation of dM^{3}SB.

We call this model a mixture of MMSBs (M^{3}SB). c_i is a $C \times 1$ cluster selection indicator for π_i, where C is the number of mixture components. Thus, π_i is drawn from a logistic normal distribution with mean and covariance selected by c_i. c_i itself is drawn from a prior multinomial distribution δ.

Like dMMSB, the M^{3}SB accounts for role correlations using its logistic normal distribution. However, the M^{3}SB also has the flexibility to fit complex data densities, by virtue of its multi-modal mixture prior. In the sequel, we shall exploit this property to design a time-varying network model that tracks *cluster trajectories*, in contrast to dMMSB which tracks a single, average trajectory.

2.3 Dynamic M^{3}SB (dM^{3}SB)

In a time-evolving network, the MM vectors $\pi^{(t)}$ and their prior $p^{(t)}(\pi)$ change with time, and the goal now is to infer their dynamic trajectories. This enables detection of large-scale network trends, e.g. a group of actors whose MM vectors π shift from one set of roles to another. For example, if politicians change party affiliations; their MM vectors should exhibit a shift in political roles over time.

In order to model time-evolving networks, we place a state-space model on *every* logistic normal distribution in the mixture prior $p(\pi)$. In contrast, dMMSB only uses a single state-space model for its prior. Let N denote the number of actors, and T the number of time points in the evolving network. Also, let K denote the number of MMSB latent roles, and C the number of mixture components. We begin with an outline of our generative process; see Figure 1 for a graphical model representation.

1. **Mixture State Space Model for MM Vectors**
   - $\mu_h^{(1)} \sim \text{Normal}(\nu, \Phi)$ for $h = 1 \dots C$. Sample mixture means for the MM prior at $t = 1$.
   - $\mu_h^{(t)} \sim \text{Normal}(\mu^{(t-1)}, \Phi)$ for $h = 1 \dots C$, $t = 1 \dots T$. Sample mixture means for $t > 1$.
2. **Mixture Component Indicators**
   - $\{c_i^{(t)}\}_{i=1}^N \sim \text{Multinomial}(\delta)$ for $t = 1 \dots T$. Sample mixture indicator for each MM vector.
3. **Mixed Membership Stochastic Blockmodel**
   - $\{\gamma_i^{(t)}\}_{i=1}^N \sim \text{Normal}(\mu_{c_i^{(t)}}^{(t)}, \Sigma_{c_i^{(t)}})$ for $t = 1 \dots T$. Sample untransformed MM vectors *according to the mixture indicated by* $c_i^{(t)}$.
   - $\pi_i^{(t)} = \text{Logistic}(\gamma_i^{(t)})$, $[\text{Logistic}(\gamma)]_k = \frac{\exp\{\gamma_k\}}{\sum_{l=1}^K \exp\{\gamma_l\}}$. Logistic transform $\gamma_i^{(t)}$ into MM vector $\pi_i^{(t)}$.
   - For every actor pair $(i, j \neq i)$ and every time point $t = 1 \dots T$:
     - $z_{\rightarrow ij}^{(t)} \sim \text{Multinomial}(\pi_i^{(t)})$. Sample role indicator for the donor i.
     - $z_{\leftarrow ij}^{(t)} \sim \text{Multinomial}(\pi_j^{(t)})$. Sample role indicator for the receiver j.
     - $e_{ij}^{(t)} \sim \text{Bernoulli}(z_{\rightarrow ij}^{(t)\top} B z_{\leftarrow ij}^{(t)})$. Sample the interaction between actors i, j.

We refer to this model as the dynamic Mixture of MMSBs (dM^{3}SB for short). The general idea is to apply the state space model (SSM) used in object tracking to the MMSB model. Specifically, the MMSB becomes the emission model to the SSM; a distinct MMSB model is "emitted" at each time point (Figure 1). Furthermore, the SSM contains C distinct trajectories μ_h, each modeling the mean trajectory for a subset of MM vectors $\pi_i^{(t)}$. The SSM has two parameters ν, Φ, representing the prior mean and variance of the C trajectories. Each trajectory evolves according to a linear transition model $\mu_h^{(t)} = A\mu_h^{(t-1)} + w_h^{(t)}$, where A is a transition matrix and $w_h^{(t)} \sim \text{Normal}(0, \Phi)$ is Gaussian transition noise. We assume A to be the identity matrix, which corresponds to random walk dynamics; generalization to arbitrary A is straightforward.

Each MM vector $\pi_i^{(t)}$ is then drawn from one of the C trajectories $\mu_h^{(t)}$. The choice of trajectory for $\pi_i^{(t)}$ is given by the indicator vector $c_i^{(t)}$, which is drawn from some prior. For simplicity, we have used a single multinomial prior with parameter δ for all $c_i^{(t)}$. Observe that $c_i^{(t)}$ *can change over time*, allowing actors to switch clusters if that would fit the data better. Given $c_i^{(t)}$, the MM vector $\pi_i^{(t)}$ is drawn according to $\mathcal{LN}(\mu_{c_i^{(t)}}^{(t)}, \Sigma_{c_i^{(t)}})$, where the variances $\Sigma_1, \dots, \Sigma_C$ are model parameters. $\mathcal{LN}$ denotes a logistic normal distribution, the result of applying a logistic transformation to a normal distribution.

Once $\{\pi_i^{(t)}\}_{i=1}^N$ have been drawn for some t, the remaining variables $z_{\rightarrow ij}^{(t)}, z_{\leftarrow ij}^{(t)}, e_{ij}^{(t)}$ follow the MMSB exactly. We assume the role compatibility B to be a model parameter, although we note that more sophisticated assumptions can be found in the literature, such as a state space model prior (Xing, Fu, and Song 2010).

Algorithm 1 Variational EM for dM^{3}SB

Input: temporal sequence of networks $\{\mathcal{G}^{(t)}\}_{t=1}^T$.
Output: variational distributions $q_z, q_\gamma, q_c, q_\mu$ and model parameters $B, \delta, \nu, \Phi, \{\Sigma_h\}_{h=1}^C$.
Initialize parameters $B, \delta, \nu, \Phi, \{\Sigma_h\}_{h=1}^C$.
Sample initial values for $\mu^{(t)}, \gamma^{(t)}, c^{(t)}$.
repeat
  repeat
    Update $q_z(z_{i\to j}^{(t)}, z_{i\leftarrow j}^{(t)})$ for all i, j, t.
    Update B.
    Update $q_\gamma(\gamma_i^{(t)})$ for all i, t.
  until convergence
  Update $q_\mu(\{\mu_h^{(t)}\}_{t,h=1}^{T,C})$.
  Update ν, Φ.
  Update $q_c(c_i^{(t)})$ for all i, t.
  Update $\delta, \{\Sigma_h\}_{h=1}^C$.
until convergence

3 dM^{3}SB INFERENCE AND LEARNING

Neither exact latent variable inference nor parameter learning are computationally tractable in dM^{3}SB. The mixture prior on $\pi_i^{(t)}$, a factorial Hidden Markov Model, presents the biggest difficulty — it is analytically un-integrable, its likelihood is subject to many local maxima, and it requires exponential time for exact inference. Moreover, its logistic normal distribution does not admit closed-form integration with the multinomial distribution of $z|\pi$. Finally, the space of possible discrete role indicators z is exponentially large in the number of actors N and time points T.

We address all these difficulties with a variational EM procedure (Ghahramani and Beal 2001) based on the generalized mean field (GMF) algorithm (Xing, Jordan, and Russell 2003), and using techniques from Ghahramani & Hinton (2000) and dMMSB (Xing, Fu, and Song 2010). Our algorithm simultaneously performs inference and learning for dM^{3}SB in a computationally-effective fashion.

3.1 Variational Inference

Let $\Theta = \{\nu, \Phi, \{\Sigma_h\}_{h=1}^C, \delta, B\}$ denote all model parameters. We approximate the joint posterior $p(\{z^{(t)}, \gamma^{(t)}, c^{(t)}, \{\mu_h^{(t)}\}_{h=1}^C\}_{t=1}^T \mid \{\mathcal{E}^{(t)}\}_{t=1}^T; \Theta)$ by a variational distribution over factored marginals,

$$q = q_\mu\left(\{\mu_h^{(t)}\}_{t,h}^{T,C}\right) \prod_{t,i=1}^{T,N} \left[q_\gamma(\gamma_i^{(t)}) q_c(c_i^{(t)}) \prod_{j=1}^{N} q_z(z_{\to ij}^{(t)}, z_{\leftarrow ij}^{(t)}) \right].$$

q_z, q_γ and q_c correspond to MMSB latent variables z, γ and mixture indicators c, while q_μ corresponds to the mixture of C SSMs over time. The idea is to approximate latent variable inference under p (intractable) with feasible inference under q. In particular, Ghahramani & Hinton (2000) have demonstrated that it is feasible to have one marginal q_μ over all μs.

The GMF algorithm maxmizes a lower bound on the marginal distribution $p(\{\mathcal{E}^{(t)}\}_{t=1}^T; \Theta)$ over arbitrary choices of $q_z, q_\gamma, q_c, q_\mu$. We use the GMF solutions to the qs as the E-step in our variational EM algorithm, and derive the M-step through direct maximization of Θ with respect to our variational lower bound. Under GMF, the optimal solution to a marginal $q(\mathbf{X})$ for some latent variable set $\mathbf{X}$ is $p(\mathbf{X}|\mathbf{Y}, \mathbb{E}_q[\phi(\mathcal{MB}_\mathbf{X})])$, the distribution of $\mathbf{X}$ conditioned on the observed variables $\mathbf{Y}$ and the *expected exponential family sufficient statistics* (under variational distribution q) of $\mathbf{X}$'s Markov Blanket variables (Xing, Jordan, and Russell 2003). Hence our E-step iteratively computes $q(\mathbf{X}) := p(\mathbf{X}|\{\mathcal{E}^{(t)}\}_{t=1}^T, \mathbb{E}_q[\phi(\mathcal{MB}_\mathbf{X})])$ for $\mathbf{X} = \{u_h^{(t)}\}_{t,h}^{T,C}$, $\gamma_i^{(t)}$, $c_i^{(t)}$ and $\{z_{\to ij}^{(t)}, z_{\leftarrow ij}^{(t)}\}$. For brevity, we present only the final E-step equations; exact derivations can be found in the Supplemental.

E-step for q_z: From here, we drop time indices t whenever appropriate. q_z is a categorical distribution over K^2 elements,

$$q_z(z_{\to ij} = k, z_{\leftarrow ij} = l) \sim \text{Multinomial}(\omega_{(ij)}) \tag{1}$$
$$\omega_{(ij)kl} \propto (B_{kl})^{e_{ij}} (1 - B_{kl})^{1-e_{ij}} \exp(\langle\gamma_{ik}\rangle + \langle\gamma_{jl}\rangle)$$

where $\omega_{(ij)}$ is a normalized $K^2 \times 1$ vector indexed[1] by (k, l). The notation $\langle X\rangle$ denotes the expectation of X under q; for example, the expectations of z under q_z are $\langle z_{(\to ij)k}\rangle := \sum_l \omega_{(ij)kl}$ and $\langle z_{(\leftarrow ij)l}\rangle := \sum_k \omega_{(ij)kl}$.

E-step for q_γ: q_γ does not have a closed form, because the logistic-normal distribution of γ is not conjugate to the multinomial distribution of z. We apply a Laplace approximation to q_γ, making it normally distributed (Xing, Fu, and Song 2010; Ahmed and Xing 2007). Define $\Psi(a, b, C) := \exp\{-\frac{1}{2}(a-b)^\top C^{-1}(a-b)\}$. The approximation to q_γ is

$$q_\gamma(\gamma_i) \propto \Psi(\gamma_i, \tau_i, \Lambda_i) \text{ where} \tag{2}$$
$$\Lambda_i = \left((2N-2) H_i + \sum_{h=1}^C \Sigma_h^{-1} \langle c_{ih}\rangle \right)^{-1},$$
$$\tau_i = u + \Lambda_i \{ \sum_{j\neq i}^N (\langle z_{\to ij}\rangle + \langle z_{\leftarrow ji}\rangle) - (2N-2)\,(g_i + H_i(u - \hat{\gamma_i}))\},$$
$$u = \left(\sum_{h=1}^C \Sigma_h^{-1} \langle c_{ih}\rangle \right)^{-1} \left(\sum_{h=1}^C \Sigma_h^{-1} \langle c_{ih}\rangle \langle \mu_h\rangle \right),$$

$\hat{\gamma_i}$ is a Taylor expansion point, and g_i and H_i are the gradient and Hessian of the vector-valued function $\log(\sum_{l=1}^K \exp\gamma_i)$ evaluated at $\gamma_i = \hat{\gamma_i}$. We set $\hat{\gamma_i}$ to $\langle\gamma_i\rangle$ from the previous E-step iteration, keeping the expansion point close to the current expectation of γ_i.

[1] k, l correspond to roles indicated by $z_{i\to j}, z_{i\leftarrow j}$.

E-step for q_c: q_c is discrete over C elements,

$$q_c(c_i = h) \propto \delta_h \left|\Sigma_h\right|^{-1/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma_h^{-1}\left(\langle\gamma_i\gamma_i^\top\rangle - \langle\mu_h\rangle\langle\gamma_i\rangle^\top - \langle\gamma_i\rangle\langle\mu_h\rangle^\top + \langle\mu_h\mu_h^\top\rangle\right)\right]\right\}$$

Note the dependency on second order moments $\langle\gamma_i\gamma_i^\top\rangle$ and $\langle\mu_h\mu_h^\top\rangle$. Since q_γ, q_μ are Gaussian, these moments are simple to compute.

E-step for q_μ: The GMF solution to q_μ factors across clusters h:

$$q_\mu\left(\{\mu_h^{(t)}\}_{t,h}^{T,C}\right) := \prod_{h=1}^{C} q_{\mu,h}\left(\{\mu_h^{(t)}\}_t^T\right) \quad \text{where} \tag{3}$$

$$q_{\mu,h}\left(\{\mu_h^{(t)}\}_t^T\right) \propto \Psi(\mu_h^{(1)}, \nu, \Phi)\mathrm{Ob}(1,h)\prod_{t=1}^{T}\Psi(\mu_h^{(t)}, \mu_h^{(t-1)}, \Phi)\mathrm{Ob}(t,h),$$

$$\mathrm{Ob}(t,h) := \Psi\left(\frac{\sum_{i=1}^N \langle c_{ih}^{(t)}\rangle\langle\gamma_i^{(t)}\rangle}{\sum_{i=1}^N\langle c_{ih}^{(t)}\rangle}, \mu_h^{(t)}, \frac{\Sigma_h}{\sum_{i=1}^N\langle c_{ih}^{(t)}\rangle}\right).$$

Notice that factor $q_{\mu,h}(\{\mu_h^{(t)}\}_t^T)$ resembles a state-space model for cluster h, with "observation probability" at time t proportional to $\mathrm{Ob}(h,t)$. Hence the mean and covariance of each μ can be efficiently computed using the Kalman Smoother algorithm.

3.2 Parameter Estimation (M-step)

Given GMF solutions to each q from our E-step, we take our variational lower bound on the log marginal likelihood, and maximize it jointly with respect to all parameters Θ (for details, refer to the Supplemental). Let $\mathbb{S}(A) := A + A^\top$. The parameter solutions are:

$$\hat{\beta}_{kl} := \frac{\sum_{t,i,j\neq i}^{T,N,N}\omega_{(ij)kl}^{(t)}e_{ij}^{(t)}}{\sum_{t,i,j\neq i}^{T,N,N}\omega_{(ij)kl}^{(t)}},\ \hat{\nu} := \sum_h^C \frac{\langle\mu_h^{(1)}\rangle}{C},\ \hat{\delta} := \sum_{t,i}^{T,N}\frac{\langle c_i^{(t)}\rangle}{TN}$$

$$\hat{\Phi} := \frac{1}{TC}\left[\sum_{h=1}^{C}\langle\mu_h^{(1)}\mu_h^{(1)\top}\rangle - \mathbb{S}\left(\langle\mu_h^{(1)}\rangle\hat{\nu}^\top\right) + \hat{\nu}\hat{\nu}^\top + \sum_{t=2}^{T}\langle\mu_h^{(t)}\mu_h^{(t)\top}\rangle - \mathbb{S}\left(\langle\mu_h^{(t)}\mu_h^{(t-1)\top}\rangle\right) + \langle\mu_h^{(t-1)}\mu_h^{(t-1)\top}\rangle\right]$$

$$\hat{\Sigma}_h := \frac{\sum_{t,i}^{T,N}\langle c_{ih}^{(t)}\rangle[\langle\gamma_i^{(t)}\gamma_i^{(t)\top}\rangle - \mathbb{S}(\langle\gamma_i^{(t)}\rangle\langle\mu_h^{(t)}\rangle^\top) + \langle\mu_h^{(t)}\mu_h^{(t)\top}\rangle]}{\sum_{t,i}^{T,N}\langle c_{ih}^{(t)}\rangle}.$$

In particular, our estimate of $\hat{\Sigma}_h$ contains second order moments of μ (full derivations are in the Supplemental). dMMSB's unimodal prior has a similar covariance parameter, but its M-step equation lacks the aforementioned moments (Xing, Fu, and Song 2010). That paper does not furnish the relevant derivations, so we cannot verify their equations.

Our full inference and learning algorithm is summarized in Algorithm 1. This algorithm interleaves the E-step and M-step equations, yielding a coordinate ascent algorithm in the space of variational and model parameters. The algorithm is guaranteed to converge to a local optimum in our variational lower bound, and we use multiple random restarts to approach the global optimum. Similar to Airoldi *et al.* (2008), we update q_z, q_γ and B more often for improved convergence. Note that each random restart can be run on a separate computational thread, making dM^{3}SB easily *parallelizable* and therefore highly *scalable*.

3.3 Suitability of the Variational Approximation

Given that our true model is multimodal, our variational approximation will only be useful if it also fits multimodal data. Historically, *naive* mean field approximations, such as used in latent space models such as MMSB (Airoldi, Blei, Fienberg, and Xing 2008) and the Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), approximate all latent variables with unimodal variational distributions.

Instead, we have employed a *structured* mean field approximation that approximates all μs with a single, multimodal switching state-space distribution $q_\mu()$, essentially a collection of C Kalman Filters. This ensures that the multimodal structure of the prior on the MM vectors $\gamma_i^{(t)}$ is not lost. Moreover, although each $q_\gamma(\gamma_i^{(t)})$ for a given i,t is a unimodal Gaussian, it can be fitted to any mode in $q_\mu()$, independently of $q_\gamma(\gamma_i^{(t)})$ for other i,t. This flexibility ensures the variational posterior over all $\gamma_i^{(t)}$s remains multimodal.

4 EXPERIMENTS

We now validate dM^{3}SB on synthetic and real-world data, showing that it improves over dMMSB (Xing, Fu, and Song 2010) in multiple respects. We then conduct a case study on a real-world dataset to demonstrate dM^{3}SB's capabilities.

In the experiments that follow, we ran our algorithm for 50 outer loop iterations per random restart, with 5 iterations per inner loop. We also fixed $\Phi = \mathbb{I}_K$ and $\delta = 1/C$ instead of running their M-steps, as the former yields more stable results. For the remaining parameters, we used their M-steps with the following initializations: $B_{kl} \sim \mathrm{Uniform}(0,1)$, $\Sigma_h = \mathbb{I}_K$. For ν, we initialized $\langle\mu_h^{(1)}\rangle \sim \mathrm{Uniform}([-1,1]^K)$ for all h and set ν to their average. The remaining variational parameters were initialized via the generative process.

4.1 Synthetic Evaluation

Xing *et al.* (2010) have established the advantages of a time-varying MMSB model (dMMSB) compared to naive MMSB. In particular, when the roles are correlated, the logistic-normal prior provides a better fit to the data than the Dirichlet prior. Moreover, for time-varying networks, dMMSB provides a better fit than disjoint MMSBs on every time point.

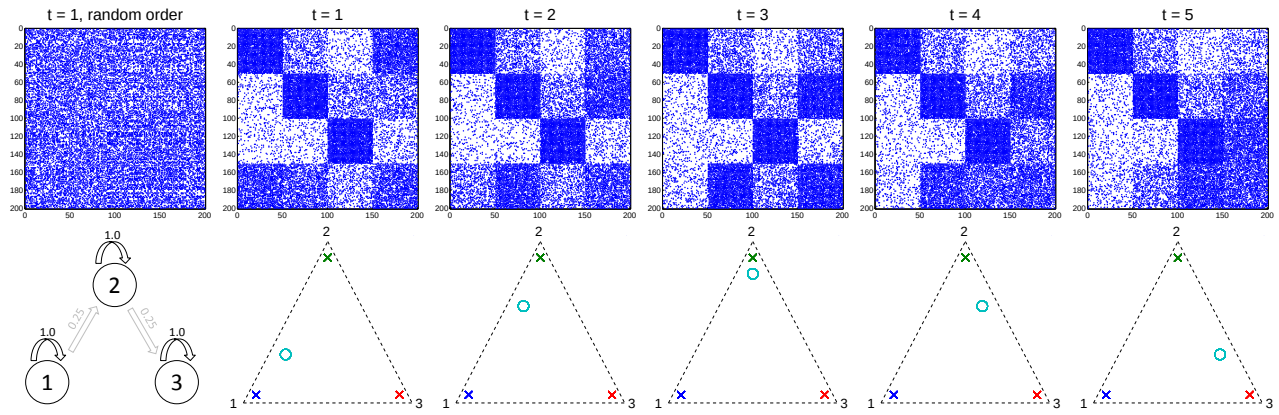

Figure 2: Synthetic data ground truth visualization. **Top Row:** Adjacency matrix visualizations, beginning on the left with $t = 1$ using random actor ordering, followed by $t = 1, \ldots, 5$ with actors grouped according to the ground truth. **Bottom left:** The role compatibility matrix B, shown as a graph. Circles represent roles, and numbered arrows represent interaction probabilities. **Bottom row:** True actor MM plots in the 3-role simplex for each t. Blue, green and red crosses denote the static MMs of the first 3 actor groups, and the cyan circle denotes the moving MM of the last actor group.

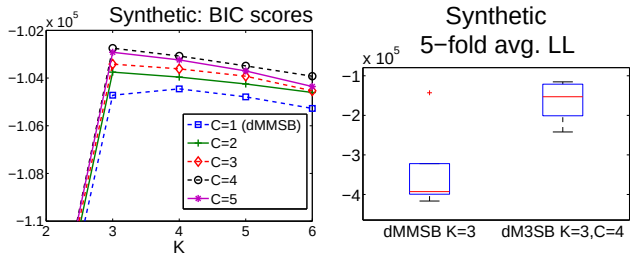

Figure 3: **Synthetic data:** BIC scores and 5-fold heldout log-likelihoods for dM^{3}SB and dMMSB.

In this experiment, we compare dM^{3}SB's performance to dMMSB, in terms of model fit (measured by the log marginal likelihood) and in terms of actor MM recovery. We generate data with $N = 200$ actors and $T = 5$ time points, and assume a $K = 3$ role compatibility matrix $B = (B_1, B_2, B_3)^\top$, with rows $B_1 = (1, .25, 0)$, $B_2 = (0, 1, .25)$, $B_3 = (0, 0, 1)$. The actors are divided into 4 groups of 50, with the first three groups having true MM vectors $(.9, .05, .05)$, $(.05, .9, .05)$ and $(.05, .05, .9)$ respectively, for all time points. The last group has MM vectors that move over time, according to the sequence $\pi^{(1)} = (.6, .3, .1)$, $\pi^{(2)} = (.3, .6, .1)$, $\pi^{(3)} = (.1, .8, .1)$, $\pi^{(4)} = (.1, .6, .3)$, $\pi^{(5)} = (.1, .3, .6)$. In Figure 2, we visualize our generated B, MM vectors π, and networks $\mathcal{E}^{(t)}$.

Thus far, we have not addressed model selection — specifically, selection of the number of roles K and the number of mixture components (clusters) C. To do so, we performed a gridsearch over $K \in \{2, 3, 4, 5, 6\}$ and $C \in \{1, 2, 3, 4, 5\}$ on the full network, using 200 random restarts per (K, C) combination. For all combinations, we observed convergence well within our limit of 50 outer iterations. Furthermore, completing all 200 restarts for each K, C took between 8 hours ($K = 2, C = 1$) and 28 hours ($K = 6, C = 5$) on a *single* processor. Since the random restarts can be run in parallel, with sufficient computing power one could easily scale dM^{3}SB to much larger time-varying networks with thousands of actors and tens of time points.

Table 1: **Synthetic data**: Estimation accuracy of dM^{3}SB ($K = 3, C = 4$) and dMMSB ($K = 3$).

dM^{3}SB role matrix B, Total Variation	0.1083
dMMSB role matrix B, Total Variation	0.0135
dM^{3}SB MMs $\pi_i^{(t)}$, mean ℓ_2 difference	0.0266
dMMSB MMs $\pi_i^{(t)}$, mean ℓ_2 difference	0.0477

For each (K, C) from the gridsearch, we selected its best random restart using the variational lower bound with a BIC penalty. The best restart BIC scores are plotted in Figure 3; note that dMMSB corresponds to the special case $C = 1$. The optimal BIC score selects the correct number of roles $K = 3$ and clusters $C = 4$, making it a good substitute for held-out model selection.

Next, using the BIC-optimal (K, C), we ran dM^{3}SB on a 5-fold heldout experiment. In each fold, we randomly partitioned the dataset's actors into two equal sets, and used the two corresponding subnetworks as training and test data. In each training fold, we selected the best model parameters Θ from 100 random restarts using the variational lower bound. We then estimated the log marginal likelihood for these parameters on the corresponding test fold, using Monte Carlo integration with 2,000 samples. This process was repeated for all 5 folds to get an average log marginal likelihood for dM^{3}SB . For comparison, we conducted the same heldout experiment for dMMSB set to K from the optimal (K, C) pair. The average log marginal likelihood for both methods is shown in Figure 3, and we see that dM^{3}SB's greater heldout likelihood makes it a better statistical fit to this synthetic dataset than dMMSB.

Finally, we compared dM^{3}SB to dMMSB in role estimation (B) and actor role recovery ($\pi_i^{(t)}$), using their best restarts on the correct (K, C) (or just K for dMMSB). Ta-

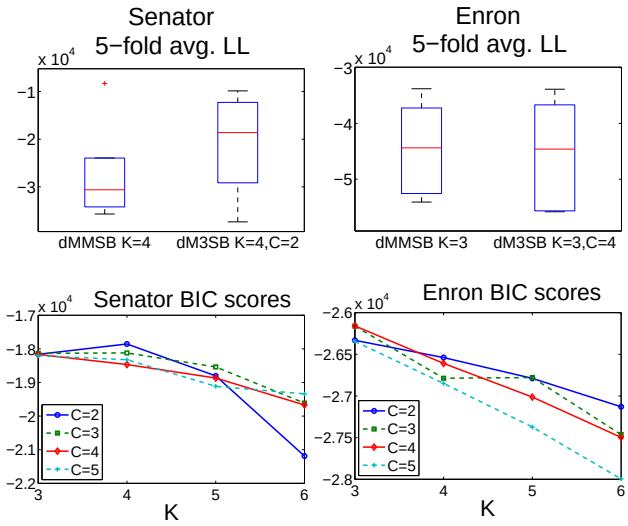

Figure 4: **Senator/Enron data:** BIC scores and 5-fold heldout log-likelihoods for dM^{3}SB and dMMSB.

ble 1 shows, for both methods versus the ground truth, the average ℓ_2 error in $\pi_i^{(t)}$ — specifically, we compared the ground truth to $\pi_i^{(t)}$'s posterior mean from either method — as well as the total variation in B. dM^{3}SB's average ℓ_2 error in $\pi_i^{(t)}$ is significantly lower than dMMSB's, at the cost of a higher total variation in B. However, dM^{3}SB's total variation of 0.1083 implies an average difference of only 0.012 in each of the 9 entries of B, which is already quite accurate. The fact that dM^{3}SB accurately recovers $\pi_i^{(t)}$ confirms that its posterior over all $\pi_i^{(t)}$ is multimodal, which validates our variational approximation.

We also note that dM^{3}SB's mean cluster trajectories $\langle \mu_h^{(t)} \rangle$ accurately estimated the four groups' mean MM vectors, with a maximum ℓ_2 error of 0.0761 for any group h and time t, except at $t = 5$ where dM^{3}SB exchanged group 3's trajectory with that of (moving) group 4.

4.2 Real Data Held-Out Comparisons

We now compare dM^{3}SB to dMMSB on two real-world data sets: a 151 actor subset of the Enron email communications dataset (Shetty and Adibi 2004) over the 12 months of 2001, and a 100 actor subset of the United States Congress voting data over the 8 quarters of 2005 and 2006 (described in the next section).

For both datasets, we first selected the optimal values of (K, C) via BIC score gridsearch with dM^{3}SB over $K \in \{3, 4, 5, 6\}, C \in \{2, 3, 4, 5\}$. Our previous synthetic experiment has demonstrated that model gridsearch using BIC produces good results. The optimal values were $K = 4, C = 2$ for the Senator dataset, and $K = 3, C = 4$ for the Enron dataset (Figure 4).

Using each dataset's optimal (K, C), we next ran dM^{3}SB on the 5-fold heldout experiment discussed in the previous section, obtaining average log marginal likelihoods. For comparison, we conducted the same heldout experiments for dMMSB set to K from the optimal (K, C) pair.

Plots of the heldout log marginal likelihoods for dM^{3}SB and dMMSB can be found in Figure 4. On the Senator dataset, dM^{3}SB has the higher log marginal likelihood, implying that it is a better statistical fit than dMMSB. For the Enron dataset, both methods have the same likelihood, showing that using dM^{3}SB with more mixture components at least incurs no statistical cost over dMMSB.

4.3 Case Study: US Congress Voting Data

We now apply dM^{3}SB to qualitatively analyze a real data set, the United States 109th Congress voting records. We shall show that dM^{3}SB not only recovers Mixed Membership (MM) vectors and a role-compatibility matrix that match our intuitive expectations of the data, but that the MM vectors are useful for identifying outliers and other unusual phenomena.

The Congress involved 100 senators and 542 bills spread over Jan 1st 2005 through Dec 31st 2006. The original voting data[2] is provided in the form of yes/no votes for each senator and each bill. In order to create a time-varying network suitable for dM^{3}SB, we applied the method of Kolar *et al.* to recreate their network result in (Kolar, Song, Ahmed, and Xing 2008).

The generated time-varying network contains 100 actors (senators), and 8 time points corresponding to 3-month epochs starting on Jan 1st 2005 and ending on Dec 31st 2006. The network is an undirected graph, where an edge between two senators indicates that their votes were mostly similar during that particular epoch. Conversely, a missing edge indicates that their votes were mostly different. Our intention is to discover how the political allegiances of different senators shifted from 2005 to 2006.

For our analysis, we used the optimal dM^{3}SB restart from the BIC gridsearch described in the previous held-out experiment. Recall that this optimal restart uses $K = 4$ roles and $C = 2$ clusters. The learned MM vectors π_i, compatibility matrix B, and most probable cluster assignments are summarized in Figure 5. The results are intuitive: Democratic party members have a high proportion of Role 1, while Republican party members have a high proportion of Role 2. Both Roles 1 and 2 interact exclusively with themselves, reflecting the tendency of both political parties to vote with their comrades and against the other party. The remaining two roles exhibit no interactions; senators with high proportions of these roles are unaligned and unlikely to vote with either political party. Observe that the two clusters perfectly capture party affiliations — Republican senators are almost always in cluster 1, while Democratic senators are almost always in cluster 2.

[2] Available at http://www.senate.gov

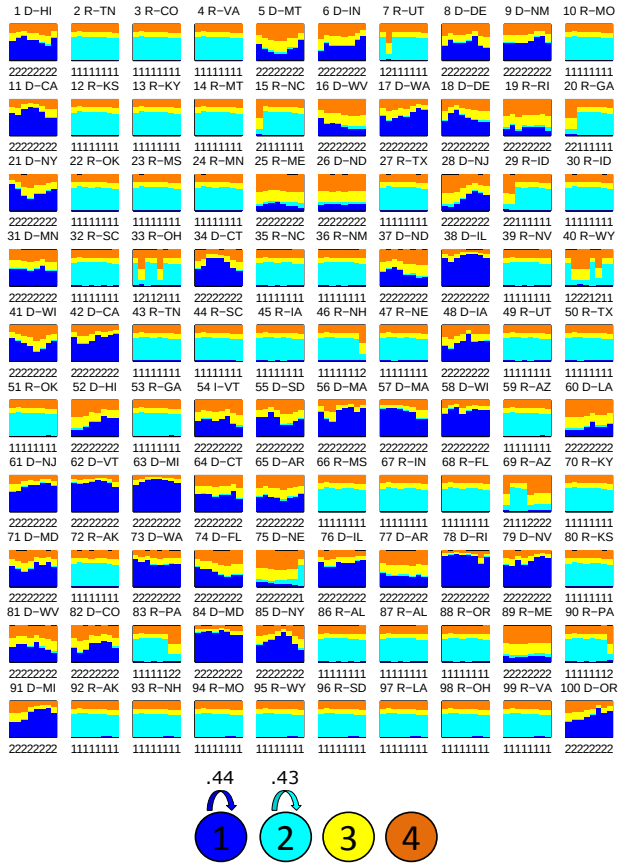

Figure 5: Congress voting network: Mixed membership vectors (colored bars) and most probable cluster assignments (numbers under bars) for all 100 senators, displayed as an 8-time-point series from left-to-right. The annotation beside a senator's number refers to that senator's political party (D for Democrat, R for Republican, I for Independent) and state (as a two-letter abbreviation). The learned role compatibility matrix is displayed at the bottom.

While it is reassuring to see results that reflect a contemporary understanding of US politics, the true value of dM^{3}SB's mixed-membership analysis lies in identifying outliers. For instance, consider the Democrat Ben Nelson (#75): from $t = 1$ through 7, his votes were unaligned with either Democrats or Republicans, though his votes were gradually shifting towards Republican. At $t = 8$ (end 2006), his voting becomes strongly Republican (Role 2), and he shifts from the Democrat cluster (1) to the Republican one (2). Ben Nelson's trajectory through the role simplex is plotted in Figure 6. Incidentally, Ben Nelson was reelected as the Senator from Nebraska in late 2006, winning a considerable percentage of his state's Republican vote.

Next, observe how the senator from New Jersey, #28, started off unaligned from $t = 1$ to 4 but ended up Democratic from $t = 5$ to 8; his role trajectory is also plotted in Figure 6. There is an interesting reason for this: the seat for New Jersey was occupied by two senators during the Congress, Jon Corzine in the first session ($t = 1$ to 4), and Bob Menendez in the second session ($t = 5$ to 8). Jon Corzine was known to have far-left views, reflected in

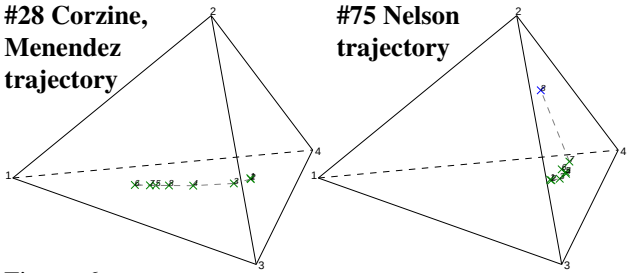

Figure 6: Congress voting network 3-simplex visualizations. Colors (green, blue) denote cluster membership. **Left:** MM vector time-trajectory for Senator #28 (D-NJ) — Jon Corzine during time points 1-4, and Bob Menendez during time points 5-8. **Right:** MM vector time-trajectory for Ben Nelson (#75, D-NE) .

#28's lack of both Republican *and* Democratic roles during his term (the Democrat role captures mainstream rather than extremist voting behavior). Once Bob Menendez took over, #28's behavior fell in line with most Democrats.

Other outliers include James Jeffords (#54), the sole Independent senator who votes like a Democrat, and three Republican senators with Democratic leanings: Lincoln Chafee #19, Susan Collins #25, and Olympia Snowe #89.

5 CONCLUSION

We have developed a probabilistic model, dM^{3}SB, for latent role analysis in time-varying networks, as well as an efficient variational EM algorithm for approximate inference and learning. Our model is distinguished by its explict modeling of actor multi-functionalities (role MMs), as well as its multimodal logistic normal mixture prior over these multi-functionalities. In particular, the latter separates dM^{3}SB from earlier models like dMMSB (Xing, Fu, and Song 2010) or MMSB (Airoldi, Blei, Fienberg, and Xing 2008), as it allows dM^{3}SB to fit complex latent role densities. This is validated by our experiments, as dM^{3}SB outperforms dMMSB on both synthetic tests and held-out experiments on real-world data. We have also demonstrated how dM^{3}SB can be used to explore actor latent roles, using the US Congress voting data as a case study.

Finally, we note that dM^{3}SB's variational inference algorithm is trivial to run in parallel, since each random restart can be run on a separate computational thread. We intend to explore larger time-varying datasets, such as gene networks, in other publication avenues.

Acknowledgements

This paper is based on work supported by NSF IIS-0713379, NSF DBI-0546594 (Career), ONR N000140910758, AFOSR FA9550010247, NIH 1R01GM093156, and an Alfred P. Sloan Research Fellowship to Eric P. Xing. Qirong Ho is supported by a graduate fellowship from the Agency for Science, Technology And Research, Singapore. Le Song is supported by a Stephenie and Ray Lane Fellowship.

References

Ahmed, A. and E. P. Xing (2007). On tight approximate inference of logistic-normal admixture model. In *AISTATS 2007*.

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

Blei, D. M., A. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

Ghahramani, Z. and M. Beal (2001). Propagation algorithms for variational bayesian learning. In *NIPS 2000*.

Ghahramani, Z. and G. E. Hinton (2000). Variational learning for switching state-space models. *Neural Computation 12*(4), 831–864.

Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model-based clustering for social networks. *J. R. Statist. Soc. A 170*(2), 1–22.

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

Kolar, M., L. Song, A. Ahmed, and E. P. Xing (2008). Estimating time-varying networks. To appear in to Annals of Applied Statistics, `arXiv:0812.5087 [stat.ML]`.

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

Xing, E. P., W. Fu, and L. Song (2010). A State-Space Mixed Membership Blockmodel for Dynamic Network Tomography. *Annals of Applied Statistics 4*(2), 535–566.

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