Statistica Sinica **32** (2022), 2147-2170
doi:https://doi.org/10.5705/ss.202020.0506

VARIATIONAL INFERENCE FOR LATENT SPACE MODELS FOR DYNAMIC NETWORKS

Yan Liu and Yuguo Chen

University of Illinois at Urbana-Champaign

Abstract: Latent space models are popular for analyzing dynamic network data. We propose a variational approach to estimate the model parameters and the latent positions of the nodes in the network. The proposed approach is much faster than Markov chain Monte Carlo algorithms, and is able to handle large networks. Theoretical properties of the variational Bayes risk of the proposed procedure are provided. We apply the variational method with the latent space model to simulated and real data to demonstrate its performance.

Key words and phrases: Bayes risk, dynamic network, latent space model, variational inference.

1. Introduction

Network data analysis has become an increasingly important research topic in various scientific disciplines in recent years. Most existing work on network data focuses on static networks, which means the inference is based on a static list of nodes and edges in an observed network at a given point in time (see Goldenberg et al. (2010) for a survey). However, the network structures of real-world systems are often time varying, or dynamic, in nature, with the set of nodes or the set of edges, or both, evolving over time. In this study, we focus on a time series of observed networks with the same set of nodes and a sequence of sets of edges observed at multiple time points. Analyzing such networks is crucial to understanding their dynamic aspect, such as how social relations and structures, gene-protein interactions, and co-authorship patterns evolve over time.

Many models for dynamic networks have been proposed in the literature. Some are extensions of existing static network models, including the dynamic versions of the stochastic blockmodel (SBM) (Yang et al. (2011); Xu and Hero (2014); Xu, Kliger and Hero III (2014); Xu (2015); Matias and Miele (2017)), degree-corrected stochastic blockmodel (Wilson, Stevens and Woodall (2019)), mixed-membership stochastic blockmodel (MMSB) (Fu, Song and Xing (2009);

Corresponding author: Yuguo Chen, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA. E-mail: yuguo@illinois.edu.

Xing, Fu and Song (2010); Ho, Song and Xing (2011)), exponential random graph model (Guo et al. (2007); Ahmed and Xing (2009); Hanneke, Fu and Xing (2010); Krivitsky and Handcock (2014); Lee, Xue and Hunter (2020)), latent factor model (Ward, Ahlquist and Rozenas (2013); Durante and Dunson (2014a,b); Hoff (2015)), latent feature relational model (Foulds et al. (2011); Heaukulani and Ghahramani (2013)), and latent space model (Sarkar and Moore (2005); Sewell and Chen (2015); Friel et al. (2016); Sewell and Chen (2017)). See Kim et al. (2018) for a review of dynamic network models with latent variables.

The latent space model embeds dynamic networks into a low-dimensional Euclidean space, and has the advantage of meaningful visualization and interpretation. The model has also been used for multilayer networks (Gollini and Murphy (2016); Durante, Dunson and Vogelstein (2017)). Although sometimes dynamic networks are treated as multilayer networks (Han, Xu and Airoldi (2015)), the temporal aspect of dynamic networks is not considered in multilayer networks.

The latent space model is a popular approach for modeling dynamic networks, but estimating the model parameters and latent positions is often computationally intensive. Sewell and Chen (2015) used a Markov chain Monte Carlo (MCMC) approach with a case-control approximate likelihood to reduce the computational cost, but it still has difficulties in handling large networks. As an alternative, the variational inference (VI) approach (Jordan et al. (1999); Wainwright and Jordan (2008)) is becoming increasingly popular in the statistics community; see Blei, Kucukelbir and McAuliffe (2017) for a comprehensive review. Daudin, Picard and Robin (2008) proposed a variational approach to estimate the parameters of the SBM, and this idea was later generalized by Mariadassou, Robin and Vacher (2010) to deal with valued graphs and possible covariates. Yang et al. (2011) and Matias and Miele (2017) designed variational expectation-maximization (EM) algorithms for the dynamic version of the SBM. Airoldi et al. (2008) used a variational approach to fit the MMSB, and Xing, Fu and Song (2010) and Ho, Song and Xing (2011) proposed variational EM algorithms for approximate inference for the dynamic version of the MMSB. Salter-Townshend and Murphy (2013) proposed a variational Bayes (VB) algorithm for a static latent space model with a community structure. Sewell and Chen (2017) proposed a VB algorithm for projection models in dynamic networks.

Despite the empirical success of variational algorithms in estimating the posterior distributions of the parameters of various network models, theoretical studies of such algorithms are limited. Some results for variational approaches have been obtained under the SBM (Celisse, Daudin and Pierre (2012); Bickel et al. (2013); Zhang and Zhou (2020)), but there are no theoretical results on variational

algorithms for latent space models. Recently, Yang, Pati and Bhattacharya (2020) proposed a more general class of VB algorithms, with the standard variational approximation as a special case. They also developed variational inequalities and convergence results on the Bayes risk of the proposed variational approximation. Their results indicate that the parameter estimates given by the variational algorithm converge to the true parameter values under certain conditions. This work provides a framework for analyzing the theoretical properties of VB algorithms for latent space models.

We consider a class of latent space models for dynamic networks, and propose a variational algorithm for performing posterior inference for large-scale networks. Furthermore, we show that the parameter estimation based on the variational posterior is consistent. The simulation studies demonstrate that the proposed variational algorithm is much faster than the MCMC algorithm, while still giving satisfactory results, so it is more suitable for large-scale dynamic networks. We also apply our method to analyze a friendship network from the "Teenage Friends and Lifestyle Study" and a Wiki-talk communication network.

The rest of the paper is organized as follows. Section 2 considers a class of latent space models for dynamic networks. Section 3 proposes a variational algorithm for posterior inference. Section 4 gives finite-sample upper bounds to the VB risk of the proposed procedure and shows the consistency of the parameter estimation. Sections 5 and 6 illustrate the performance of the proposed method using simulation studies and real-world examples, respectively. Section 7 concludes the paper with a discussion.

2. Dynamic Latent Space Model

Latent space models for network data are a popular class of models first proposed by Hoff, Raftery and Handcock (2002) for static networks, and later generalized by Sewell and Chen (2015) and Sewell and Chen (2016) to dynamic networks. This class of models embed the nodes of a network into an unobserved latent space, which can provide visualization and insight into the evolution of the network.

Suppose there are n nodes and T time steps in a dynamic network. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ be the observed $n \times n$ adjacency matrices at the T time steps, where $Y_{ijt} = 1$ if there is an edge from node i to node j at time t, and zero otherwise. Throughout this paper, the latent space we consider is the d-dimensional Euclidean space $\mathbb{R}^d$. Let $\boldsymbol{X}_{it} \in \mathbb{R}^d$ be the latent position of the ith node at time t, for $1 \leq i \leq n$ and $1 \leq t \leq T$. The dynamic latent space model rep-

resents the time series of networks using a hidden Markov model. The latent positions of the nodes (which are the hidden states) evolve independently of each other. The latent position of each node is modeled by a Markov process with the initial distribution $\boldsymbol{X}_{i1} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \mathbb{I}_d)$ $(i = 1, \ldots, n)$ and transition distribution $\boldsymbol{X}_{it}|\boldsymbol{X}_{i(t-1)} \sim \mathcal{N}(\boldsymbol{X}_{i(t-1)}, \tau^2 \mathbb{I}_d)$ $(t = 2, \ldots, T,\ i = 1, \ldots, n)$, where $\mathbb{I}_d$ is the $d \times d$ identity matrix.

The observed networks $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ are conditionally independent, given the latent positions across the time span. In addition, for the network $\boldsymbol{Y}_t$ at time t, we assume that all edges Y_{ijt} are formed independently, conditioning on the latent positions at time t. A general expression for the probability that there is an edge between nodes i and j can be written as $p(Y_{ijt} = 1|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \boldsymbol{\beta}) = h(||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||, \boldsymbol{\beta}), \quad 1 \leq i \neq j \leq n, \quad 1 \leq t \leq T$, where $||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||$ is the Euclidean distance between nodes i and j at time t, $\boldsymbol{\beta} \in \mathbb{R}^p$ are the model parameters that do not change with t, and $h : \mathbb{R}^{p+1} \to [0, 1]$ is a function that is strictly decreasing in its first argument. This model assumes that a smaller distance between two nodes in the latent space yields a larger link probability between them.

Different forms of the function h have been suggested in the literature. The distance model in Hoff, Raftery and Handcock (2002) assumes that $\boldsymbol{\beta}$ is a one-dimensional intercept term, and $\text{logit}[h(||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||, \beta)] = \beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||$. However, the variation inference based on this formulation involves several additional approximation steps (Salter-Townshend and Murphy (2013)). Gollini and Murphy (2016) suggested using $\text{logit}[h(||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||, \beta)] = \beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2$, which can reduce the number of approximation steps in VI. Sewell and Chen (2015) proposed a more complicated h function to distinguish the effects of the sender and the receiver in edge formation.

Here, we adopt the h function suggested by Gollini and Murphy (2016) to derive the VI. More precisely, for any $1 \leq i \neq j \leq n$ and $1 \leq t \leq T$, $\text{logit}[p(Y_{ijt} = 1|\boldsymbol{X}_{it}, \boldsymbol{X}_{jt}, \beta)] = \beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2$. Let $\boldsymbol{\mathcal{X}} = \{\boldsymbol{X}_{it} : 1 \leq i \leq n,\ 1 \leq t \leq T\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{Y}_t : 1 \leq t \leq T\}$. Then, the model can be written as

$$p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \beta) = \prod_{t=1}^{T} \prod_{1 \leq i \neq j \leq n} \frac{\exp\{Y_{ijt}\left(\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2\right)\}}{1 + \exp\{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2\}}. \tag{2.1}$$

3. A Variational Algorithm for Posterior Inference

VI or VB (Jordan et al. (1999); Wainwright and Jordan (2008)) is a powerful tool for approximating intractable complex distributions. The basic idea of VB

is to approximate the posterior distribution by the closest member in a certain family of distributions (which is usually called the variational family). The closest member, which is referred to as the variational distribution, is then used for posterior inference. Thus, the posterior inference problem becomes an optimization problem of finding the member in the variational family that minimizes a divergence measure between the approximate posterior and the true posterior.

The most popular approach for VI is the mean-field method, which approximates the target distribution by a fully factorized distribution. In this section, we further restrict each component of the factorized distribution to be in a family of tractable distributions indexed by variational parameters. These variational parameters are chosen to minimize the Kullback–Leibler (KL) divergence between the approximate posterior and the true posterior.

Now, we derive a variational algorithm for posterior inference of the dynamic latent space model described in Section 2. We are interested in $p(\boldsymbol{\mathcal{X}}, \beta|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T)$, the posterior distribution of the intercept β, and the latent positions of the nodes $\boldsymbol{\mathcal{X}}$. We assign a normal prior $\mathcal{N}(\xi, \psi^2)$ for the intercept β, and view ξ, ψ^2, σ^2, and τ^2 as hyperparameters. Then, $p(\boldsymbol{\mathcal{X}}, \beta|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T) \propto p(\beta)p(\boldsymbol{\mathcal{X}}) \prod_{t=1}^{T} p(\boldsymbol{Y}_t|\boldsymbol{\mathcal{X}}, \beta)$. We approximate the posterior by the following family of distributions:

$$q(\boldsymbol{\mathcal{X}}, \beta) = q(\beta = \cdot|\tilde{\xi}, \tilde{\psi}^2) \prod_{t=1}^{T} \prod_{i=1}^{n} q(\boldsymbol{X}_{it} = \cdot|\tilde{\boldsymbol{\mu}}_{it}, \tilde{\Sigma}),$$

where $q(\beta = \cdot|\tilde{\xi}, \tilde{\psi}^2)$ is a normal distribution with mean $\tilde{\xi}$ and variance $\tilde{\psi}^2$, and $q(\boldsymbol{X}_{it} = \cdot|\tilde{\boldsymbol{\mu}}_{it}, \tilde{\Sigma})$ is a d-dimensional normal distribution with mean vector $\tilde{\boldsymbol{\mu}}_{it}$ and covariance matrix $\tilde{\Sigma}$. Note that we can also allow the covariance matrix $\tilde{\Sigma}$ to vary with i and t; the derivation of the variational algorithm is similar.

The main hindrance in deriving the analytical form of the KL divergence between $q(\boldsymbol{\mathcal{X}}, \beta)$ and $p(\boldsymbol{\mathcal{X}}, \beta|\boldsymbol{\mathcal{Y}})$ is the expectation of the log-likelihood $\mathbb{E}_q[\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \beta)]$, which does not have an analytical form. Therefore, instead of working with the original KL divergence, we approximate it by the following lower bound:

$$\begin{aligned}
&\mathbb{E}_q[\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \beta)] \\
&= \sum_{t=1}^{T} \sum_{i \neq j} \mathbb{E}_q \left[Y_{ijt}(\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2) - \log\left(1 + e^{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}\right) \right] \\
&\geq \sum_{t=1}^{T} \sum_{i \neq j} \left[Y_{ijt}(\tilde{\xi} - \mathbb{E}_q[||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2]) \right] - \log\left(1 + \mathbb{E}_q\left[e^{\beta - ||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||^2}\right]\right)
\end{aligned}$$

$$= \sum_{t=1}^{T}\sum_{i\neq j}\Bigg\{ Y_{ijt}\left(\tilde{\xi} - 2\,\mathrm{tr}(\tilde{\Sigma}) - ||\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt}||^2\right) - \log\left(1 + \frac{\exp\{\tilde{\xi} + (1/2)\tilde{\psi}^2\}}{\det(\mathbb{I} + 4\tilde{\Sigma})^{1/2}} \cdot \exp\{-(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})^T(\mathbb{I} + 4\tilde{\Sigma})^{-1}(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})\}\right)\Bigg\},$$

where the inequality is from Jensen's inequality. Then, the KL divergence $D[q(\boldsymbol{\mathcal{X}}, \beta)||p(\boldsymbol{\mathcal{X}}, \beta|\boldsymbol{\mathcal{Y}})]$ between the approximate posterior and the true posterior can be approximated by an upper bound, as follows:

$$\begin{aligned}
&D\left[q(\boldsymbol{\mathcal{X}}, \beta)||p(\boldsymbol{\mathcal{X}}, \beta|\boldsymbol{\mathcal{Y}})\right] \\
&:= \mathbb{E}_q[\log q(\boldsymbol{\mathcal{X}})] - \mathbb{E}_q[\log p(\boldsymbol{\mathcal{X}})] + \mathbb{E}_q[\log q(\beta)] - \mathbb{E}_q[\log p(\beta)] \\
&\quad - \mathbb{E}_q[\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}}, \beta)] + constant \\
&\leq -\frac{nT}{2}\log(\det(\tilde{\Sigma})) + \left(\frac{n}{2\sigma^2} + \frac{n(T-1)}{\tau^2}\right)\mathrm{tr}(\tilde{\Sigma}) \\
&\quad + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|\tilde{\boldsymbol{\mu}}_{i1}\|^2 + \frac{1}{2\tau^2}\sum_{t=2}^{T}\sum_{i=1}^{n}\left\|\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{i(t-1)}\right\|^2 \\
&\quad + \frac{1}{2}\left(\frac{\tilde{\psi}^2}{\psi^2} - \log\frac{\tilde{\psi}^2}{\psi^2} + \frac{(\tilde{\xi} - \xi)^2}{\psi^2}\right) - \sum_{t=1}^{T}\sum_{i\neq j}\Bigg\{ Y_{ijt}\left(\tilde{\xi} - 2\,\mathrm{tr}(\tilde{\Sigma}) - ||\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt}||^2\right) \\
&\quad - \log\left(1 + \frac{\exp\{\tilde{\xi} + (1/2)\tilde{\psi}^2\}}{\det(\mathbb{I} + 4\tilde{\Sigma})^{1/2}} \cdot \exp\{-(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})^T(\mathbb{I} + 4\tilde{\Sigma})^{-1}(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})\}\right)\Bigg\} \\
&\quad + constant. \qquad (3.1)
\end{aligned}$$

Two constant terms in the above derivation are not written out explicitly and will be omitted later, because they do not play a role in the optimization procedure. Compared with the multiple approximation steps in the derivation of the variational algorithm in Salter-Townshend and Murphy (2013), only one approximation step based on Jensen's inequality is used in our derivation. Note too that this Jensen's bound is tighter than the lower bound given by a first-order approximation.

Our goal is to minimize the approximated KL divergence (3.1) over the variational parameters $\{\tilde{\boldsymbol{\mu}}_{it}\}$, $\tilde{\Sigma}$, $\tilde{\xi}$, and $\tilde{\psi}^2$. For convenience, we define

$$\begin{aligned}
&f(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}, \tilde{\xi}, \tilde{\psi}^2) := \\
&\sum_{t=1}^{T}\sum_{i\neq j}\log\left(1 + \frac{\exp\{\tilde{\xi} + (1/2)\tilde{\psi}^2\}}{\det(\mathbb{I} + 4\tilde{\Sigma})^{1/2}} \cdot \exp\{-(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})^T(\mathbb{I} + 4\tilde{\Sigma})^{-1}(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})\}\right)\Bigg\}.
\end{aligned}$$

In order to avoid searching for a numerical root in the optimization over $\{\tilde{\boldsymbol{\mu}}_{it}\}$, we use the Taylor expansion to approximate the gradient function $f'(x) = f'(x_0) + f''(x_0)(x - x_0) + o(x - x_0)$.

Assume we have estimates of the variational parameters after s iterations. Then, the update equation for each variational parameter in the $(s+1)$th iteration is as follows:

- Update of $\tilde{\Sigma}$:

$$\tilde{\Sigma}^{(s+1)} \leftarrow \frac{nT}{2}\left[\left(\frac{n}{2\sigma^2} + \frac{n(T-1)}{\tau^2} + \sum_{t=1}^{T}\sum_{i\neq j} 2Y_{ijt}\right)\mathbb{I}_d + J(\tilde{\Sigma}^{(s)})\right]^{-1},$$

  where $J(\tilde{\Sigma}^{(s)})$ is the Jacobian matrix of f evaluated at $\tilde{\Sigma}^{(s)}$, and has the following expression:

$$J(\tilde{\Sigma}^{(s)}) = \sum_{t=1}^{T}\sum_{i\neq j}\frac{1}{A_{ijt}}\Big(- 8(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1}(\tilde{\boldsymbol{\mu}}_{it}^{(s)} - \tilde{\boldsymbol{\mu}}_{jt}^{(s)}) (\tilde{\boldsymbol{\mu}}_{it}^{(s)} - \tilde{\boldsymbol{\mu}}_{jt}^{(s)})^T(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1}\Big),$$

  where $A_{ijt} = 1 + (\exp\{-\tilde{\xi} - (1/2)\tilde{\psi}^2\}/\det(\mathbb{I} + 4\tilde{\Sigma}^{(s)})^{-1/2}) \cdot \exp\{(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})^T (\mathbb{I} + 4\tilde{\Sigma}^{(s)})^{-1}(\tilde{\boldsymbol{\mu}}_{it} - \tilde{\boldsymbol{\mu}}_{jt})\}$.

- Update of $\tilde{\boldsymbol{\mu}}_{it}$:

$$\tilde{\boldsymbol{\mu}}_{it}^{(s+1)} \leftarrow \left(H(\tilde{\boldsymbol{\mu}}_{it}^{(s)}) + \left(\frac{2}{\tau^2} + \sum_{j\neq i} 2(Y_{ijt} + Y_{jit})\right)\mathbb{I}_d\right)^{-1} \cdot \left[\sum_{j\neq i} 2(Y_{ijt} + Y_{jit})\tilde{\boldsymbol{\mu}}_{jt}^{(s)} + \frac{1}{\tau^2}(\tilde{\boldsymbol{\mu}}_{i(t-1)}^{(s)} + \tilde{\boldsymbol{\mu}}_{i(t+1)}^{(s)}) + H(\tilde{\boldsymbol{\mu}}_{it}^{(s)})\tilde{\boldsymbol{\mu}}_{it}^{(s)} - G(\tilde{\boldsymbol{\mu}}_{it}^{(s)})\right],$$

  where $G(\tilde{\boldsymbol{\mu}}_{it}^{(s)})$ is the gradient of f evaluated at $\tilde{\boldsymbol{\mu}}_{it}^{(s)}$, and $H(\tilde{\boldsymbol{\mu}}_{it}^{(s)})$ is the Hessian matrix of f evaluated at $\tilde{\boldsymbol{\mu}}_{it}^{(s)}$; that is,

$$G(\tilde{\boldsymbol{\mu}}_{it}^{(s)}) = \sum_{i\neq j}\frac{-2}{A_{ijt}}(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1}(\tilde{\boldsymbol{\mu}}_{it}^{(s)} - \tilde{\boldsymbol{\mu}}_{jt}^{(s)}),$$

$$H(\tilde{\boldsymbol{\mu}}_{it}^{(s)}) = \sum_{i\neq j}\frac{4}{1 + A_{ijt} + 1/(A_{ijt} - 1)}(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1} (\tilde{\boldsymbol{\mu}}_{it}^{(s)} - \tilde{\boldsymbol{\mu}}_{jt}^{(s)})(\tilde{\boldsymbol{\mu}}_{it}^{(s)} - \tilde{\boldsymbol{\mu}}_{jt}^{(s)})^T(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1} - \frac{2}{1 + A}(\mathbb{I}_d + 4\tilde{\Sigma}^{(s)})^{-1}.$$

- Update of $\tilde{\xi}$: $\tilde{\xi}^{(s+1)} \leftarrow (1+\psi^2 f''(\tilde{\xi}^{(s)}))^{-1}[\xi+\psi^2(\sum_{t=1}^{T}\sum_{i\neq j} Y_{ijt}+f''(\tilde{\xi}^{(s)})\tilde{\xi}^{(s)}-f'(\tilde{\xi}^{(s)}))]$.
- Update of $\tilde{\psi}^2$: $\tilde{\psi}^{2\ (s+1)} \leftarrow (1/\psi^2 + 2f'(\tilde{\psi}^{2\ (s)}))^{-1}$.

The algorithm converges when the relative change in the log-likelihood at two consecutive steps is smaller than some threshold value. After convergence, the variational parameter $\tilde{\boldsymbol{\mu}}_{it}$ is the estimated posterior mean of the latent position of node i at time t, and $\tilde{\xi}$ is the estimated posterior mean of the intercept.

4. Theoretical Properties

In this section, we provide theoretical properties of variational algorithms for latent space models, which have not previously been studied in the literature. In particular, we show that under certain regularity conditions, the point estimates of the model parameters from the VB procedure converge to the true parameters as the number of nodes goes to infinity. We achieve this result by showing that the VB risk goes to zero as the number of nodes goes to infinity.

We follow the framework of Yang, Pati and Bhattacharya (2020) in which the authors studied a slightly modified VB procedure, called α-VB, with the standard VB algorithm as a special case. The authors showed that the point estimates given by this VB procedure converge to the true parameters. The main idea of the proof is to establish a finite-sample upper bound of the VB risk using the VB objective function, and then to give the convergence rate of this upper bound. Note that the theoretical results in Yang, Pati and Bhattacharya (2020) are mainly for independent and identically distributed (i.i.d.) data, whereas our results are tailored to dependent random variables in the network setting.

We first introduce some notation. Let $D(p||q) = \int p\log(p/q)d\mu$ denote the KL divergence between the probability density functions p and q with respect to a measure μ. For $\alpha \in (0,1)$, let $D_\alpha(p||q) := (1/\alpha)\log\int p^\alpha q^{1-\alpha}d\mu$ denote the α-divergence between the probability density functions p and q. In this section, we use $\theta := (\beta, \sigma^2, \tau^2)$ to denote all the parameters in the model, and $\pi := (\sigma^2, \tau^2)$ to denote the parameters related to the latent variables. We use q_θ and $q_{\boldsymbol{\mathcal{X}}}$ to denote the variational distribution of the model parameters θ and the variational distribution of the latent variables $\boldsymbol{\mathcal{X}}$, respectively. In addition, for the rest of this section, we assume θ has the true value θ^*.

Now, we follow the framework of Yang, Pati and Bhattacharya (2020) to derive a further decomposition of the VB objective function. If we adopt the standard VB algorithm, we minimize the KL divergence between the variational posterior and the true posterior; that is, we minimize the following objective

function over the members of the variational family:

$$\begin{aligned}
&D(q(\theta,\boldsymbol{\mathcal{X}})||p(\theta,\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}))\\
&=\mathbb{E}_q[\log q(\boldsymbol{\mathcal{X}})]+\mathbb{E}_q[\log q(\theta)]-\mathbb{E}_q[\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\theta)]-\mathbb{E}_q[\log p(\boldsymbol{\mathcal{X}}|\theta)]\\
&\quad-\mathbb{E}_q[\log p(\theta)]+constant\\
&=D(q(\theta)||p(\theta))-\int\int\log\left[\frac{p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{X}},\theta)p(\boldsymbol{\mathcal{X}}|\theta)}{q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}})}\right]q_\theta(d\theta)q_{\boldsymbol{\mathcal{X}}}(d\boldsymbol{\mathcal{X}})+constant.
\end{aligned}\tag{4.1}$$

Let $l_n(\theta):=\log p(\boldsymbol{\mathcal{Y}}|\theta)=\log(\int p(\boldsymbol{\mathcal{Y}}|\theta,\boldsymbol{\mathcal{X}})p(\boldsymbol{\mathcal{X}}|\theta)d\boldsymbol{\mathcal{X}})$ and $\hat{l}_n(\theta):=\int\log($ $p(\boldsymbol{\mathcal{Y}}|\theta,\boldsymbol{\mathcal{X}})p(\boldsymbol{\mathcal{X}}|\theta)/q_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\mathcal{X}}))q_{\boldsymbol{\mathcal{X}}}(d\boldsymbol{\mathcal{X}})$. By Jensen's inequality, $l_n(\theta)\geq\hat{l}_n(\theta)$. Then, the KL divergence (4.1) can be decomposed into three parts, as follows:

$$\begin{aligned}
&D(q(\theta,\boldsymbol{\mathcal{X}})||p(\theta,\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{Y}}))\\
&=D(q(\theta)||p(\theta))+\int\Big(l_n(\theta)-\hat{l}_n(\theta)\Big)q_\theta(d\theta)-\int l_n(\theta)q_\theta(d\theta)+constant,
\end{aligned}\tag{4.2}$$

where the first term is the discrepancy between the variational distribution and the prior of the model parameters, the second term is the average Jensen gap (denoted by $\Delta_J(q_\theta,q_{\boldsymbol{\mathcal{X}}})$) from the variational approximation (which is the only term that involves the variational distribution $q(\boldsymbol{\mathcal{X}})$), and the third term is the integrated log-likelihood. The constant term will be omitted later.

Define the following objective function:

$$\Psi_n(q_\theta,q_{\boldsymbol{\mathcal{X}}}):=-\int_\Theta(l_n(\theta)-l_n(\theta^*))q(d\theta)+\Delta_J(q_\theta,q_{\boldsymbol{\mathcal{X}}})+D(q(\theta)||p(\theta)),$$

where the subscript n indicates the dependence of the objective function on the number of nodes n. Note that minimizing Ψ_n over $(q_\theta,q_{\boldsymbol{\mathcal{X}}})$ is equivalent to minimizing the KL divergence (4.2) over $(q_\theta,q_{\boldsymbol{\mathcal{X}}})$, because $l_n(\theta^*)$ does not depend on the variational distribution $q(\boldsymbol{\mathcal{X}})$.

Yang, Pati and Bhattacharya (2020) proposed a slightly different procedure, α-VB, in which a stronger penalty on the discrepancy between the variational distribution and the prior is introduced into the objective function,

$$\Psi_{n,\alpha}(q_\theta,q_{\boldsymbol{\mathcal{X}}}):=-\int_\Theta(l_n(\theta)-l_n(\theta^*))q(d\theta)+\Delta_J(q_\theta,q_{\boldsymbol{\mathcal{X}}})+\frac{1}{\alpha}D(q(\theta)||p(\theta)),\tag{4.3}$$

where $\alpha\in(0,1]$ is a tuning parameter, q_θ and $q_{\boldsymbol{\mathcal{X}}}$ are variational distributions that are restricted in some variational families Γ_θ and $\Gamma_{\boldsymbol{\mathcal{X}}}$, respectively. Note that

the α-VB objective function $\Psi_{n,\alpha}$ reduces to Ψ_n when $\alpha = 1$. The α-VB solution is defined by $(\hat{q}_{\theta,\alpha}, \hat{q}_{\boldsymbol{\mathcal{X}},\alpha}) := \operatorname{argmin}_{q_\theta \in \Gamma_\theta, q_{\boldsymbol{\mathcal{X}}} \in \Gamma_{\boldsymbol{\mathcal{X}}}} \Psi_{n,\alpha}(q_\theta, q_{\boldsymbol{\mathcal{X}}})$. Note that all of the theoretical results presented here are on this global optimum $(\hat{q}_{\theta,\alpha}, \hat{q}_{\boldsymbol{\mathcal{X}},\alpha})$, which may not be achieved in practice.

We use the average α-divergence $(1/n(n-1)T)D_\alpha^{(n)}(\theta, \theta^*) := (1/n(n-1)T)$ $D_\alpha[p_\theta^{(n)} || p_{\theta^*}^{(n)}]$ as the loss function, where $p_\theta^{(n)}$ denotes the distribution of $\boldsymbol{\mathcal{Y}}$ given the model parameter θ. This loss function measures the discrepancy between the distribution of $\boldsymbol{\mathcal{Y}}$ with model parameter θ and the distribution of $\boldsymbol{\mathcal{Y}}$ with the true model parameter θ^*.

The following theorem gives a finite-sample upper bound of the VB risk for the case $\alpha < 1$.

Theorem 1. *With a certain choice of variational family $\Gamma_{\boldsymbol{\mathcal{X}}}$ and the variational distribution $q_\theta(\theta)$ restricted to a certain KL-neighborhood (defined later), for any $\zeta \in (0,1)$, $D > 1$, and $(\epsilon_\beta, \epsilon_\pi) \in (0,1)^2$, it holds with probability at least $(1 - 2/(D-1)^2 n(\epsilon_\beta^2 + \epsilon_\pi^2))$ that*

$$\int \frac{D_\alpha^{(n)}(\theta, \theta^*)}{n(n-1)T} \hat{q}_{\theta,\alpha}(d\theta) \leq \frac{D\alpha}{1-\alpha}(\epsilon_\pi^2 + \epsilon_\beta^2) - \frac{\log P_\pi(\mathcal{B}_n(\pi^*, \epsilon_\pi))}{n(n-1)T(1-\alpha)} - \frac{\log P_\beta(\mathcal{B}_n(\beta^*, \epsilon_\beta))}{n(n-1)T(1-\alpha)}, \tag{4.4}$$

where P_π and P_β are probability measures corresponding to the prior densities p_π and p_β, respectively. Here, $\mathcal{B}_n(\pi^*, \epsilon_\pi)$ and $\mathcal{B}_n(\beta^*, \epsilon_\beta)$ are KL-neighborhoods for model parameters defined in the following way:

$$\mathcal{B}_n(\pi^*, \epsilon_\pi) := \left\{ \pi : D\left(p(\boldsymbol{\mathcal{X}}_1|\pi^*) || p(\boldsymbol{\mathcal{X}}_1|\pi)\right) \leq \epsilon_\pi^2, \quad V\left(p(\boldsymbol{\mathcal{X}}_1|\pi^*) || p(\boldsymbol{\mathcal{X}}_1|\pi)\right) \leq \epsilon_\pi^2 \right\},$$

$$\mathcal{B}_n(\beta^*, \epsilon_\beta) := \Bigg\{ \beta : \sup_{\boldsymbol{X}_{11}, \boldsymbol{X}_{21}} D(p(Y_{121}|\beta^*, \boldsymbol{X}_{11}, \boldsymbol{X}_{21}) || p(Y_{121}|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})) \leq \epsilon_\beta^2,$$
$$\sup_{\boldsymbol{X}_{11}, \boldsymbol{X}_{21}} V(p(Y_{121}|\beta^*, \boldsymbol{X}_{11}, \boldsymbol{X}_{21}) || p(Y_{121}|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})) \leq \epsilon_\beta^2 \Bigg\},$$

where $\mathcal{X}_1 = \{\boldsymbol{X}_{1t} : 1 \leq t \leq T\}$, and $V(p||q) := \int p \log^2(p/q) d\mu$ denotes the discrepancy measure between two probability density functions p and q.

The left-hand side of inequality (4.4) is the VB risk. The upper bound is obtained based on a certain choice of variational family $\Gamma_{\boldsymbol{\mathcal{X}}}$ and a theorem in Yang, Pati and Bhattacharya (2020) that connects the VB risk to the α-VB objective function (4.3). The proofs of all theorems are provided in the Supplementary Material.

The following theorem gives the convergence rate of the VB risk.

Theorem 2. *Assume that the prior densities p_β and p_π are thick and continuous at β^* and π^*, respectively (here, "thick" means $p_\beta(\beta^*) > 0$ and $p_\pi(\pi^*) > 0$). Then, there exists a constant $C > 0$, such that as $n \to \infty$, it holds with probability tending to one that*

$$\int \frac{1}{n(n-1)T} D_\alpha^{(n)}(\theta, \theta^*) \hat{q}_{\theta,\alpha}(d\theta) \lesssim \frac{C}{n}.$$

Theorem 2 implies that the point estimate of the model parameter based on optimizing the α-VB objective function ($\alpha < 1$) converges to the true parameter value as $n \to \infty$; that is, α-VB provides consistent parameter estimation.

For the usual VB ($\alpha = 1$), stronger conditions are required to obtain the variational risk bound. Note that for this part, we let the loss function be the squared Hellinger distance $h^2(p||q) := \int (\sqrt{p} - \sqrt{q})^2 d\mu$. In addition, we define $h^2(\theta||\theta^*) := h^2(p(\boldsymbol{\mathcal{Y}}|\theta)||p(\boldsymbol{\mathcal{Y}}|\theta^*))$ as the squared Hellinger distance between the distributions of the observed networks generated by the parameters θ and θ^*. We restrict the model parameters in the compact set $[-M, M] \times [m, M]^2$, where $M > m > 0$ are some constants. We first state the assumptions.

Assumption 1. *The prior densities p_β and p_π satisfy $\inf_\beta p_\beta(\beta) > 0$ and $\inf_\pi p_\pi(\pi) > 0$.*

Assumption 2. *There exists a constant $C > 0$ such that the following inequalities hold for any (β, π) and (β', π') in the parameter space:*

$$\begin{aligned}
D\left(p(Y_{121} = \cdot|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})||p(Y_{121} = \cdot|\beta', \boldsymbol{X}_{11}, \boldsymbol{X}_{21})\right) &\le C|\beta - \beta'|^2, \\
V\left(p(Y_{121} = \cdot|\beta, \boldsymbol{X}_{11}, \boldsymbol{X}_{21})||p(Y_{121} = \cdot|\beta', \boldsymbol{X}_{11}, \boldsymbol{X}_{21})\right) &\le C|\beta - \beta'|^2, \\
D\left(p(\boldsymbol{\mathcal{X}}_1 = \cdot|\pi)||p(\boldsymbol{\mathcal{X}}_1 = \cdot|\pi')\right) &\le C\|\pi - \pi'\|^2, \\
V\left(p(\boldsymbol{\mathcal{X}}_1 = \cdot|\pi)||p(\boldsymbol{\mathcal{X}}_1 = \cdot|\pi')\right) &\le C\|\pi - \pi'\|^2.
\end{aligned}$$

Assumption 1 is the prior thickness condition. Assumption 2 is a regularity condition that justifies the prior concentration condition, which ensures that the prior mass of the KL neighborhood around the true parameter values is not too small. The priors used in our implementation can be shown to satisfy these assumptions by simple calculation.

The following theorem gives a high probability VB risk bound and the asymptotic variational posterior concentration result for the usual VB.

Theorem 3. *Under Assumptions 1 and 2, for any $D > 1$, it holds with probability at least $(1 - 1/(D-1)^2 n\epsilon_n^2)$ that for any $\epsilon \in [\epsilon_n, e^{cn(n-1)\epsilon_n^2}]$,*

$$\hat{Q}_\theta \left(h^2(\theta||\theta^*) \leq \epsilon^2 \right) \to 1 \text{ as } n \to \infty,$$

where $\hat{Q}_\theta$ is the probability measure corresponding to the variational distribution $\hat{q}_\theta$ given by the VB procedure.

Using this result, we can obtain the convergence rate for a truncated version of the VB risk. See the proof of Theorem 3 in the Supplementary Material for details.

5. Simulation Results

All computations are performed on a Linux machine with 2.20 GHz processors. In all simulation and real-data analyses, we set the dimension of the latent space $d = 2$ for better visualization. More details of the implementation and additional simulation studies can be found in the Supplementary Material.

We evaluate the performance using the area under the ROC curve (AUC) values of in-sample predictions. To calculate this criterion, we plugged the estimated posterior means of the model parameters and latent positions into (2.1) and calculated the estimated link probabilities. Then, we compared the estimated link probabilities with the observed data $\mathcal{Y}$. A value of one implies perfect model fitting, and a value of 0.5 implies random predictions. Note that although the AUC criterion does not measure the discrepancy between the approximate posterior and the true posterior directly, it measures the goodness-of-fit of the model to the observed data.

We carried out simulation studies for networks with $n = 100$ and $n = 1{,}000$ nodes under various settings (see the Supplementary Material for the simulation with $n = 5{,}000$ nodes). We simulated 20 dynamic networks with the number of time steps $T = 10$ for each case. The average AUC values are reported in Figure 1 for $n = 100$ and in Figure 2 for $n = 1{,}000$. The variational method performed well in all cases. Although the variance of the transition distribution does not seem to have much effect on performance, the variational method performed better on dense networks than it did on sparse networks.

Figure 3 plots the distribution of the pairwise distance ratios for each simulated network in the dense, small transition case with $n = 100$. That is, we calculated the ratio $||\hat{\boldsymbol{\mu}}_{it} - \hat{\boldsymbol{\mu}}_{jt}||/||\boldsymbol{X}_{it} - \boldsymbol{X}_{jt}||$ for each (i, j, t) of each simulated network, and plotted the density curve of these ratios. Most of these distributions are narrow and centered around one, which indicates that the estimated latent positions are close to the truth.

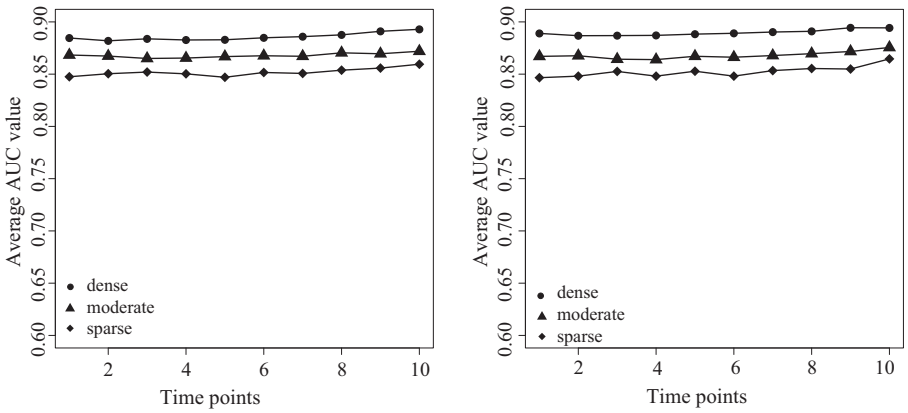

Figure 1. The average AUC values for VB on simulated networks with $n = 100$ nodes and (left) small transition and (right) large transition.

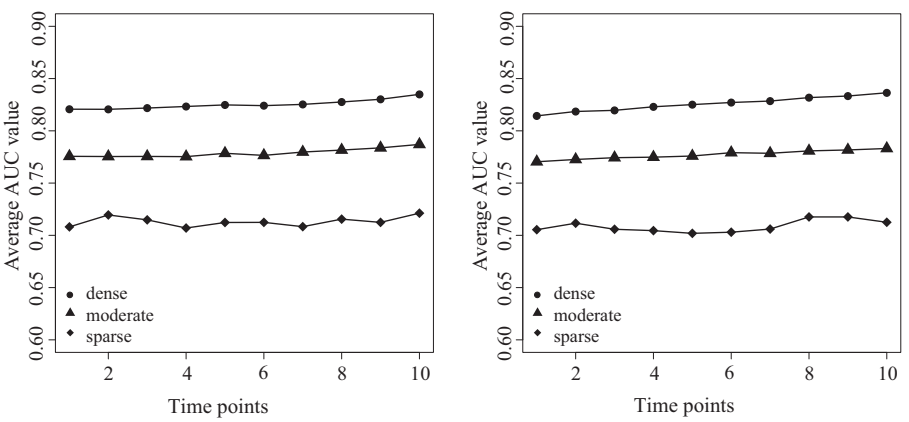

Figure 2. The average AUC values for VB on simulated networks with $n = 1{,}000$ nodes and (left) small transition and (right) large transition.

6. Real-Data Analysis

6.1. Dynamic networks from "Teenage Friends and Lifestyle Study"

Here, we analyze a sequence of directed networks of friendship relations from the "Teenage Friends and Lifestyle Study" data set (Michell and Amos (1997); Michell and West (1996); Pearson, Steglich and Snijders (2006)). In this longitudinal study, a total of 160 pupils were studied over a three-year period from January 1995. At the measurement time point in each year, the pupils were asked

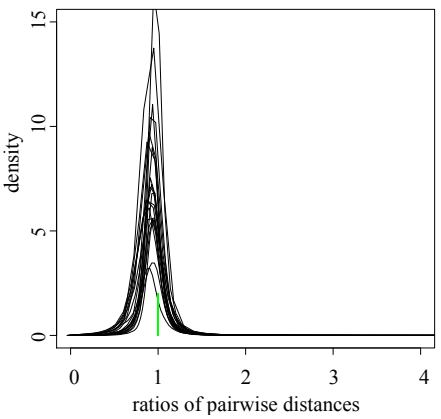

Figure 3. Density plots of ratios of pairwise distances, comparing the estimated latent positions with true latent positions.

to name up to 12 best friends, yielding three adjacency matrices. The (i, j)th entry of the tth adjacency matrix is one if pupil i named pupil j as one of his or her best friends at the tth measurement time point, and zero otherwise. The study also collected information on substance use and adolescent behavior, such as music preference and tobacco, alcohol, and cannabis consumption. We focus on the networks formed by the 129 pupils who were present at all three measurement time points (see the networks in the Supplementary Material). The average edge density of this network is 0.0274.

We fitted the dynamic latent space model to the sequence of networks. To implement the proposed VB algorithm, we set a normal prior $\mathcal{N}(0, 2)$ for the intercept β and the hyperparameters $\sigma^2 = 0.5$ and $\tau^2 = 0.1$. The variational parameters $\{\tilde{\boldsymbol{\mu}}_{it}\}$ were randomly initialized, and the initial values of the other variational parameters were set to $\tilde{\Sigma}_0 = \mathbb{I}_2$, $\tilde{\xi}_0 = 0$, and $\tilde{\psi}_0^2 = 2$. The algorithm converged in less than five seconds. The approximate posterior distribution of the intercept β is $\mathcal{N}(-1.5092, 0.0001)$. The AUC values of the in-sample predictions for the three time points are 0.9364, 0.9497, and 0.9681, respectively.

Figure 4 shows the estimated latent positions at the first time point based on the approximate posterior, as well as where the actors are moving to in the next two time points (indicated by the arrows). A longer arrow indicates a larger move in the latent space. Filled squares denote boys, and filled circles denote girls. We can see that the girls lie on the top-left side of the latent space, whereas the boys lie on the bottom-right side. Only a few actors (actors 6, 8, 32, 93, 95) are close to the opposite gender in the social space. This coincides with the claim

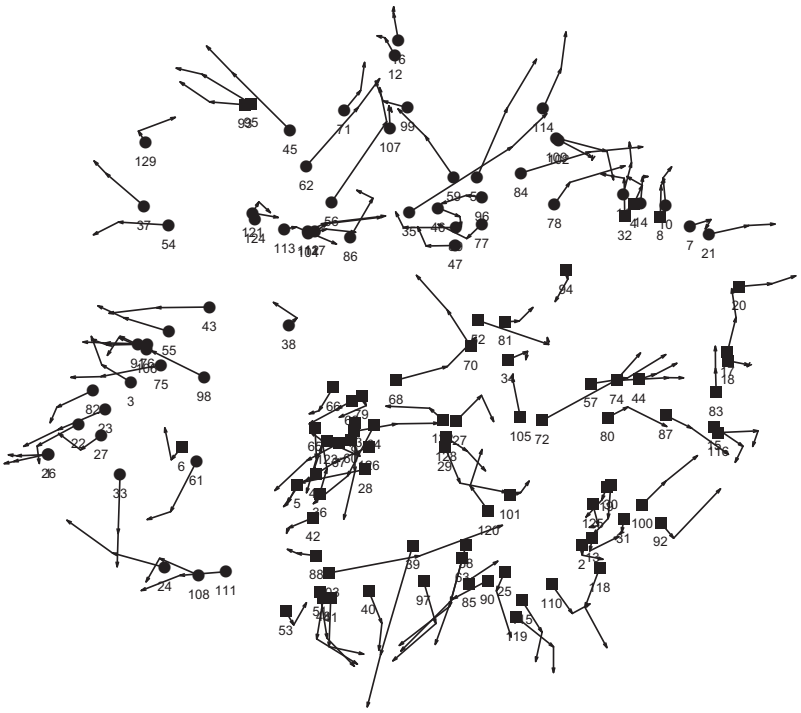

Figure 4. Estimated latent positions of each node at the first time point. The arrows denote where the actors are moving to in the next two time points. Filled squares denote boys, and filled circles denote girls.

in Pearson, Steglich and Snijders (2006) that there is strong gender homophily in friendship selection. From the trajectories of the nodes, we can also see the formation of groups over time. For example, actors 15, 80, 87, 92, 100, and 116 were forming a new group, whereas actors 35 and 50 were leaving their original social group.

To study the dynamics of the network, we calculated the squared distances of the movements for each node during the two transitions and created box plots for them (Figure 5). From these plots, we can see that the median and variation of the moving distances of the first transition are larger than those of the second transition, which indicates that the friendship network changed more during the first transition. Pearson, Steglich and Snijders (2006) analyzed the same data set, also claiming that the rate of network change is larger in the first transition than it is in the second transition.

We also examined tobacco and cannabis consumption of these 129 pupils. Figures 6 and 7 give the latent positions of the pupils, as well as their tobacco

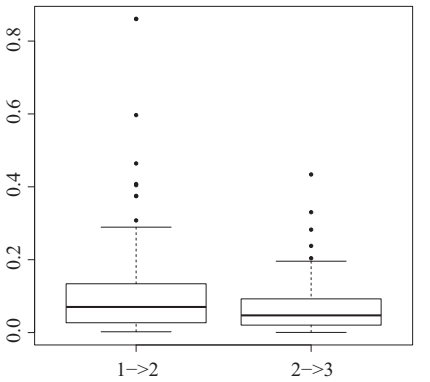

Figure 5. Box plots for the squared distances of movements for each node during two transitions.

and cannabis consumption status at each time point. Note that pupils with similar substance consumption behavior tended to move closer to each other. For example, actors 24, 61, 108, and 111 occasionally or regularly used tobacco and cannabis at the first time point, and they moved closer to each other in the latent space during the two transitions. Another interesting observation is that as actor 63, who occasionally or regularly used cannabis at the first time point, moved into the nearby social group, the whole group of pupils (actors 40, 41, 48, 58, 90, and 97) became occasional or regular cannabis users by time point 3. This observation corroborates the conclusion in Pearson, Steglich and Snijders (2006) that there is a significant positive influence effect of friends on cannabis use.

6.2. The wiki-talk temporal network

In this section, we analyze the Wiki-talk temporal network (Leskovec, Huttenlocher and Kleinberg (2010); Paranjape, Benson and Leskovec (2017)). Wiki-talk pages are part of the Wikipedia administration system, where users are able to communicate with each other on possible improvements to Wikipedia pages. Each user page has an associated talk page. Wikipedia users can edit a user's talk page by leaving a message.

The temporal network we analyzed represents the editing activities of Wikipedia users on each other's Wiki-talk pages. Nodes represent Wikipedia users, and an edge from node i to node j at time point t means that user i edited user j's Wiki-talk page at time t. The original data set contains 1,140,149 nodes and spans 2,277 days. We took a subset of 5,294 nodes and aggregated the temporal edges between these nodes in each of the last three years to form three

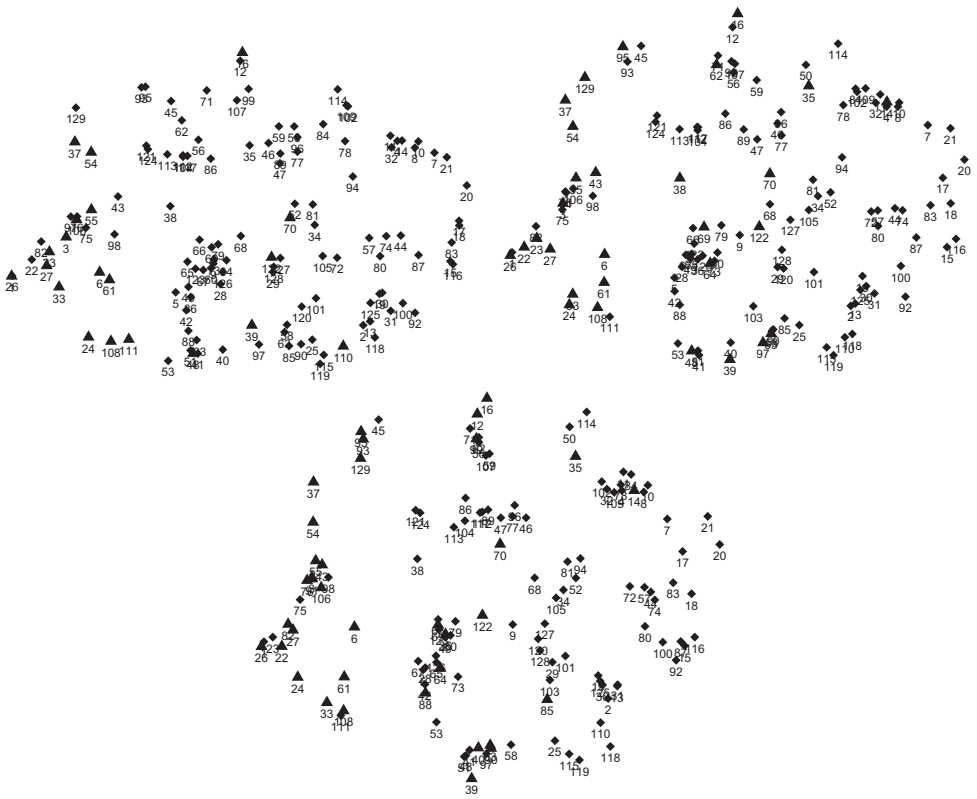

Figure 6. The latent positions and tobacco consumption of the 129 pupils at times 1, 2 (top), and 3 (bottom). The filled triangles represent actors who used tobacco occasionally or regularly. The filled diamonds represent actors who never used tobacco or only tried once.

networks. All 5,294 nodes are present at all three time points. The edge densities of the observed network at the three time points are 0.0022, 0.0030, and 0.0012, respectively.

We implemented the proposed VB algorithm with a normal prior $\mathcal{N}(0, 0.01)$ on the intercept β and the hyperparameters $\sigma^2 = 5$ and $\tau^2 = 0.01$. The variational parameters $\tilde{\boldsymbol{\mu}}_{it}$ were initialized randomly, and the initial values of the other variational parameters were set as $\tilde{\Sigma}_0 = \mathbb{I}_2$, $\tilde{\xi}_0 = 0$, and $\tilde{\psi}_0^2 = 0.01$. The computation took about 20 minutes. The variational posterior of the intercept β is $\mathcal{N}(-3.3054,\ 5.5278 \times 10^{-6})$. The AUC values of the in-sample predictions at each time step are 0.7260, 0.7486, and 0.6955, respectively.

Figure 8 shows the estimated latent positions at each time step. Most nodes are concentrated around the center. With such a large number of nodes in the area close to the center, it is not easy to plot the movements of the nodes or

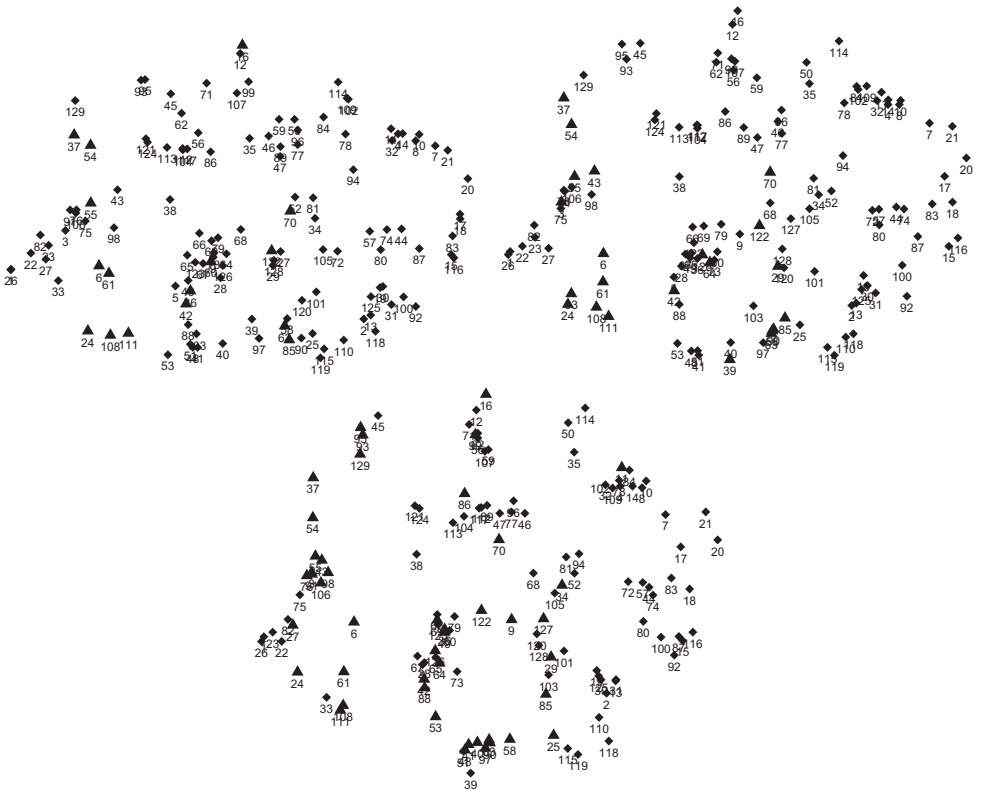

Figure 7. The latent positions and cannabis consumption of the 129 pupils at times 1, 2 (top), and 3 (bottom). The filled triangles represent actors who used cannabis occasionally or regularly. The filled diamonds represent actors who never used cannabis or only tried once.

their trajectories. Instead, we plot the latent positions of the nodes with squared moving distances greater than 5 or smaller than 0.01 during the two transitions (Figures 9 and 10). For both transitions, actors with large moving distances are more spread out in the latent space, whereas actors with small moving distances are all concentrated around the center of the latent space. They are very likely to be the most active users in terms of contributing to the website or administrative users who have been granted certain privileges.

To study the dynamics of the network, we also calculated the squared distances of the movements for each node during the two transitions and created box plots for them (Figure 11). From these plots, we can see that the ranges of these distances are larger in the second transition than they are in the first transition. The dynamics of the network did not seem to be stable during these transitions.

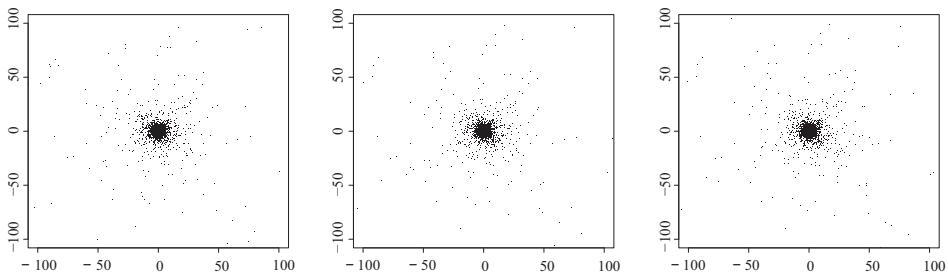

Figure 8. The estimated latent positions for the Wiki-talk data set at times 1, 2, and 3.

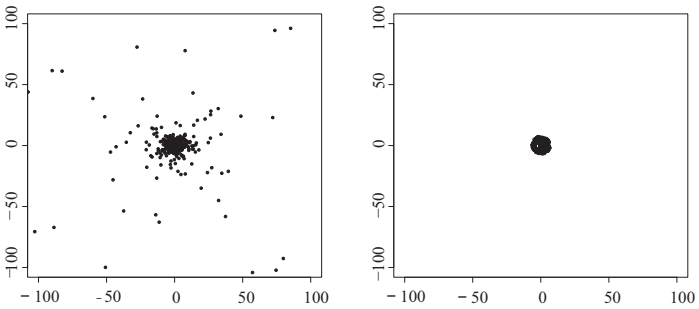

Figure 9. The initial latent positions of the nodes with squared moving distances greater than 5 (left) or smaller than 0.01 (right) during the first transition.

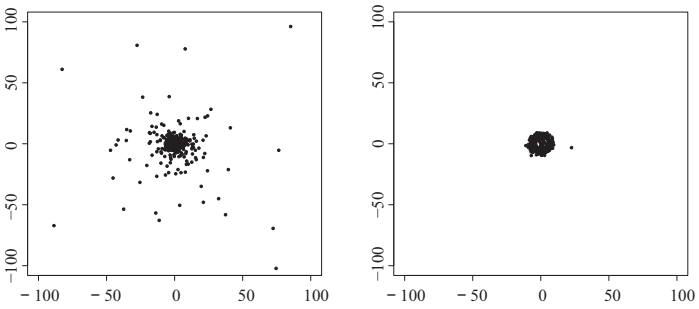

Figure 10. The latent positions at the second time point of the nodes with squared moving distances greater than 5 (left) or smaller than 0.01 (right) during the second transition.

7. Discussion

We have proposed a VB algorithm for dynamic latent space models. The proposed algorithm is able to handle large-scale networks, and performs well for

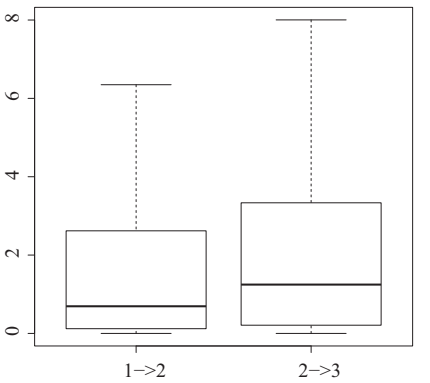

Figure 11. Box plots of the distances of the movements of the nodes in the latent space during the two transitions.

simulated and real-world networks. Furthermore, we have proved that under certain conditions, the VB risk of the VB procedure goes to zero as the number of nodes goes to infinity. To the best of our knowledge, this is the first study to propose a variational algorithm for a dynamic distance model and address its theoretical properties.

Note that the lower bound of the expected log-likelihood is not unique (see Jaakkola and Jordan (2000) for another possible lower bound). However, the one used in our derivation requires fewer approximation steps, and the resulting objective function is easier to optimize. It is also well known that VI has a tendency to underestimate the posterior variance because it approximates the posterior using a factorized distribution. Because we focus on point estimates, the underestimation of the variance does not affect our performance metric. It is of interest to investigate the uncertainty estimation based on VB for the latent space model.

In a dynamic network context, our mean-field approximation may not be the most suitable for the time series aspect of the model. For example, the generalized mean-field approach in Xing, Fu and Song (2010) factorizes the approximate posterior distribution into the product of several modules, and models each module by a state-space model. Such variation is also possible for dynamic networks. However, whereas state-space models can better capture the dependence structure, it is harder to study the theoretical properties of such algorithms.

The proposed VB algorithm can be extended to more complicated models. For example, if we assume that the initial latent positions come from a mixture

of Gaussian distributions and that their cluster assignments change over time, our method can be modified to address the community detection problem for dynamic latent space models. Additional global parameters, such as the parameters characterizing popularity and social activity in Sewell and Chen (2015), can also be incorporated into the model. Another possible extension is to incorporate dyadic-level covariate information into the dynamic network model, extending the model proposed by Krivitsky et al. (2009). It is also possible to generalize the proposed method to dynamic multilayer latent space models. It is of interest to develop a VB algorithm to accelerate the computation of such models.

In latent space models, the choice of the likelihood function is flexible. The model we used can be categorized into a larger class of latent variable models (LVMs) (see Rastelli, Friel and Raftery (2016)). Rastelli, Friel and Raftery (2016) also introduced the Gaussian latent position model (GLPM) as a special class of the LVM, which replaces the logistic link function for the edges with a non-normalized multivariate Gaussian density. The proposed VB algorithm can be modified to apply to the GLPM and dynamic GLPM. The modified algorithm will involve a similar Jensen approximation to that of the proposed VB algorithm.

Supplementary Material

The online Supplementary Material contains proofs of all theorems and details of the implementation for the simulations, as well as additional simulation studies that compare VB with MCMC, apply VB to networks with $n = 5{,}000$ nodes, investigate the effect of α in α-VB, and verify the asymptotic behavior of the proposed algorithm.

Acknowledgments

This work was supported in part by National Science Foundation grant DMS-2015561. The authors would like to thank Professor Yun Yang for the helpful discussions related to the theoretical results.

References

Ahmed, A. and Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences* **106**, 11878–11883.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics* **41**, 1922–1943.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

Celisse, A., Daudin, J.-J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6**, 1847–1899.

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

Durante, D. and Dunson, D. B. (2014a). Bayesian dynamic financial networks with time-varying predictors. *Statistics & Probability Letters* **93**, 19–26.

Durante, D. and Dunson, D. B. (2014b). Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101**, 883–898.

Durante, D., Dunson, D. B. and Vogelstein, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association* **112**, 1516–1530.

Foulds, J., DuBois, C., Asuncion, A., Butts, C. and Smyth, P. (2011). A dynamic relational infinite feature model for longitudinal social networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* **15**, 287–295.

Friel, N., Rastelli, R., Wyse, J. and Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. In *Proceedings of the National Academy of Sciences* **113**, 6629–6634.

Fu, W., Song, L. and Xing, E. P. (2009). Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning* 329–336. Association for Computing Machinery, New York.

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

Gollini, I. and Murphy, T. B. (2016). Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics* **25**, 246–265.

Guo, F., Hanneke, S., Fu, W. and Xing, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th International Conference on Machine Learning*, 321–328. Association for Computing Machinery, New York.

Han, Q., Xu, K. and Airoldi, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning* **37**, 1511–1520.

Hanneke, S., Fu, W. and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics* **4**, 585–605.

Heaukulani, C. and Ghahramani, Z. (2013). Dynamic probabilistic models for latent feature propagation in social networks. In *Proceedings of the 30th International Conference on Machine Learning* **28**, 275–283.

Ho, Q., Song, L. and Xing, E. (2011). Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* **15**, 342–350.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9**, 1169–1193.

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

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Kim, B., Lee, K. H., Xue, L. and Niu, X. (2018). A review of dynamic network models with latent variables. *Statistics Surveys* **12**, 105–135.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 29–46.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E. and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31**, 204–213.

Lee, K. H., Xue, L. and Hunter, D. R. (2020). Model-based clustering of time-evolving networks through temporal exponential-family random graph models. *Journal of Multivariate Analysis* **175**, 104540.

Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010). Governance in social media: A case study of the Wikipedia promotion process. In *Proceedings of the Fourth International AAAI Conference on Web and Social Media* **4**, 98–105.

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

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

Michell, L. and Amos, A. (1997). Girls, pecking order and smoking. *Social Science & Medicine* **44**, 1861–1869.

Michell, L. and West, P. (1996). Peer pressure to smoke: The meaning depends on the method. *Health Education Research* **11**, 39–49.

Paranjape, A., Benson, A. R. and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 601–610. Association for Computing Machinery, New York.

Pearson, M., Steglich, C. and Snijders, T. (2006). Homophily and assimilation among sport-active adolescent substance users. *Connections* **27**, 47–63.

Rastelli, R., Friel, N. and Raftery, A. E. (2016). Properties of latent variable network models. *Network Science* **4**, 407–432.

Salter-Townshend, M. and Murphy, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis* **57**, 661–671.

Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. In *ACM SIGKDD Explorations Newsletter* **7**, 31–40.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association* **110**, 1646–1657.

Sewell, D. K. and Chen, Y. (2016). Latent space models for dynamic networks with weighted edges. *Social Networks* **44**, 105–116.

Sewell, D. K. and Chen, Y. (2017). Latent space approaches to community detection in dynamic networks. *Bayesian Analysis* **12**, 351–377.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**, 1–305.

Ward, M. D., Ahlquist, J. S. and Rozenas, A. (2013). Gravity's rainbow: A dynamic latent space model for the world trade network. *Network Science* **1**, 95–118.

Wilson, J. D., Stevens, N. T. and Woodall, W. H. (2019). Modeling and detecting change in temporal networks via the degree corrected stochastic block model. *Quality and Reliability Engineering International* **35**, 1363–1378.

Xing, E. P., Fu, W. and Song, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics* **4**, 535–566.

Xu, K. (2015). Stochastic block transition models for dynamic networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 1079–1087.

Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing* **8**, 552–562.

Xu, K. S., Kliger, M. and Hero III, A. O. (2014). Adaptive evolutionary clustering. *Data Mining and Knowledge Discovery* **28**, 304–336.

Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks: A Bayesian approach. *Machine Learning* **82**, 157–189.

Yang, Y., Pati, D. and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics* **48**, 886–905.

Zhang, A. Y. and Zhou, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics* **48**, 2575–2598.

Yan Liu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: yanl5@illinois.edu

Yuguo Chen

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu

(Received March 2020; accepted March 2021)